Bayesian Decision Making

for

postgraduate

Akira Imada
Brest State Technical University, Belarus

(last modified on)

May 9, 2012

Bibliography

This lectures is partly based on the wonderful book:

- R. O. Duda, P. E. Hart and D. G. Stork (2000) "Pattern Classification." 2nd Edition, John Wiley & Sons.

Also

- S. Theodoridis and K. Koutroumbas (1998) "Pattern Recognition" Academic Press.
- K. B. Korb and A. E. Nicholson (2003) "Bayesian Artificial Intelligence." (Available from Internet without its reference list though.)
- E. Charniak (1991) "Bayesian Networks without Tears." AI MAGAZINE Vol. 12 No. 4, pp. 50-63. (Available from Internet.)

are referred.

Index

PART I
BAYESIAN CLASSIFICATION

1. Assuming all distributions are Gaussian p.d.f. with known parameters.
2. Assuming we know the form of distribution but don't know its parameters.
3. Assuming we don't know the distribution of dataset given.

1 An example to start with.

In general, we need what *features* can be used for pattern classification purpose. This is, of course, one of big issues in pattern classification, but here we assume that we already know *features* which efficiently can be used to classify our target patterns.

1.1 Which do you like better female or male?

I mean not human, but salmon. Well, we now assume, as an example, we want to classify salmons into female and male. The situation is, we have thousands of salmons and our task is to classify salmons into female and male only by observing one feature x — length of the salmon.

1.2 If we only know prior probability.

If we don't know any prior information except for how many we've had so far are female and male. Now we denote the class of female salmon as ω_1 and the class of male salmon as ω_2. Then above mentioned "how many we've had so far are female and male" are denoted as $P(\omega_1)$ and $P(\omega_2)$, respectively, and called *prior probability*.

Note here $p(\cdot)$ denotes a probability and $P(\cdot)$ denotes a density distribution.

In this case, all we should act is with the following rule.

Rule 1 (Classification only with prior probability) *If $P(\omega_1) > P(\omega_2)$ then classify it to ω_1 otherwise ω_2.*

This might seem to be a trivial reaction - not a clever strategy at all.

1.3 If we also know posterior distribution of features.

If we assume that we know further distribution of length s in each of female and male salmon, that is, $p(x|\omega_1)$ and $p(x|\omega_2)$, then our task of classify salmons with an observation of x is a little more sophisticated.

Note here $p(\cdot)$ denotes a probability and $P(\cdot)$ denotes a density distribution.

Instinctively, we might put the threshold θ as in Fig. 1. If both classes are equally important then we put it at the center of those two distribution functions. But sometimes we might put it in different way like in Fig. 2, if we take it into account that the risk of misclassify x into ω_2 while x is ω_2 is important[1].

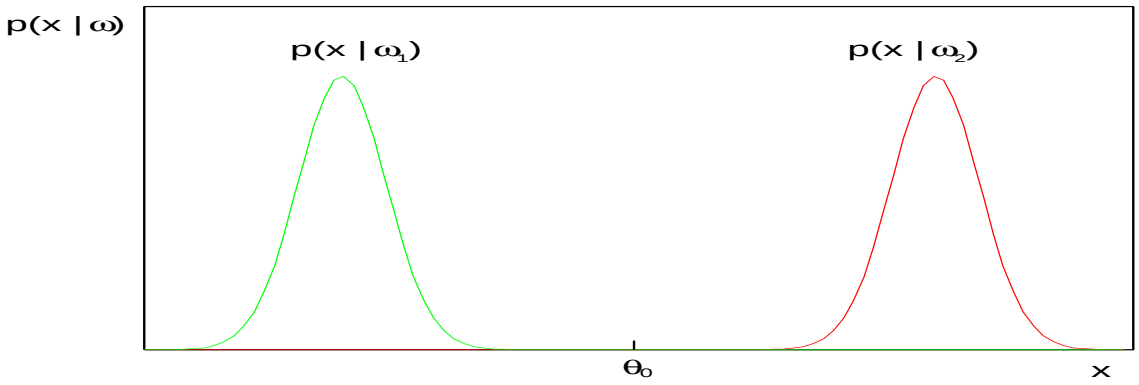

Figure 1: An example of two distribution of the observation x w.r.t. each class of ω_1 and ω_2. Here reader might image that distribution of length of female salmon and male salmon.

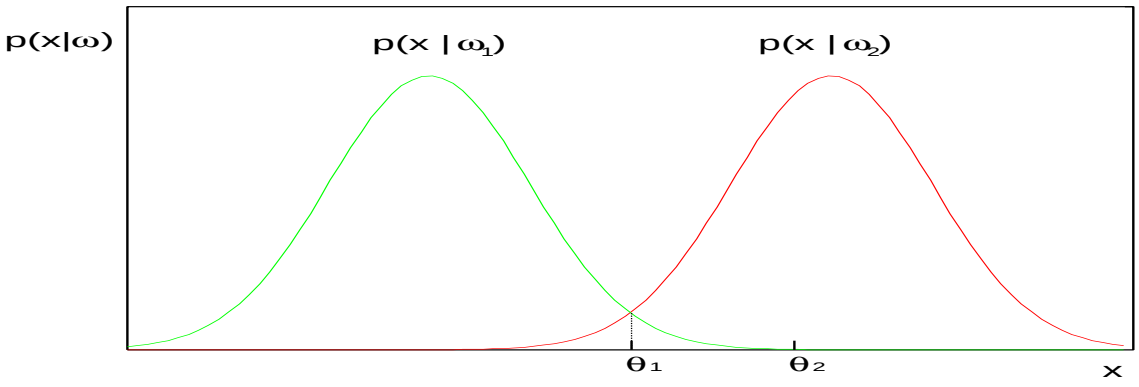

Figure 2: Yet another example where two distributions are closer than the first example. Two threshold are shown, considering a risk of miss-classification.

In the example of Fig. 1.3, you might imagine *a classification for human female and male by one feature – frequency of her/his voice.* In this case $p(x|\omega_1)$ and $p(x|\omega_1)$ would be well separated and easy to classify.

Rule 2 (Classification with posterior probability) *If* $P(x|\omega_1) > P(x|\omega_2)$ *then classify it to* ω_1 *otherwise* ω_2.

[1] In the case of salmon, we tend to be happier if the one we bought a male salmon with cheaper price than female due to eggs, was mistakenly female.

2 Before moving on further

2.1 Example of Bayesian Rule

• Example-1
We have two bags of no difference from its outlook. One bag called R has 70 red balls and 30 blue balls. The other bag called B has 30 red balls and 70 blue balls. When we take one bag at random and pick up 12 balls, returning it to the bag at each time. The result was 8 red balls and 4 blue one. Then was the bag estimated to be R or B, and how probable the estimate is?

• Example-2
In a courtroom trial with a male defendant, CCTV evidence from a pub shows a crime being committed by an individual carrying a bottle, but it is unclear whether this is a male or a female. A witness reports seeing the defendant carrying a bottle around the time of the offense. Suppose that examination of the tape shows 60% of customers were male, and 2% of males and 1% of females carried bottles at any given moment. How strong is the evidence against the defendant?

• Example-3
Suppose the AIDS positive is one in 100. Suppose the test has a false positive rate of 0.2 (that is, 20% of people without HIV will test positive for HIV) and that it has a false negative rate of 0.1 (that is, 10% of people with HIV will test negative). The laws of probability dictate from this last fact that the probability of a positive test given HIV is 90%. Now suppose that you are such a person who has just tested positive. What is the probability that you have HIV?

• Example-4[2]
The legal system is replete with misapplication of probability and with incorrect claims of the irrelevance of probabilistic reasoning as well. In 1964 an interracial couple was convicted of robbery in Los Angeles, largely on the grounds that they matched a highly improbable profile, a profile which fit witness reports. In particular, the two robbers were reported to be A man with a mustache

- Who was black and had a beard
- And a woman with a pony tail
- Who was blonde
- The couple was interracial
- And were driving a yellow car

The prosecution suggested that these characteristics had the following probabilities of being observed at random in the LA area.

- A man with a mustache 1/4

[2]Taken from the book "Bayesian Artificial Intelligence" by Kevin B. Korb & Ann E. Nicholson (2004)

- Who was black and had a beard 1/10
- And a woman with a pony tail 1/10
- Who was blonde 1/3
- The couple was interracial 1/1000
- And were driving a yellow car 1/10

• Example-5
Three prisoners (**A**, **B**, and **C**) are in a prison. **A** knows the fact that the two out of the three are to be executed tomorrow, and the rest becomes free. **A** thought either one of **B** or **C** is sure to be executed. Then, **A** asked a guard "even if you tell me which of **B** and **C** is executed, that will not give me any information as for me. So please tell it to me." The guard answers that **C** will. ⇒ data D Now, **A** knows one of **A** or **B** is sure to be free.

2.2 Multidimensional Gaussian p.d.f.

2.2.1 1-D Gaussian

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\}$$

Exercise 1 *Create 100 points $x_i$which are distributed following 1-D Gaussian in which $\mu = 5$ and $\sigma = 2$.*

2.2.2 2-D Gaussian

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t {\Sigma_i}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\}$$

Exercise 2 *Create 100 points (x_i, y_i) which are distributed following 2-D Gaussian in which ... $\mu_1 = (2.5, 2.5)$ and $\mu_2 = (7.5, 7.5)$ and*

$$\Sigma = \begin{pmatrix} 0.2 & 0.4 \\ 0.7 & 0.3 \end{pmatrix},$$

and

$$\Sigma = \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.1 \end{pmatrix}.$$

3 What will borders look like on what condition?

Now that we restrict our universe in two-dimensional space, we use a notation (x, y) instead of (x_1, x_2). So we now express $\mathbf{x} = (x, y)$. Furthermore, both of our two classes are assumed to follow the Gaussian p.d.f. whose μ are $\boldsymbol{\mu}_1 = (0, 0)$ and $\boldsymbol{\mu}_2 = (1, 0)$, and Σ are

$$\Sigma_1 = \begin{pmatrix} a_1 & 0 \\ 0 & b_1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} a_2 & 0 \\ 0 & b_2 \end{pmatrix}$$

Under this simple condition, our inverse matrix is simply, $|\Sigma_1| = a_1 b_1$ and $|\Sigma_2| = a_2 b_2$. So, we now know

$$\Sigma_1^{-1} = \frac{1}{a_1 b_1} \begin{pmatrix} b_1 & 0 \\ 0 & a_1 \end{pmatrix} = \begin{pmatrix} 1/a_1 & 0 \\ 0 & 1/b_1 \end{pmatrix}$$

and in the same way

$$\Sigma_2^{-1} = \frac{1}{a_2 b_2} \begin{pmatrix} b_2 & 0 \\ 0 & a_2 \end{pmatrix} = \begin{pmatrix} 1/a_2 & 0 \\ 0 & 1/b_2 \end{pmatrix}$$

Now our Gaussian equation is more specifically

$$p(\mathbf{x}|\omega_1) = \frac{1}{2\pi\sqrt{a_1 b_1}} \exp\{-\frac{1}{2}(x \;\; y) \begin{pmatrix} 1/a_1 & 0 \\ 0 & 1/b_1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}\}$$

and

$$p(\mathbf{x}|\omega_2) = \frac{1}{2\pi\sqrt{a_2 b_2}} \exp\{-\frac{1}{2}(x-1 \;\; y) \begin{pmatrix} 1/a_2 & 0 \\ 0 & 1/b_2 \end{pmatrix} \begin{pmatrix} x-1 \\ y \end{pmatrix}\}.$$

Then we can define our discriminant function $g_i(\mathbf{x})$ $(i = 1, 2)$ taking logarithm based natural number e as

$$g_1(\mathbf{x}) = -\frac{1}{2}(x \;\; y) \begin{pmatrix} 1/a_1 & 0 \\ 0 & 1/b_1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \ln(2\pi) + \frac{1}{2}\ln(a_1 b_1)$$

and

$$g_2(\mathbf{x}) = -\frac{1}{2}(x-1 \;\; y) \begin{pmatrix} 1/a_2 & 0 \\ 0 & 1/b_2 \end{pmatrix} \begin{pmatrix} x-1 \\ y \end{pmatrix} + \ln(2\pi) + \frac{1}{2}\ln(a_2 b_2)$$

Neglecting here the common term for both equation $\ln(2\pi)$, our new discriminant functions are

$$g_1(\mathbf{x}) = -\frac{1}{2}\{\frac{x^2}{a_1} + \frac{y^2}{b_1}\} + \frac{1}{2}\ln(a_1 b_1)$$

and

$$g_2(\mathbf{x}) = -\frac{1}{2}\{\frac{(x-1)^2}{a_2} + \frac{y^2}{b_2}\} + \frac{1}{2}\ln(a_2 b_2)$$

Finally, we obtain the border equation from $g_1(\mathbf{x}) - g_2(\mathbf{x}) = 0$.

$$(\frac{1}{a_1} - \frac{1}{a_2})x^2 + \frac{2}{a_2}x + (\frac{1}{b_1} - \frac{1}{b_2})y^2 = \frac{1}{a_2} + \ln\frac{a_1 b_1}{a_2 b_2} \tag{1}$$

We now know that the shape of the border will be either of the following five cases: (i) straight line (ii) circle; (iii) ellipse; (iv) parabola; (v) hyperbola; (Vi) two straight lines, depending on how the points distribute, that is, depending on a_1, b_1, b_1 and b_2 in our situation above.

3.1 Examples

Let's try following calculations,

$$(1) \quad \Sigma_1 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.10 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.10 \end{pmatrix}$$

$$(2) \quad \Sigma_1 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.10 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.20 & 0 \\ 0 & 0.20 \end{pmatrix}$$

$$(3) \quad \Sigma_1 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.15 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.20 & 0 \\ 0 & 0.25 \end{pmatrix}$$

$$(4) \quad \Sigma_1 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.15 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.10 \end{pmatrix}$$

$$(5) \quad \Sigma_1 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.20 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.10 \end{pmatrix}$$

The next example is somewhat tricky. I wanted an example in which the right-hand side of the equation (6) becomes zero and the left-hand side is a product of one-order equations of x and y. As you might know, this is the case where border equation will be made up of two straight lines.

$$(6) \quad \Sigma_1 = \begin{pmatrix} 2e & 0 \\ 0 & 0.5 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

My quick calculation tentatively results in as follows. See also the Figure below.

(1) $2x = 1$
(2) $5(x+1)^2 + 5y^2 = 10 - \ln 4$
(3) $5(x+1)^2 + (8/3)y^2 = 10 - \ln(10/3)$
(4) $5(x+1)^2 - (10/3)y^2 = 10$
(5) $20x - 5y^2 = 10 - \ln 2$
(6) $(1 - 1/2e)x^2 - x - y^2 = 0$

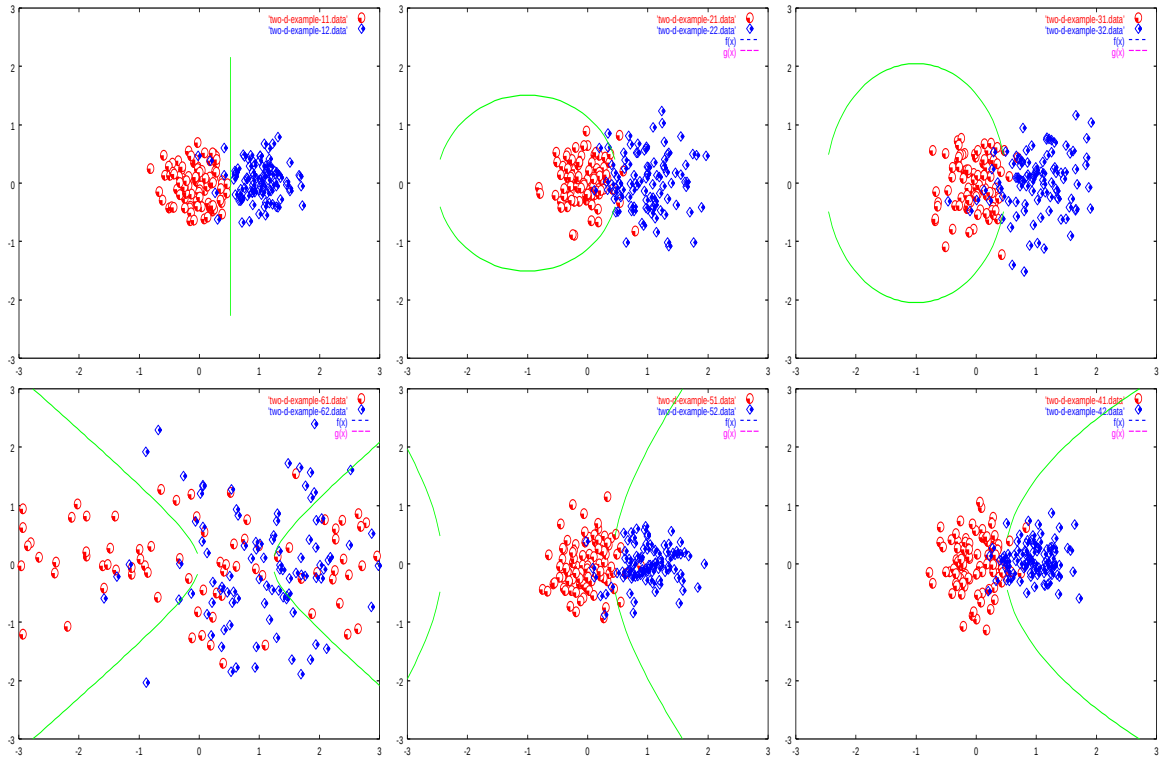

Figure 3: A cloud of 100 points each extracted from a set of two classes and border of the two classes calculated on six different conditions. (Results of (5) and (6) are still fishy and under another trial.)

3.2 3-D Gaussian case

• **When all** $\Sigma_i = \Sigma$

Here we study only one example. We assume two classes where $P(\omega_1) = P(\omega_1) = 1/2$. In each class, the patterns are distributed with Gaussian p.d.f both have the same covariance matrix

$$\Sigma_i = \Sigma_j = \Sigma = \begin{pmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.3 & -0.1 \\ 0.1 & -0.1 & 0.3 \end{pmatrix}$$

and means of the distribution are $(0, 0, 0)^T$ and $(1, 1, 1)^T$. We now take a look at what our discriminant function

$$g_i(\mathbf{x}) \quad = \quad -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \tag{2}$$

leads to?

Since we calculate (See APPENDIX for detail)

$$\Sigma^{-1} = \frac{1}{30} \begin{pmatrix} 5 & -3 & -3 \\ -2 & 6 & 3 \\ -1 & 3 & 6 \end{pmatrix}$$

Now our discriminant equation $g_1(\mathbf{x}) = g_2(\mathbf{x})$ is

$$(x_1 x_2 x_3) \begin{pmatrix} 5 & -3 & -3 \\ -2 & 6 & 3 \\ -1 & 3 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} =$$

$$((x_1 - 1)(x_2 - 1)(x_3 - 1)) \begin{pmatrix} 5 & -3 & -3 \\ -2 & 6 & 3 \\ -1 & 3 & 6 \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \\ x_3 - 1 \end{pmatrix}$$

Further calculation leads to

$$((5x_1 - 2x_2 - x_3)(-3x_1 + 6x_2 + 3x_3)) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} =$$

$$((5x_1 - 2x_2 - x_3 - 2)(-3x_1 + 6x_2 + 3x_3 - 6)(-3x_1 + 3x_2 + 6x_3 - 6)) \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \\ x_3 - 1 \end{pmatrix}$$

All the 2nd-order terms are canceled and we obtain,

$$7x_1 + 13x_2 - 20x_3 = 14$$

We now know that it is the plane which discriminates two of these classes ω_1 and ω_2.

3.3 A Higher order Gaussian case

The Equation

$$g_i(\mathbf{x}) \quad = \quad \ln(p(\mathbf{x}|\omega_i) + \ln P(\omega_i)) \tag{3}$$

still holds, of course. Now let's recall that the Gaussian p.d.f. is

$$p(\mathbf{x}|\omega) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\} \tag{4}$$

and as such

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}|\Sigma_i| + \ln P(\omega_i) \tag{5}$$

We know take a look at cases which simplify situation more or less.

• **When** $\Sigma_i = \sigma^2 I$

In this case, it's easy to guess samples fall in equal diameter hyper-shapers. Note, first of all $|\Sigma_i| = \sigma^{2d}$ and $\Sigma_i^{-1} = (1/\sigma^2)I$. So, we assume $g_i(\mathbf{x})$ here to be

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i) \tag{6}$$

or, equivalently

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i) + \ln P(\omega_i) \tag{7}$$

Neglecting the terms those no affecting to the relation $g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$ our $g_i(\mathbf{x})$ is now

$$g_i(\mathbf{x}) = \frac{1}{\sigma^2}\boldsymbol{\mu}_i^t\mathbf{x} - \frac{1}{2\sigma_i^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i) \tag{8}$$

Then $g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$ leads to

$$\frac{1}{\sigma^2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i)^t\mathbf{x} - \frac{1}{2\sigma_i^2}(\|\boldsymbol{\mu}_i\|^2 - \|\boldsymbol{\mu}_j\|^2) + \ln\frac{P(\omega_i)}{P(\omega_j)} = 0 \tag{9}$$

If we carefully modify Eq. (9) we will obtain

$$\mathbf{W}\cdot(\mathbf{x}-\mathbf{x}_0) = 0. \tag{10}$$

In this case our classification rule will be

Rule 3 (Minimum Distance Classification) *Measure Euclidean distance* $\|\mathbf{x}-\boldsymbol{\mu}\|$ *for* $\forall i$*, then classify* $\mathbf{x}$ *to the class whose mean is nearest to* $\mathbf{x}$*.*

Exercise 3 *Derive the Eq. (10) specifying* $\mathbf{W}$ *and* $\mathbf{x}_0$*.*

Eq. (10) is the equation which can be interpret as

"A hyperplane through $\mathbf{x}_0$ *perpendicular to* $\mathbf{W}$*."*

- **When all** $\Sigma_i = \Sigma$

This condition implies that the patterns in each of both classes distribute like hyper-ellipsoid. Now that our discriminant function is

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

We again obtain

$$\mathbf{W}\cdot(\mathbf{x}-\mathbf{x}_0) = 0$$

where $\mathbf{W}$ and $\mathbf{x}_0$ are

$$\mathbf{W} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{11}$$

and

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln P(\omega_i)/P(\omega_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \tag{12}$$

Notice here that $\mathbf{W}$ is no more perpendicular to the direction between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$.

Exercise 4 *Derive* $\mathbf{w}$ *and* $\mathbf{x}_0$ *above.*

So we modify the above rule to

Rule 4 (Classification by Mahalanobis distance) *Assign* $\mathbf{x}$ *to* ω_i *in which Mahalanobis distance from* $\boldsymbol{\mu}_i$ *is minimum for* $\forall i$*.*

Yes! This *Mahalanobis distance* between $\mathbf{a}$ and $\mathbf{b}$ is defined as

$$(\mathbf{a}-\mathbf{b})^t\Sigma^{-1}(\mathbf{a}-\mathbf{b}) \tag{13}$$

The final example in this sub-section is more general 2-dimensional case, but (artificially) devised so that calculations won't become very complicated. We now assume $\boldsymbol{\mu}_1 = (0,0)$ and $\boldsymbol{\mu}_2 = (1,0)$, and we both classes share the same Σ:

(7) $\Sigma_1 = \begin{pmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{pmatrix}$

• **When all Σ_i's are arbitrary**

When no such restriction as above to simplify situation, the discriminant function is

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + lnP(\omega_i)$$

Only the term we can neglect now is $(d/2)\ln 2\pi$. We now apply the identity

$$(\mathbf{x}-\mathbf{y})^t A(\mathbf{x}-\mathbf{y}) = \mathbf{x}^t A\mathbf{x} - 2(A\mathbf{y})^t\mathbf{x} + \mathbf{y}^t A\mathbf{y}.$$

Then, we get the following renewed discriminant function

$$g_i(\mathbf{x}) = \mathbf{x}^t W_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + w_{i0} \tag{14}$$

where

$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$\mathbf{w}_i = \Sigma_i^{-1}\boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i).$$

Hence, $g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$ leads us to a *hyper quadratic form.* Or, if you want, we can express it as

$$(a_1x_1 + a_2x_2 + \cdots + a_nx_n)(b_1x_1 + b_2x_2 + \cdots + b_nx_n) = const.$$

Namely, the border is either of (i) Hyper-planes; (ii) a pair of hyper-planes; (iii) hyper-sphere; (iv) hyper-ellipsoid; (v) hyper paraboloid; (vi) hyper-hyperboloid.

4 Parameter Estimation

Let's go back to one dimensional case, that is, p.d.f is defined on 1-D variable x. Now we assume p.d.f is not only Gaussian but also many others. Then the goal here is to know the form of $p(x|\omega_i)$ with its parameters included. We also assume that we know which of the p.d.f. among others for some reason to believe. What we don't know are the parameters of the density function.

We now name a few of such p.d.f.s other than Gaussian p.d.f. all of which need to be specified by some parameters.

- Discrete x
  - Binomial
    - $f(x) = (m!/x!(m-x)! \cdot \theta^x(1-\theta)^{m-x}$
  - Poisson
    - $f(x) = (\theta^x/x!)\exp(-\theta)$
- Continuous x
  - Uniform
    - $p(x) = 1/\theta$ ... if $x > 0$ otherwise 0
  - Exponential
    - $f(x) = \theta\exp(-\theta x)$ ... if $x > 0$ otherwise 0
  - Rayleigh
    - $f(x) = 2\theta x\exp(-\theta x^2)$ ... if $x > 0$ otherwise 0

4.1 Overviews of each p.d.f.

First, for a function $f(x)$ to be a p.d.f. it should fulfill

(1) $f(x) \geq 0$ for $\forall x$

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$ or $\sum f(x_i) = 1$ for discrete variable x

Then the probability of $a \leq x \leq b$ is

$$p(a \leq x \leq b) = \int_a^b f(x)dx$$

for continuous x.

4.1.1 Binomial

Imagine an experiment that has only two possible outcomes. We now suppose a probability that a particular event will occur is p, and as such a probability that the event will not occur $(1 - p)$. An example of those two events are 'Success' or 'Failure.' The p.d.f. models the probability that we will observe r successes and $n - r$ failures in a total of n-trials.

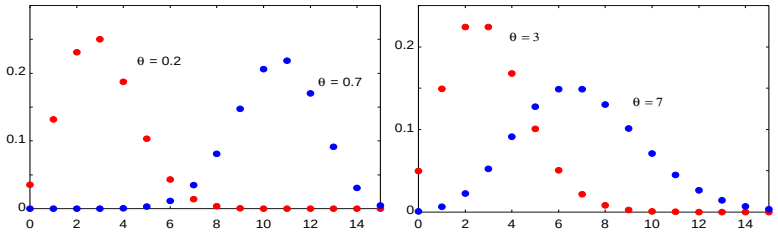

Figure 4: Binomial p.d.f. (left) and Poisson p.d.f. (right).

4.1.2 Poisson

Poisson distribution expresses the probability of how often the events we concern will occur in a fixed interval of time or space if these events occur with a known average rate and independently of the time since the last event.

This is the limit of the binomial distribution as the number of possible events approaches infinity while the probability of any specific event approaches zero, maintaining the correct average number of events in a specific interval.

4.1.3 Uniform

When the distribution is uniform in the region $a \leq x \leq b$, the p.d.f. is $f(x) = 1/(b - a)$. For example, any result of fair roulette should be uniformly distributed.

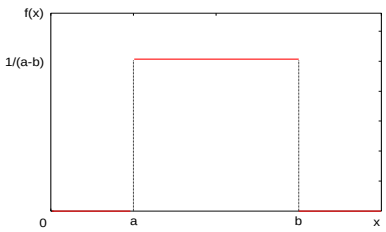

Figure 5: Yet another example where two distributions are closer than the first example. Two threshold are shown, considering a risk of miss-classification.

4.1.4 Exponential

This models events that occur continuously and independently at a constant average rate. For example, it is used to model a behavior with a constant failure rate. In the above equation, θ is a constant failure rate per unit mesurement, e.g. per hour, while x is time, e.g., hour.

It might be easy to understand if you think of its cumulative distribution function $F(x) = 1 - f(x)$. [3]

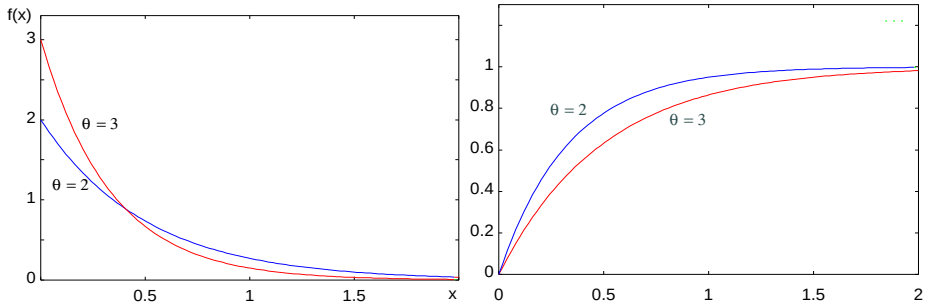

Figure 6: Exponential p.d.f. (left) and its cumulative distribution function (right).

4.1.5 Rayleigh

One example where the Rayleigh distribution naturally arises is when wind speed is analyzed into its orthogonal 2-dimensional vector components. Assuming that the magnitude of each component is uncorrelated and normally distributed with equal variance, then the overall wind speed (vector magnitude) will be characterized by a Rayleigh distribution.

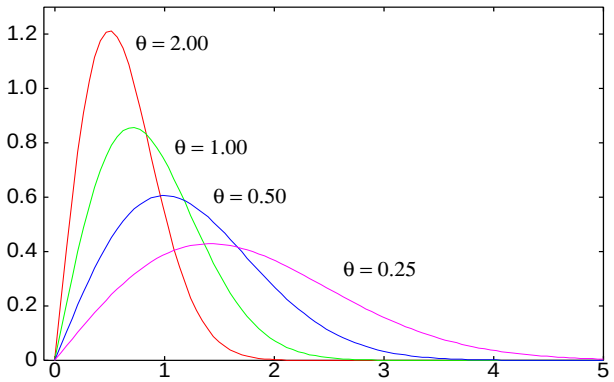

Figure 7: Rayleigh distribution function with four different values of θ.

[3] Cumulative distribution function expresses the probability that a random variable X is less than x for any x. Hence, $p(X > x) = 1 - F(x)$ because $p(X \leq x)$ and $p(X > x)$ are mutually exclusive, and as such, $p(X \leq x) + p(X > x) = 1$.

4.2 Maximum Likelihood Estimation

Before entering this topic, try the following case where we have two classes each of which has only 4 training samples.

Example 1 *Assuming in a 2-dimensional space, what if we have a pair of 4 samples from each of two classes? The patterns we have are* $(8,3),(4,3),(6,2),(6,4)$ *from* ω_1 *and* $(0,3),(-2,1),(-4,3),(-2,5)$ *from* ω_2*. Try to guess the border between two classes, and then classify a new point* $(5,3)$*, for example.*

We now further more simplify the situation.

Example 2 *Our data is now* $x_1 = 3$*,* $x_2 = 8$*,* $x_3 = 2$ *and* $x_4 = 5$*. Guess what distribution function of these 4 data follows?*

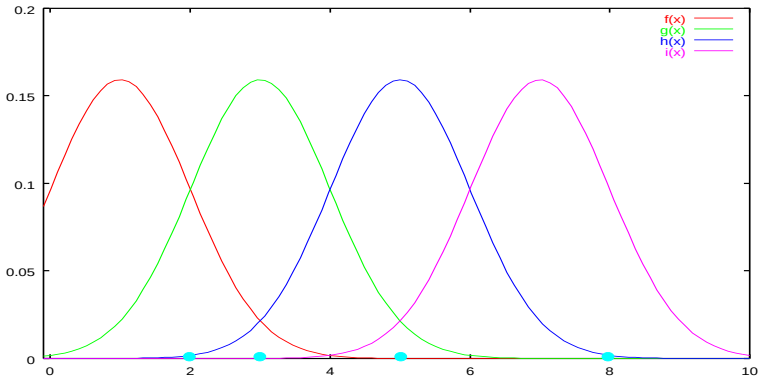

Figure 8: Assuming Gaussian which one is more likely?

This might be a Gaussian distribution but just a brief look at it suggests it more like a Rayleigh distribution.

Exercise 5 *(1) We now assume Gaussian with* $\sigma = 1$ *with* μ *being unknown. Then estimate* μ *from the same 4 data, i.e.* $x_1 = 3$*,* $x_2 = 8$*,* $x_3 = 2$ *and* $x_4 = 5$*, in the same way as above. (2) Next, create 10, 20, 100 data randomly from* $N(1,3)$*, and plot* $p(D|\mu)$ *as a function of* μ *in each of the three cases.*

Rule 5 (An assumption of independence of data) *Choose θ which maximizes*

$$P(D|\theta) = \prod_{k=1}^{n} P(x_k|\theta) \tag{15}$$

Then let's calculate our previous example of one-dimensional four data $D = \{2, 3, 5, 8\}$ by changing θ from 0.00 to 1.00 with an interval of 0.01. The results are shown the left-most graph of the Fig. 4.2. It might be interesting what if we have more data to estimate. The same procedures are made with the number of data 10, 20, 100 and the results are shown in the same Figure.

All we don't know is its parameters to specify the function. But we have an information of prior probability $p(\theta)$ and training sample give us $p(\theta|D)$ which we expect to have a sharp peak at the true θ.

Important thing is is

"The more data we have, the sharper the graph becomes."

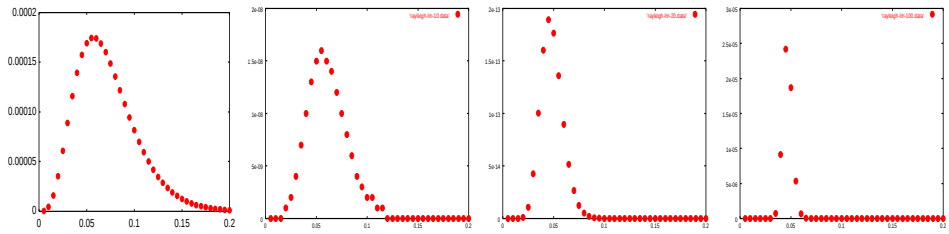

Figure 9: An example of estimation when the number of sumple is 4, 10, 20, 100

5 Density function estimation – Parzen Window Method

We have so far assumed that the distributions is known in advance (e.g. Gaussian, Rayleigh etc.), So all we have to estimate are the parameters. However, many scenarios in reality, the distributions is unknown. Then what we need is to estimate the distribution. This is sometimes called non-parametric estimation. We now explore one of such method known as Parzen Windows.

Once we know the distribution with this method (or others), the parameters are again estimated by the maximal likelihood method, numerically this time.

Here we study only distribution function defined on one-dimensional space.

Given a data sample $x_1, x_2, ..., x_n,$, we use each data, one by one, to estimate distribution function.

Parzen-windowing essentially superposes kernel functions placed at each data (as a window). We now suppose we are estimating the value of the distribution function $p(x)$ at point X. Then we place a window function at X and determine how many other points fall within the window. Thus, the value is the sum of the contributions from those points to this window. To be more specific

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x - x_i}{h})$$

where $K(x)$ is the window function called kernel, and h is the window width usually chosen based on the number of available data n.

The Gaussian is a popular kernel for Parzen-window density estimation.

$$K(x) = (1/\sqrt{2\pi}\exp(-x^2/2)$$

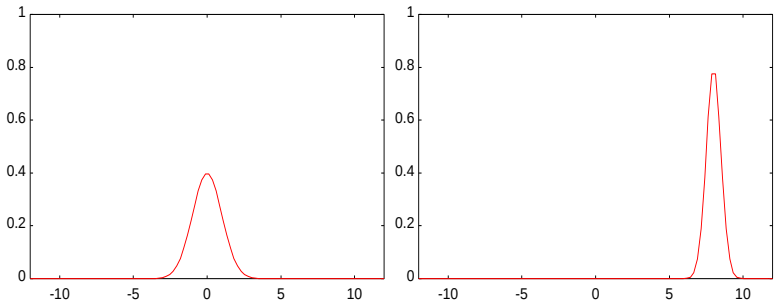

Figure 10: Gaussain kernel functions and its transposition.

Then

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{x-x_i}{h})^2)$$

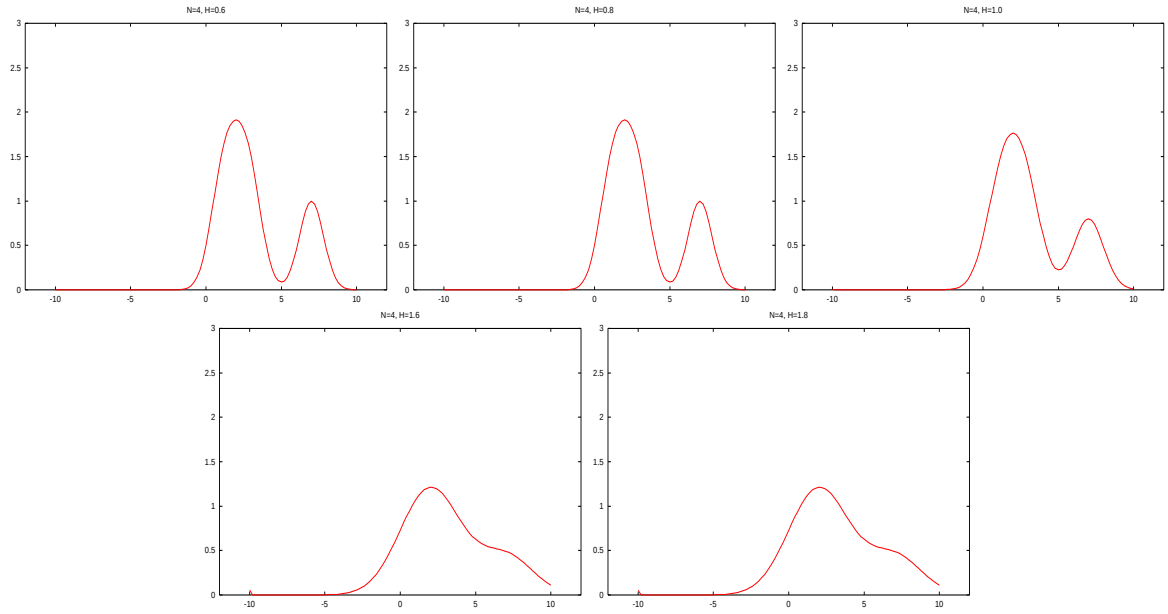

Figure 11: An example of application of Rayleigh kernel to a date with 4 points.

We now try to apply Gaussian kernel function to the following data set of 100 points.

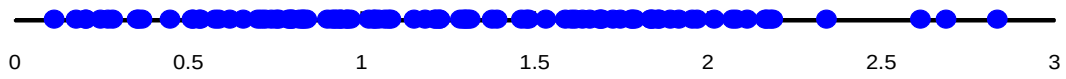

Figure 12: Randomly created 100 one-dimensional points following the Rayleigh p.d.f.

Numerical of these points from Rayleigh p.d.f. with θ being 3 are:

0.28 3.12 3.82 3.23 0.70 0.20 2.71 2.63 2.18 1.15 0.35 3.49 0.11 0.10 2.68 0.80 0.59 2.41 2.94
2.34 1.98 2.72 3.58 0.38 1.54 2.67 2.07 0.62 3.62 1.85 2.96 2.32 3.59 1.94 1.04 3.26 0.51 2.74
3.94 2.20 0.74 0.19 2.02 4.00 2.95 0.75 0.53 2.66 0.20 2.57 0.27 3.76 1.09 0.94 0.53 2.76 0.06
2.94 0.53 4.04 4.03 1.75 2.88 2.24 3.81 3.59 3.12 0.11 2.59 3.44 1.18 0.09 0.90 0.72 0.33 3.09
4.04 1.31 0.18 2.17 3.31 1.48 1.78 3.43 0.78 0.71 1.30 2.99 1.96 3.38 1.73 0.71 1.78 1.92 0.92
1.59 2.32 0.91 4.01 2.26

Also random point from Gaussian distribution with μ and σ being .. and .. respectively.

0.70 0.20 1.15 0.80 0.59 0.62 1.04 0.51 0.74 1.09 0.94 0.53 1.18 0.90 1.31 0.71 1.30 0.92 0.91
1.46 0.83 1.28 1.48 0.96 0.58 0.82 0.79 1.07 1.38 1.03 1.61 1.21 1.69 1.30 0.66 0.83 1.05 0.45

1.30 0.94 0.79 0.80 1.39 0.84 0.97 1.02 1.22 1.08 0.93 1.53 0.96 1.25 2.03 0.88 0.07 1.05 1.40
1.28 1.27 0.02 1.05 0.99 0.91 0.97 0.39 1.15 1.09 0.54 0.84 1.25 1.26 0.94 0.86 0.89 1.43 0.73
1.79 0.90 0.81 1.39 1.01 0.62 1.29 0.78 1.40 1.17 1.34 1.28 0.64 0.57 0.69 1.33 1.19 0.99 0.87
1.08 0.75 0.60 0.23 1.49

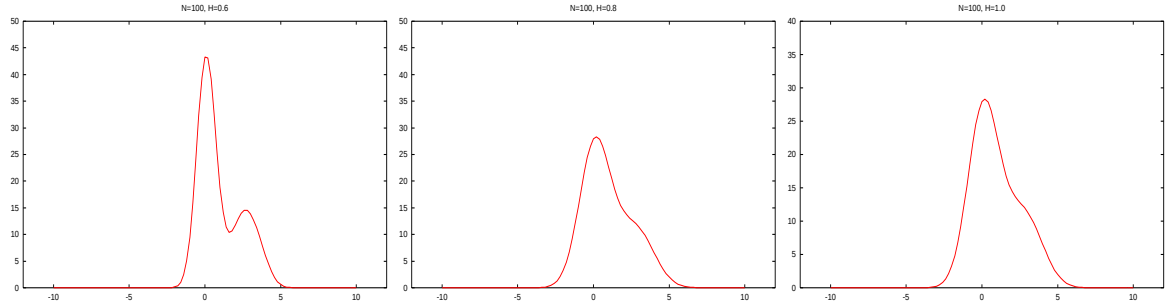

Figure 13: An example of application of Gaussian kernel to a date with 100 points.

Or, more simply, we may use uniform distribution $K(x) = 1/2$ for $|x| \leq h$ otherwise $K(x) = 0$, which results in histogram as a resultant probability distribution function.

Exercise 6 *(1) Assume we have only 4 sample data $x_1 = 3$, $x_2 = 8$, $x_3 = 2$ and $x_4 = 5$, apply Parzen method with its kerner being uniform p.d.f. (2) Then try with Gaussian p.d.f.*

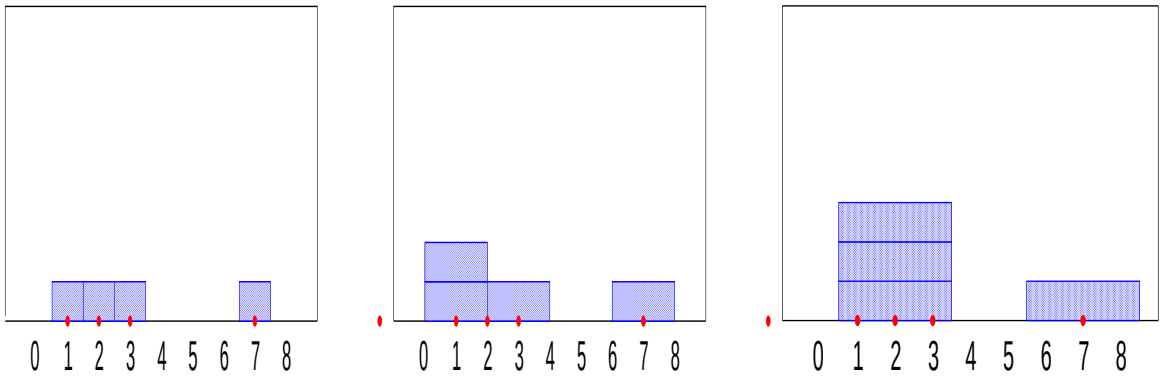

Figure 14: With kernel being a uniform p.d.f.

Exercise 7 *Create 100 points distributed Rayleigh p.d.f., and then apply Parzen with Gaussian p.d.f.*

5.1 Kernel functions

A kernel function is a non-negative, real-valued, and integrable function satisfying the

- $\int_{-\infty}^{\infty} K(x)dx = 1$.
- $K(-x) = K(x)$ for $\forall x$.

5.1.1 Other kernel function commonly used

The following seven functions are all defined in the region $|x| \leq 1$.

- Triangular
  - $K(x) = 1 - |x|$
- Epanechnikov
  - $K(x) = (3/4)(1 - x^2)$
- Biweight (or Quartic)
  - $K(x) = (15/16)(1 - x^2)^2$
- Triweight
  - $K(x) = (35/32)(1 - x^2)^3$
- Tricube
  - $K(x) = (70/81)(1 - |x|^3)^3$
- Cosine
  - $K(x) = (\pi/4)\cos(\pi x/2)$

5.1.2 Why are we happy with this?

Thus we can know the function form $p(x|\omega_i)$ for each family ω_i, and we can use our formula .

Rule (Classification with posterior probability) If $P(x|\omega_1) > P(x|\omega_2)$ then classify it to ω_1 otherwise ω_2.

Thus far, we are again happy to be able to perform Bayesian classification as described earlier.

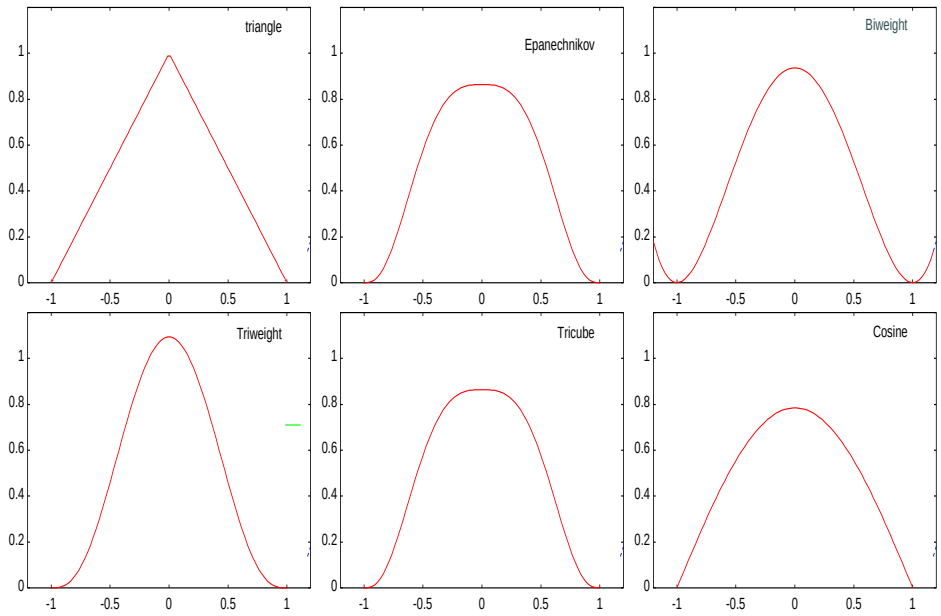

Figure 15: Six different kernel functions.

PART II
BAYESIAN NETWORK

1. What is Bayesian Network?
2. Inferring from Bayesian Network.
3. Learning of Bayesian Network.

6 Bayesian Network

So far our conditional probabilities are sometimes probability distribution function, such as $p(x|\omega)$, not a numerical value of probability. From now on, all the notations will be numerical value of the probability of some event. E.g. $p(A|B)$ means the probability of A under the condition of B, or equivalently, the probability of A given B. Specifically we call them (A and B here) *variables*.

Then Bayesian network is a graph which represent dependence of these variables. Nodes represent these variables, and arcs represent the probability of these dependencies. That is, the arc from node A to B is $p(B|A)$.

The objective of the Bayesian network is to infer a probability of some variable whose probability is unknown from the information of a set of value of the other variables. The former variable is called *hypothesis* and the latter are called *evidences*. Hence we may say this objective is to:

> infer the probability of hypotheses from evidences given.

So our most frequent notation of a probability will be described $p(h|e)$. Sometimes this probability is called *belief* and also described as

$$Bel(h|e).$$

To simply put, this is, "how much is our belief for the hypothesis given those evidences."

6.1 Examples

6.1.1 Flu & Temperature

This example is taken from Korb et al.[4]

Flu causes a high temperature by and large. We now suppose the probability that we are flu is $p(\text{Flu}) = 0.05$, the probability that we have High-temperature when we are flu is $p(\text{High-temperature}|\text{Flu}) = 0.9$ and the probability of we still have a high temperature even when we are not flu (false alarm) is $p(\text{High-temperature}|\neg\text{Flu}) = 0.2$. See Figure 16.

Exercise 8 *Now we assume to have an evidence that one guy has a high-temperature, then how much is a belief of this guy is Flu?*

Or conversely,

Exercise 9 *We have an evidence that one guy is flu then how much is a belief of this guy has a high-temperature?*

[4]K. B. Korb and A. E. Nicholson (2003) "Bayesian Artificial Intelligence."

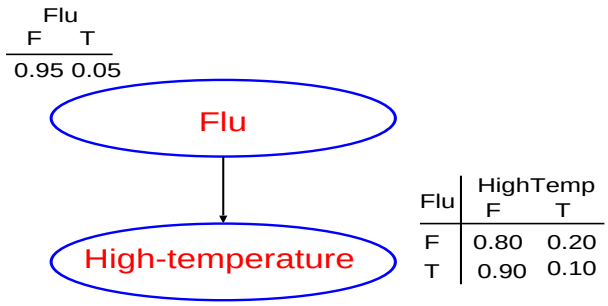

Figure 16: Flu causes high-temperature. Redrawn from Korb et al. (Sorry but without permission.)

6.1.2 Season & Rain

In the example of the previous subsection, all the variables take a binary value. Sometimes we want variables which takes more cases. Here we have such an example. Again a simple example of two variables but one is about season which takes 4 values: {winter, spring, summer and autumn}, and weather which takes also 4 values: {fine, cloudy, rain, and snow} See Figure 17.

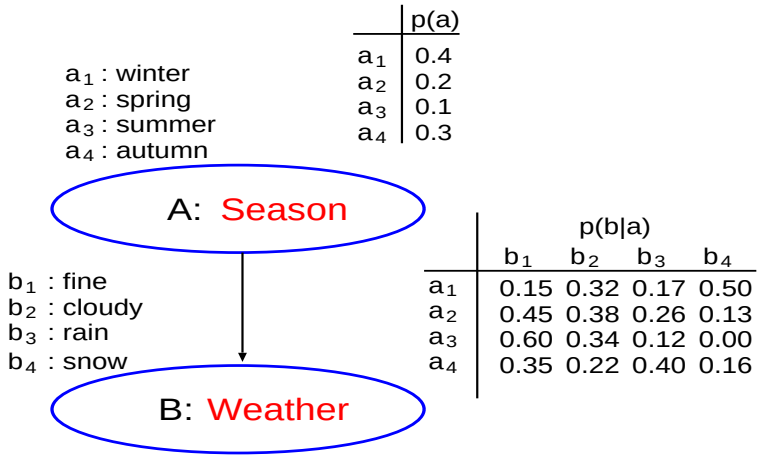

Figure 17: Season & Weather

Now try the following two inferences. The first one is very direct.

Exercise 10 *Assume it's now Summer (evidence), then how much is the probability that it's rain?*

The next one is not such straight forward but still quite easy.

Exercise 11 *Now it's snow, then how much is the probability of being autumn now?*

6.1.3 Flu, Temperature & Thermometer

Now we move on to a case of three variables. First one is simple enough like $A \to B \to C$. Let's call A "parent of B," and C "child of B." This example is again taken from Korb et al.

Relation of Flu & High Temperature is the same as before. Now we have a thermometer whose rate of false negative reading is 5% and false positive reading is 14 %, that is,

$$p(\text{HighTherm} = True|\text{HighTemp} = True) = 0.95$$
$$p(\text{HighTherm} = True|\text{HighTemp} = False) = 0.15$$

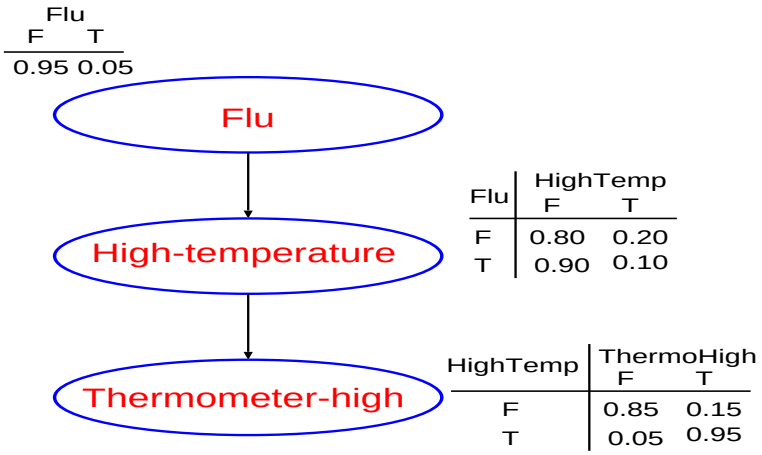

Figure 18: Flu → HighTemp → ThermoHigh. Redrawn from Korb et al. (Sorry but without permission.)

Exercise 12 *Evidence now is, he is Flu and Thermometer suggests HighTemp, then how much is the probability of hypothesis that he has a High temperature?*

Or, lack of one evidence

Exercise 13 *Now thermometer suggests HighTmp, then how much is the probability of his being Flu?*

6.1.4 Grass are soaked then it's rain or sprinkler

Now we proceed a more complicated case in which dependency is not linear. This example is taken from Wikipedia.[5]
The situation is described in the page as:

[5] at *http : //en.wikipedia.org/wiki/Bayesian_network*

Suppose that there are two events which could cause grass to be wet: either the sprinkler is on or it's raining. Also, suppose that the rain has a direct effect on the use of the sprinkler (namely that when it rains, the sprinkler is usually not turned on). All three variables have two possible values, T for true and F for false.

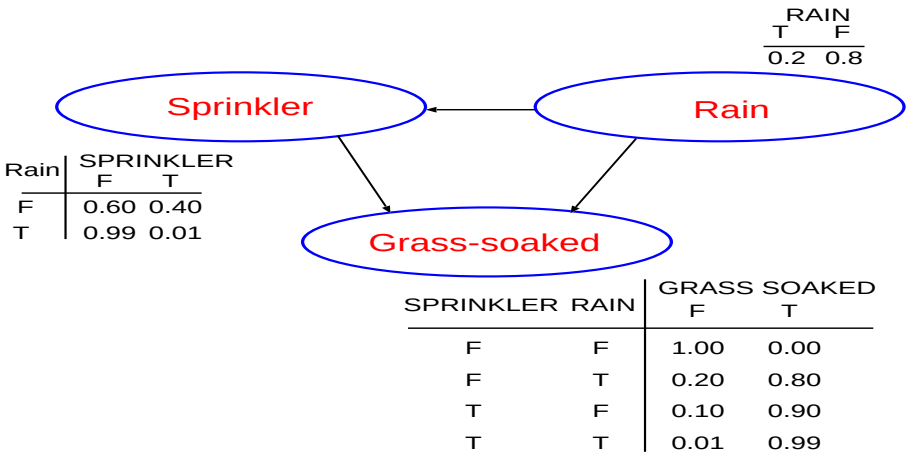

Figure 19: Grass are soaked because it's rain and/or sprinkler?

Let's calculate the joint probability function. First, recall that the joint probability of A and B, in general, can be expressed as:

Rule 6 (Joint probability) $p(A, B) = p(A|B)p(B)$.

We can extend this formula as:

Rule 7 (Extended joint probability) $p(A, B, C) = p(A|B, C)p(B, C) = p(A|B, C)p(B|C)p(C)$

We now regard $A = G$, and $B = (S, R)$. Then,

$$P(G, S, R) = P(G|S, R)P(S|R)P(R)$$

Notice here G means $G = True$ unless otherwise mentioned. Also S means $S = True$ and R means $R = True$.

Let us now denote probability of $G = True$, $S = False$ and $R = True$, for example, as $p(G, S, R)_{TFT}$.

Then we can assume

$$P(S,R) = p(G,S,R)_{TFT} + p(G,S,R)_{TTT}$$

because G is either $True$ or $False$. And similarly,

$$P(R) = p(G,S,R)_{TFT} + p(G,S,R)_{TTT} + p(G,S,R)_{TFT} + p(G,S,R)_{TTT}.$$

For example,

$$p(G,S,R)_{TFT} = 0.80 \times 0.99 \times 0.20 = 0.1584.$$

Also recall the basic formula

$$p(X|Y) = \frac{p(X,Y)}{p(Y)}.$$

And

$$p(X,Y) = \sum_Z p(X|Z,Y)p(Z|Y)$$

Thus, the probability of it's rain under the observation of the grasses are wet is:

$$\begin{aligned} p(R|G) &= \frac{p(G,R)}{p(G)} \\ &= \frac{p(G,S,R)_{TFT} + p(G,S,R)_{TTT}}{p(G,S,R)_{TFF} + p(G,S,R)_{TFT} + p(G,S,R)_{TTF} + p(G,S,R)_{TTT}} \\ &= \frac{0.1584 + 0.00198}{0 + 0.1584 + 0.288 + 0.00198} \\ &\sim 0.3577. \end{aligned} \tag{16}$$

Exercise 14 *What is the probability that sprinkler is working, given the grass is wet?*

Exercise 15 *If, on the other hand, we wish to answer an interventional question: "What is the likelihood that it would rain, given that we wet the grass?"*

6.1.5 My family now goes out?

This example is taken from Charniak.[6]

[6] E. Charniak (1991) "Bayesian Networks without Tears." AI MAGAZINE Vol. 12 No. 4, pp. 50-63.

Suppose when I go home at night, I want to know if my family is home before I try the doors. (Perhaps the most convenient door to enter is double locked when nobody is home.) Now, often when my wife leaves the house, she turns on an outdoor light. However, she sometimes turns on this light if she is expecting a guest. Also, we have a dog. When nobody is home, the dog is put in the back yard. The same is true if the dog has bowel troubles. Finally, if the dog is in the backyard, I will probably hear her barking (or what I think is her barking), but sometimes I can be confused by other dogs barking.

Exercise 16 *Try to draw the Bayesian network.*

6.1.6 Pearl's earthquake Bayesian Network

This is a very popular example to show how Bayesian network looks like by Pearl.[7]

You have a new burglar alarm installed. It reliably detects burglary, but also responds to minor earthquakes. Two neighbors, John and Mary, promise to call the police when they hear the alarm. John always calls when he hears the alarm, but sometimes confuses the alarm with the phone ringing and calls then also. On the other hand, Mary likes loud music and sometimes doesn't hear the alarm. Given evidence about who has and hasn't called, you'd like to estimate the probability of a burglary.

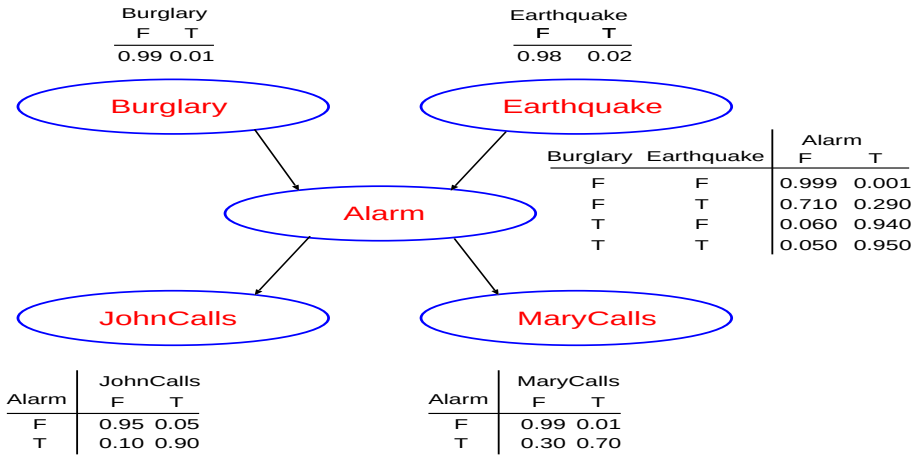

Figure 20: Pearl's Earthquake

[7]Pearl, J. (1988) "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference." San Mateo, Morgan Kaufmann.

7 Inference

Inference task is collecting evidence and updating belief.

7.1 Flu → HighTemp → HighThermo revisit

We now take a look at it by revisiting the simplest example we saw in the previous section.

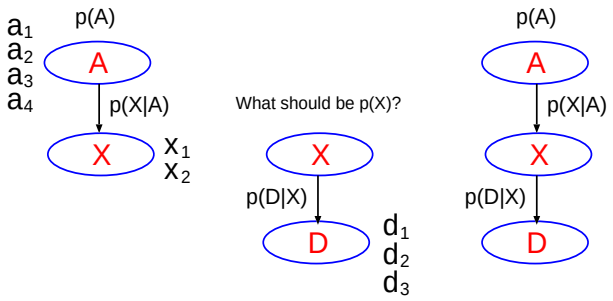

Figure 21: A simple Bayesian Network A → X → D

Assume now we want to know the probability of hypothesis HighTemp with the evidence HighThermo. Let us now denote it as $A \rightarrow X \rightarrow D$. Then we can calculate it as $p(X|D) \propto p(D|X)p(X)$. We will notice that $p(X)$ is not explicitly known. But its implication is given by $p(X|A)p(A)$. So, we can conclude that

$$p(x|A, D) = p(D|X)p(X|A)p(A).$$

7.2 With a little more extension

We now take an example where target variable X has three parent variables.

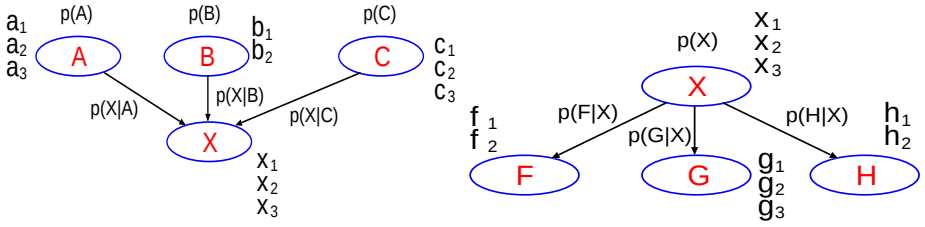

Figure 22: The variable of our concern X has three parent variables A, B, and C as well as three child variables F, G and H.

$$\begin{aligned} p(x|\mathbf{e}^P) &= p(x|e_A, e_B, e_C) \\ &= \sum_{\text{all possible } i,j,k} p(x|a_i, b_j, c_k) p(a_i, b_j, c_k|e_A, e_B, e_C) \\ &= \sum_{\text{all possible } i,j,k} p(x|a_i, b_j, c_k) p(a_i|e_A) p(b_j|e_B) p(c_k|e_C) \end{aligned}$$

$$p(\mathbf{e}^C|x) = p(e_A, e_B, e_C|x) = p(e_A|x)p(e_B|x)p(e_C|x)$$

So, by combining the above two, we obtain

$$\begin{aligned} & p(x|e_A, e_B, e_C, e_F, e_G, e_H) \\ = \; & p(x|\mathbf{e}^P)p(\mathbf{e}^C|x) \\ = \; & p(e_A|x)p(e_B|x)p(e_C|x) \cdot \sum_{i,j,k} p(x|a_i, b_j, c_k) p(a_i|e_A) p(b_j|e_B) p(c_k|e_C) \end{aligned}$$

Here we must notice that we used the following formula:

$$p(X|Y) = \sum_Z p(X|Z, Y)p(Z|Y)$$

7.3 Salmon or Sea-bass?

The example of this subsection including three exercises is totally taken, with minor modifications, from Duda et al. [8]

Try to think of the following Bayesian network shown in Figure 23

First of all, let's apply the formula described in the previous subsection. Assume now hypothesis is a value of X, and evidence is one of the values of each variable A, B, C and D.

So, $p(x|\text{four evidences})$ is

$$p(x|\mathbf{e}^P)p(\mathbf{e}^C|x)$$

where parent evidences $\mathbf{e}^P$ are now e_A and e_B, while child evidences $\mathbf{e}^C$ are e_C and e_D.

While $p(\mathbf{e}^C|x)$ will be simply calculated as

$$p(\mathbf{e}^C|x) = p(e_A, e_B|x) = p(e_A|x)p(e_B|x),$$

calculation of $p(x|\mathbf{e}^P)$ will not be so simple, but any way:

[8] R. O. Duda, P. E. Hart and D. G. Stork (2000) "Pattern Classification." 2nd Edition, John Wiley & Sons.

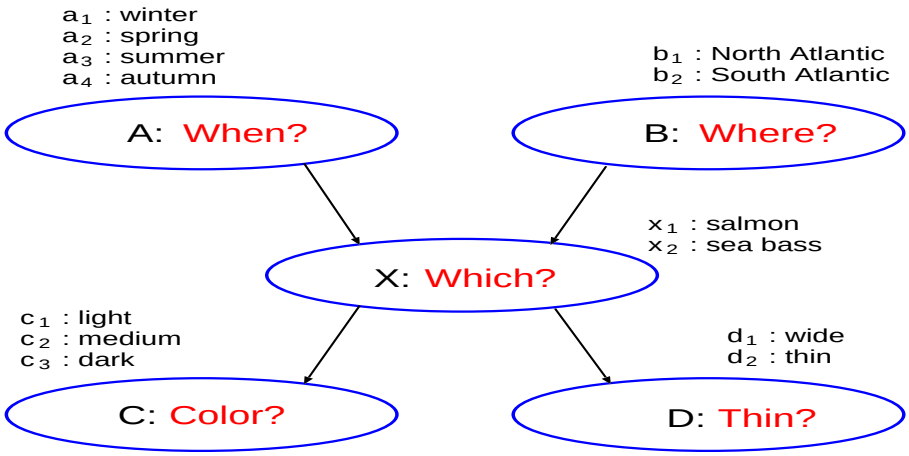

Figure 23: A Bayesian network as to classify the fish caught to Salmon or Sea-bass.

$$
\begin{aligned}
p(x|\mathbf{e}^P) &= p(x|e_A, e_B) \\
&= \sum_{\text{all possible } i,j} p(x|a_i, b_j)p(a_i, b_j|e_A, e_B) \\
&= \sum_{\text{all possible } i,j} p(x|a_i, b_j)p(a_i|e_A)p(b_j|e_B)) \\
&= p(x|a_1, b_1)p(a_1|e_A)p(b_1|e_B) \\
&+ p(x|a_1, b_2)p(a_1|e_A)p(b_2|e_B) \\
&+ p(x|a_2, b_1)p(a_2|e_A)p(b_1|e_B) \\
&+ p(x|a_2, b_2)p(a_2|e_A)p(b_2|e_B) \\
&+ p(x|a_3, b_1)p(a_3|e_A)p(b_1|e_B) \\
&+ p(x|a_3, b_2)p(a_3|e_A)p(b_2|e_B) \\
&+ p(x|a_4, b_1)p(a_4|e_A)p(b_1|e_B) \\
&+ p(x|a_4, b_2)p(a_4|e_A)p(b_2|e_B)
\end{aligned}
$$

Now the arc of the network, that is, the conditional probability of each of the arcs are given in Figure24.

We now assume that the evidences we have is, e_A is Winter, e_B = South Pacific, e_C = light, and e_D = thin. Then assume our concern is, how much is the probability of hypothesis that fish is salmon under these evidences?

Our scenario is as follows: it's Winter now, so $p(a_1|e_A) = 1$ and $p(a_i|e_A) = 0$ for $i = 2, 3, 4$. Suppose we don't know from which sea the boat came from, but the chief of the fishing crew prefers to fish in South Pacific Ocean, so assume $p(b_1|e_B) = 0.2$ and $p(b_2|e_B) = 0.8$.

p(when)

Winter	Spring	Summer	Autumn
0.30	0.25	0.20	0.25

p(where)

North	South
0.60	0.40

p(which | when)

	Salmon	Seabass
Winter	0.90	0.10
Spring	0.30	0.70
Summer	0.40	0.60
Autumn	0.80	0.20

p(which | where)

	Salmon	Seabass
North	0.65	0.35
South	0.25	0.75

p(color | which)

	Light	Medium	Dark
Salmon	0.33	0.33	0.34
Seabass	0.80	0.10	0.10

p(thickness | when)

	Salmon	Seabass
Salmon	0.40	0.60
Seabass	0.95	0.05

Figure 24: Given primer and conditional probabilities of Salmon and Sea-bass.

Further, the fish is fairly light, so $p(e_C|c_1) = 1$, $p(e_C|c_2) = 0.5$ and $p(e_C|c_3) = 0$. For some reason we cannot measure the thickness of the fish, so assume $p(e_D|d_1) = p(e_D|d_2) = 0.5$.

Then try to calculate the probability of hypothesis that fish is salmon under these evidences.

We can change the scenario as those in the following three exercise.

Exercise 17 *Suppose it's November 10 – the end of autumn and the beginning of winter – and thus let $p(a_1) = p(a4) = 0.5$. Furthermore it is know that the fish was caught in North Atlantic, that is $p(b_1) = 1$. Suppose the color of the fish was not measured, but it is known that fish is thin, that is, $p(d_2) = 1$. Classify the fish as salmon or sea-bass. What is the expected error rate of the estimate?*

Exercise 18 *Suppose all we know about fish is thin and medium light color. What season is now most likely? And what is the probability it's being correct?*

Exercise 19 *Suppose the fish is thin and medium lightness and that it was caught in North Atlantic. Then the same question as above, what season is now most likely? And what is the probability it's being correct?*

8 Algorithms for inference

This section is paraphrased all from the section 3.3 in Korb et al.[9]

8.1 Exact inference

8.1.1 Kim and Pearl's message passing algorithm

Assume X is the query node, and there is some set of evidence nodes E (not including X). The task is to update $Bel(X|E)$ by computing $p(X|E)$.

A node X is with all its connections to parents (the U_i), children (the Y_j), and the children's other parents (the Z_{ij}). The local belief updating for X must incorporate evidence from all other parts of the network.

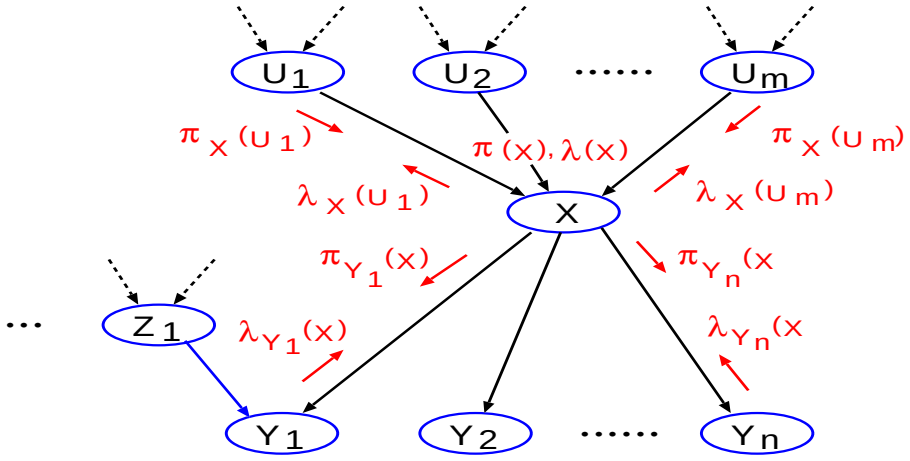

Figure 25: Extended Pearl's Earthquake

Evidence can be divided into:

- The predictive support for X, from evidence nodes connected to X through its parents, $U_1, ..., U_m$.
- The diagnostic support for X, from evidence nodes connected to X through its children $Y_1, ..., Y_n$.

We use two types of messages π and *lambda*. The messages π are sent in the direction of the arc, from parent to child, hence we can notate

$$\pi_{\text{receiver}}(\text{sender}),$$

while the messages π are sent against the direction of the arc, from child to parent, hence we can notate

$$\lambda_{\text{sender}}(\text{receiver}).$$

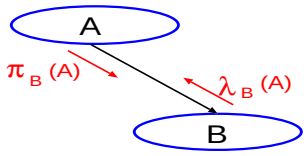

Figure 26: Notation of two messages λ and π between two nodes.

Note that π is from *prior* and λ is from *likelihood* in Bayes' theorem.

[ALGORITHM]

0. Initialization
(1) Set all λ values $\lambda(x_i)$ to 1.
(2) Set all λ messages $\lambda_X(U_i)$ and $\lambda_{Y_j}(X)$ to 1. (3) Set all π messages $\pi_X(U_i)$ and $\pi_{Y_j}(X)$ to 1.

1. Bottom-up propagation
Node X computes new λ messages to send to its parents

$$\pi(x_i) = \sum_{u_1,...,u_n} p(x_i|u_1,...,u_n) \prod_i \pi_X(u_i)$$

2. Top-down propagation
Node X computes new π messages to send to its children.

$$\pi_{Y_j}(x_i) = \alpha \prod_{k \neq j} \lambda_{Y_k}(x_i) \sum_{u_1,...,u_n} p(x_i|u_1,...,u_n) \prod_i \pi_X(u_i)$$

3. Belief updating
At each iteration of the algorithm, $Bel(X)$ is updated as:

$$Bel(x_i) = \alpha \lambda(x_i) \pi(x_i)$$

where

$$\lambda(x_i) = \prod_j \lambda_{Y_j}(X)$$

and

$$\pi(x_i) = \sum_{u_1,...u_n} p(x_i|u_1,...,u_n) \prod_i \pi_X(U_i).$$

Note that α is a normalizing constant so that $\sum_{x_i} Bel(X = x_i) = 1$.

[9]K. B. Korb and A. E. Nicholson (2003) "Bayesian Artificial Intelligence."

Note also that $p(X|U_1, ..., U_n)$ is given as the conditional probability table in the original Bayesian network.

An example – Extended Pearl's Earthquake

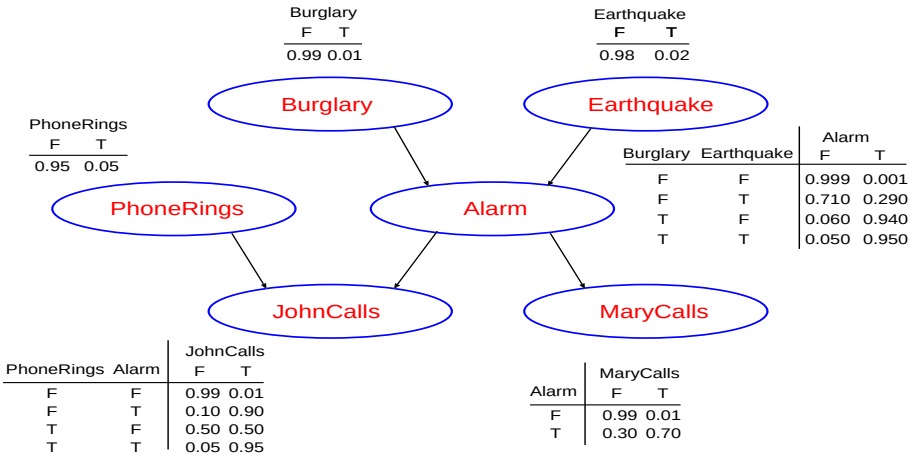

Figure 27: Extended Pearl's Earthquake

8.2 Approximate inference

8.2.1 Logic Sampling

Repeatedly a case is generated by selecting values for each node at random. In each cycle, the nodes are traversed from the roots down to the leaves. Selecting values is by following either the prior or the conditional probability table already given. When all the nodes have been visited, we have a case, or equivalently, an instantiation of all the nodes in the Bayesian network.

To estimate $p(X|E)$ with a sample value $'(X|E)$, we compute the ratio of cases where both X and E are true to the number of cases where just E is true.

8.2.2 Likelihood Weighting

This is a modified version of Logic Sampling described above. Modification is to avoid unlikely evidences. That is, if a node is the one evidence is specified choose the evidence, otherwise one of the values will be selected following the prior, if it's a root, or conditional probability table.

9 Bayesian network for decision making

9.1 Utility

When we make a decision of an action, we might consider our preferences among different possible outcomes of those available actions. In the Bayesian decision theory this preference is called *utility*, or we may rephrase it as "usefulness," "desirability," or simply "value" of the outcome.

Introducing this concept of utility allows us to calculate which action is expected to result in the most valuable utility given any available evidence E.

We now define *expected utility* as:

$$eu(A|E) = p(O_i|E, A)u(O_i|A),$$

where A is an action with possible outcome O_i. E is the available evidence. $U(O_i)|A$ is the utility of each of the outcome under the action A. $p(O_i|E, A)$ is the conditional probability distribution over the outcome O_i under the action A with the evidence E. E is the available evidence.

9.1.1 Three different nodes to express network

- Chance nodes
- Decision nodes:
  - The decision being made at a particular point in time. The values of a decision node are the actions
- Utility nodes:
  - Each utility node has an associated utility table with one entry for each possible instantiation of its parents, perhaps including an When there are multiple utility nodes, the overall utility is the sum of the individual utilities.

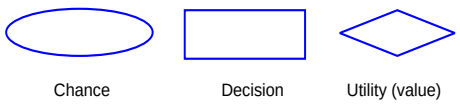

Figure 28: Symbol to express BDN – Chance, Decision, and Utility

9.2 Example-1: To bet or not to my football team?

Clare's football team, Melbourne, is going to play her friend John's team, Carlton. John offers Clare a friendly bet: whoever's team loses will buy the wine next time they go out for dinner. They never spend more than $15 on wine when they eat out. When deciding whether to accept this bet, Clare will have to assess her team's chances of winning (which will vary according to the weather on the day). She also knows that she will be happy if her team wins and miserable if her team loses, regardless of the bet.

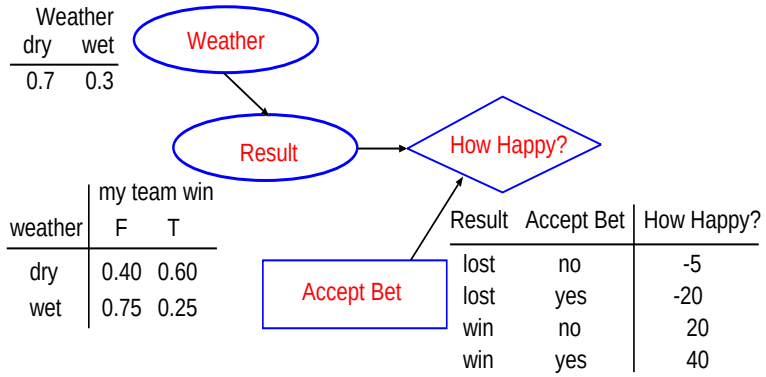

Figure 29: To bet or not to my football team?

9.2.1 Information links

There may be arcs from chance nodes to decision nodes - these are called information links.

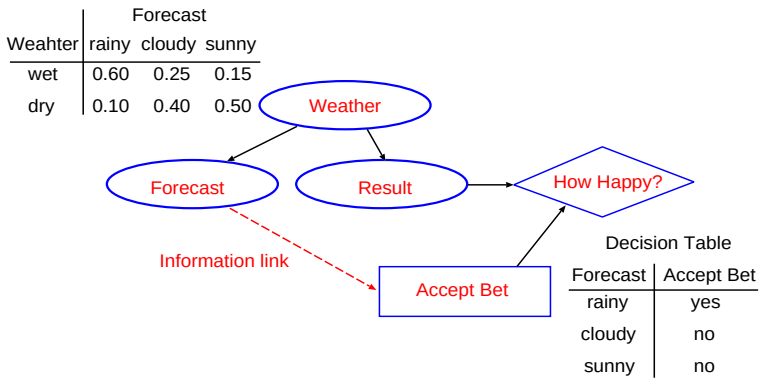

Figure 30: An example of information link.

9.3 Sequential decision making

9.3.1 Revisit to the Flu example

Suppose that you know that a fever can be caused by the flu. You can use a thermometer, which is fairly reliable, to test whether or not you have a fever. Suppose you also know that if you take aspirin it will almost certainly lower a fever to normal. Some people (about 5% of the population) have a negative reaction to aspirin. You'll be happy to get rid of your fever, as long as you don't suffer an adverse reaction if you take aspirin.

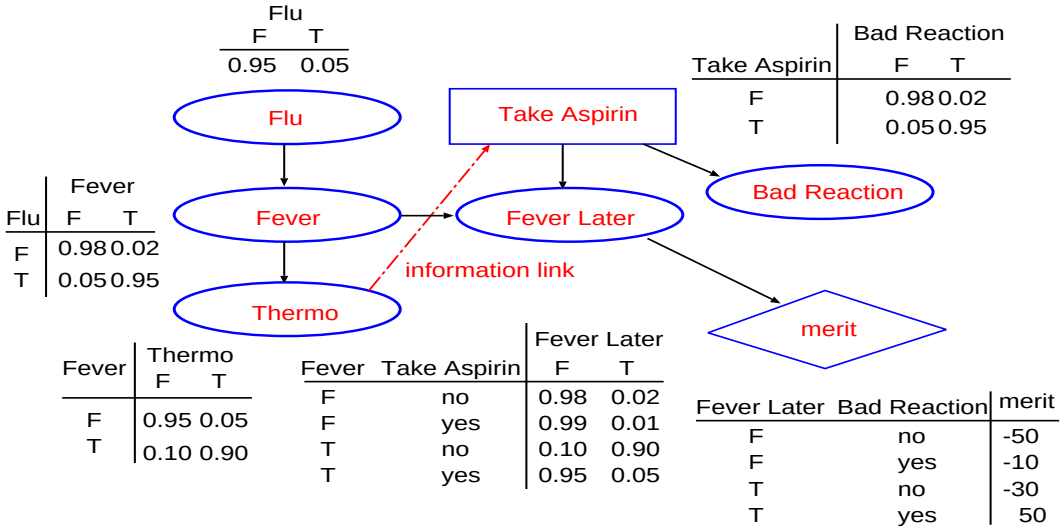

Figure 31:

9.3.2 An investment to a Real estate

Paul is thinking about buying a house as an investment. While it looks fine externally, he knows that there may be structural and other problems with the house that aren't immediately obvious. He estimates that there is a 70% chance that the house is really in good condition, with a 30% chance that it could be a real dud. Paul plans to resell the house after doing some renovation. He estimates that if the house really is in good condition (i.e., structurally sound), he should make a $5,000 profit, but if it isn't, he will lose about $3,000 on the investment. Paul knows that he can get a building surveyor to do a full inspection for $600. He also knows that the inspection report may not be completely accurate. Paul has to decide whether it is worth it to have the building inspection done, and then he will decide whether or not to buy the house.

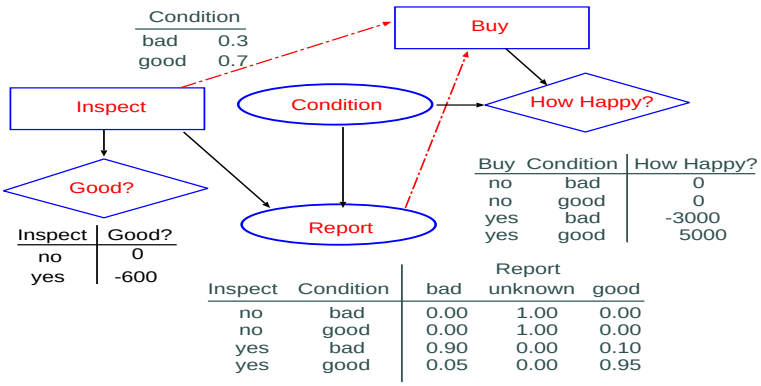
Condition
bad 0.3
good 0.7
Buy
Inspect
Condition
How Happy?
Good?
Report
Buy Condition How Happy?
no bad 0
no good 0
yes bad -3000
yes good 5000
Inspect Good?
no 0
yes -600
Report
Inspect Condition bad unknown good
no bad 0.00 1.00 0.00
no good 0.00 1.00 0.00
yes bad 0.90 0.00 0.10
yes good 0.05 0.00 0.95

Figure 32:

9.4 Dynamic Bayesian network (DBN)

When we say Bayesian network, usually all events are static. In other words, all the probability value do not change as time goes by. But as we see in the Flu-Fever-Aspirin example above, taking an aspirin influence the fever tomorrow. Now we study the probabilities are dynamically change as a function of time with the structure of the network basically remaining the same. The structure of the network at time t is called a *time-slice*. Arcs in one time-slice is called *inter-slice* arcs while arcs link to the next time-slice are called *intra-slice* arcs.

Intra-slice are usually not from all the nodes to the corresponding nodes in the next time-slice. Only sometimes we have such connections as $X_i(T) \rightarrow X_i(t+1)$. Or sometimes one node in one time-slice links different node in the next time-slice $X_i(T) \rightarrow X_j(t+1)$.

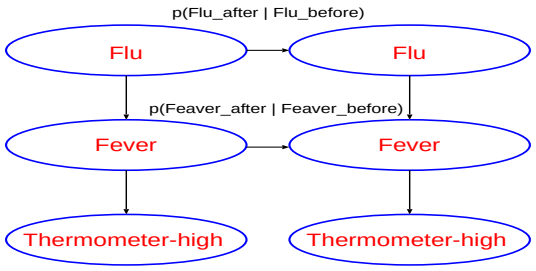

Figure 33: A simple example of Dynamic Bayesian Network.

Sometimes some nodes wield an observation, and as a result we see a time series of observation. In this scenario the node that wields the observation is called *state*.

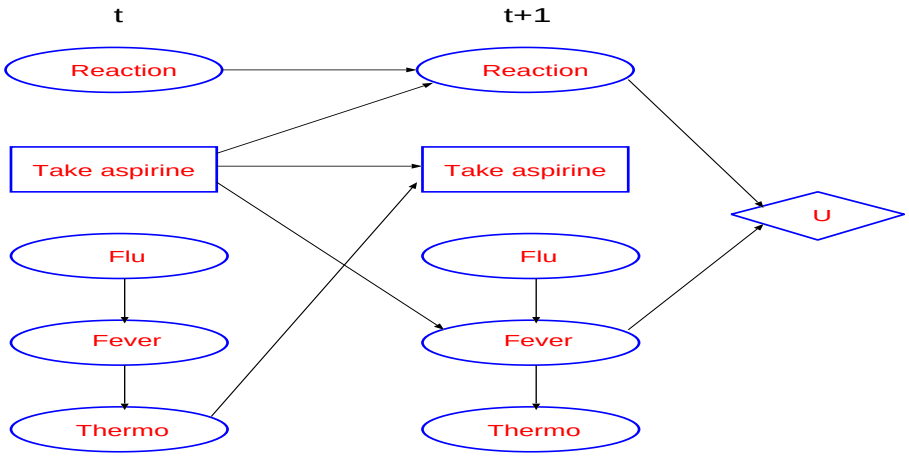

Figure 34: Still simple but more realistic Dynamic Bayesian Network.

Sometimes we want to call a node which creates an result that we can observe. From the other field like a Model of Automaton or the Hidden Markov Model, it might be

convenient to call such nodes *state* and *observation*.

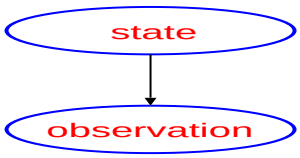

Figure 35: Node "*state*" and node "*observation*"

Such Dynamic Bayesian Network are useful when we must make a decision making in an uncertainty. That is to say, it's a good tool for *Sequential design making* or *planning under uncertainty*. Let's recall the example of decision making: to take an aspirin or not to take being afraid of it bad reaction to a body.

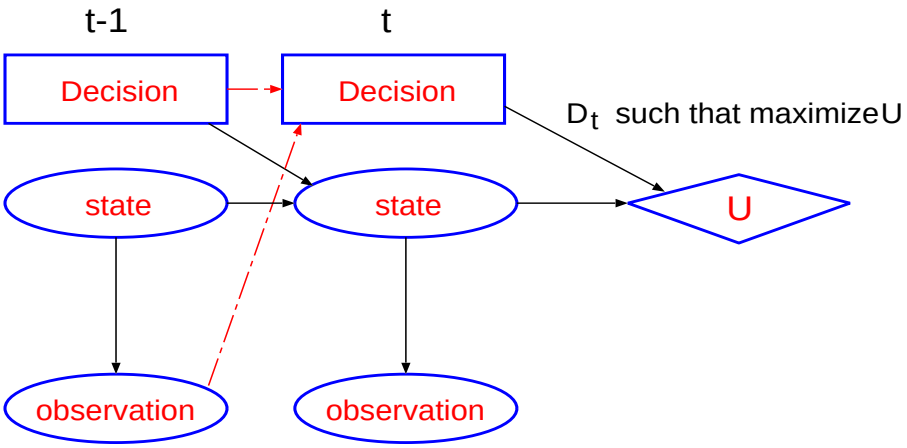

Figure 36: A Dynamic Bayesian Network for decision making.

9.4.1 Mobile robot example

We now assume a mobile robot whose task is to detect and chase a moving object. The robot should reassess its own position as well as the information where the target object is. The robot observes at any slice of time its position with respect to walls and corners and the target position with respect to the robot.

We denote the real location of own and target at time t as $S_T(t)$ and $S_R(t)$, and the observation of location of own and target at time t as $O_T(t)$ and $O_R(t)$. Utility is the distance from own to target.

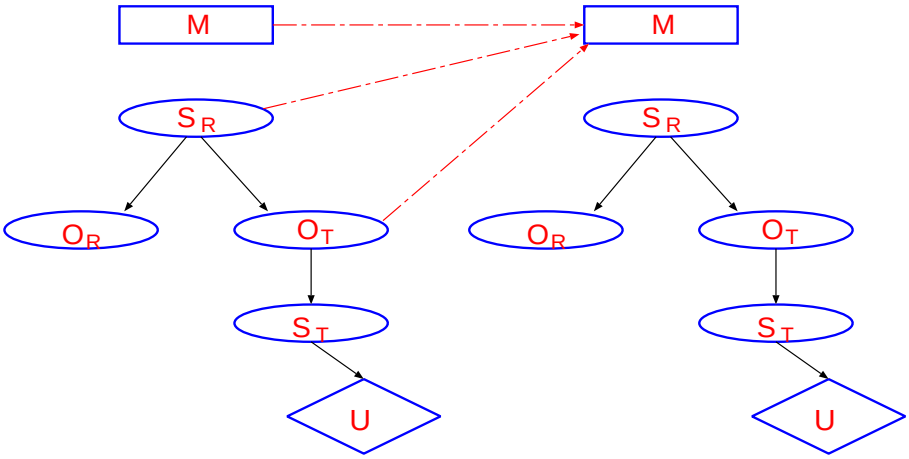

Figure 37: Dynamic Bayesian Network for a mobile robot.

PART III

HIDDEN MARKOV MODEL & BAYESIAN NETWORK

10 Hidden Markov Model

This subsection describes what is Hidden Markov Model[10], taking 3 examples from the wonderful tutorial on Hidden Markov Model by Ravinar[11]. Nowadays almost all speech recognition systems use this model. Or many areas in computational molecular biology such as grouping amino-acid sequences into protein families, or fault detection.

10.1 Markov Process

Before we study Hidden Markov Model, let's see Markov process in general. Image a system with N states, each of which can transfers to another state, or to itself, with a probability. To be more specific, state i transfer to the state j with a probability a_{ij}.

$$a_{ij} = p(q_t = S_i | q_{t-1} = S_j)$$

This is called N-state Markov Model. For example, assume now $N = 3$, we have nine such transition probability and we can represent them as the matrix A.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

10.1.1 A 3-state Markov model of the weather – Not yet "hidden" though

We now take a daily weather transition as an example. Imagine the weather at noon everyday. One day it's rain, the next day it's cloudy, the third day it's cloudy again, the fourth day it's fine, the fifth day it's rain again, the sixth day it's fine, and so on.

Let's assume the transition probability is:

[10]Named after a mathematician Andrei Andreyevich Markov who were born in Ryazan, Russia in 1856 and died in Petrograd, Russia in 1922.

[11]L. R. Ravinar (1989) "a tutorial on Hidden Markov Models and selected applications in speech recognition." Proceedings of the IEEE Vol. 77 No. 22.

$$A = \left(\begin{array}{c|ccc} & Fine & Cloudy & Rain \\ \hline Fine & 0.7 & 0.1 & 0.2 \\ Cloudy & 0.5 & 0.4 & 0.1 \\ Rain & 0.6 & 0.3 & 0.1 \end{array} \right).$$

Or, schematically,

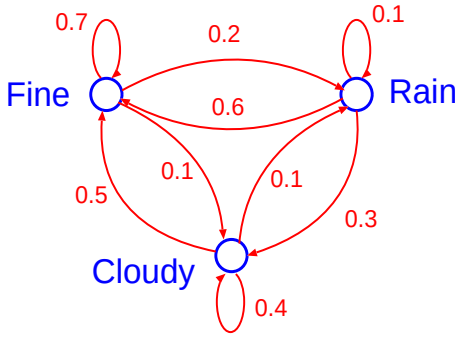

Figure 38: A toy example of State transitions between three distinct weathers.

Exercise 20 *Calculate the probability of the observation during a 7 day in a row.*

sunny – sunny – sunny – rainy – rainy – sunny – cloudy – sunny

10.2 Hidden Markov Model

Then what is Hidden Markov Model? Assume now we have N states. The probability of the first state to be started with is i-th state is denoted.

$$\pi_i, \quad i = 1, 2, \cdots, N.$$

Then the probability of transition from state i to state j is a_{ij}. Let's denote this with the matrix A:

$$A = \{a_{ij}\}, \quad i = 1, 2, \cdots, N.$$

We further assume each state results in one of M distinct observations:

$$o_1, o_2, o_3, \cdots, o_M.$$

The probability that state S_i results in the observation o_j is denoted as b_{ij} also denoted with the matrix B:

$$B = \{b_{ij}\}, \quad i = 1, 2, \cdots, M.$$

The important thing is, all we can know is only a series of observations. we cannot know the state that each of those observations has been resulted from. The state transition is

hidden to us. This is the reason of the we call it Hidden Markov Model.

Let's take an example when $N = 2$ and $M = 4$.

$$B = \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{pmatrix}.$$

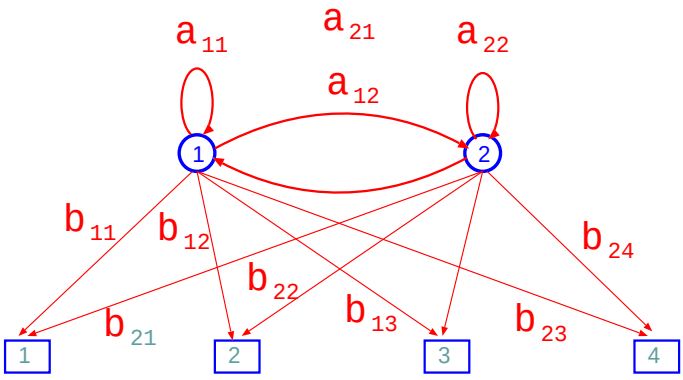

Figure 39: A schematic diagram of Hidden Markov model with 2 states and 4 observations.

Now assume a series of observations was:

$$O = \{o_1, o_2, o_3, \cdots, o_T\}$$

Let us now denote the state at time t as q_t. Generally speaking, the transition probability depends on all those states it has shifted from the start. That is,

$$a_{ij} = p(S_j).$$

Then let's think of just a special case where the state at time t is only affected by the state at time $t-1$. Then Notation

$$\lambda = \{\pi, A, B\}$$

We express the state at time t as q_t, that is, $q_t = S_i$ is the state at time t is S_i. o_t is observation at time t.

Then the probability of this observation given the model λ is

$$p(O|\lambda) = \sum_q \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}} a_{q_t} b_{q_t} o_t$$

10.2.1 An example: Fair dice and biased dice

Now let's see an example. Assume we have two dices. One is fair dice and the other is biased one. We have two states. The first state is the fair dice, and the second state is the biased one. Each gives one of 6 observations with a probability distribution. For example, fair coin (state 1), of course, results in either of 1, 2, 3, 4, 5, or 6 with the equal probability, namely 1/6, while the biased coin (state 2) results in 1, 2, 3, 4, and 5 with the probability of 1/10 and 1/2 for 6. Assume transition probability between two states is

$$A = \begin{pmatrix} 0.90 & 0.05 \\ 0.10 & 0.95 \end{pmatrix}.$$

See the Figure below.

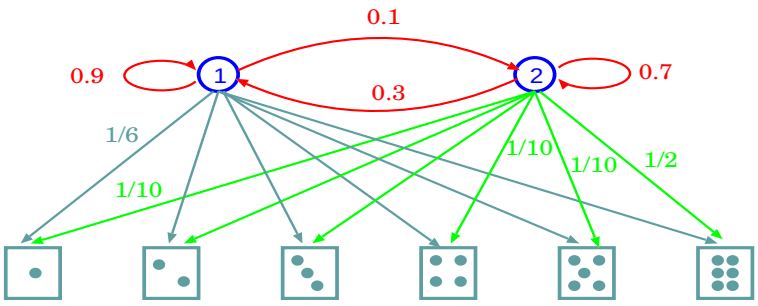

Figure 40: A toy example of a fair dice and biased one.

Assume we cannot know which dice is used. Behind the curtain a hidden man changes the dice from time to time according to the probability distribution A, with the first dice is chosen according to the probability π. All we can observe is the result of the dice, such as:

Then, a question is, e.g., "How likely the following result will occur?"

Or

Or, we can imagine the situation at a casino, where a roulette dealer tries to switch his roulette machine a fair mode and a biased mode.

10.2.2 Another example: Coin toss models

Now behind the curtain a dealer tosses coins several times. And each time the dealer tells us the result is *"Head"* or *"Tail."*

The dealer has a multiple coins each has a different probability to result in Head or Bottom. The dealer changes from one coin to another using a prefixed transition probability table. Then assume the result was

Head, Head, Head, Tail, Head, Tail, Head, Head, Head, Tail, Tail, Head

We might guess that the occurrences of *Head* are too many for the coins are fair. Then the question is, "what happens behind the curtain?"

Two coin model
We suppose the dealer has two unfair coins and the dealer follows the transition rule

$$A = \left(\begin{array}{c|cc} & Coin-1 & Coin-2 \\ \hline Coin-1 & a_{11} & a_{12} \\ Coin-2 & a_{21} & a_{22} \end{array} \right).$$

And each coin result in *Head* or *Tail* is from the probability table:

$$A = \left(\begin{array}{c|cc} & Coin-1 & Coin-2 \\ \hline Head & b_{11} & b_{12} \\ Tail & b_{21} & b_{22} \end{array} \right).$$

Further, the dealer start the tossing with i-th coin $(i = 1, 2)$ with the probability π_i. See the Figure below.

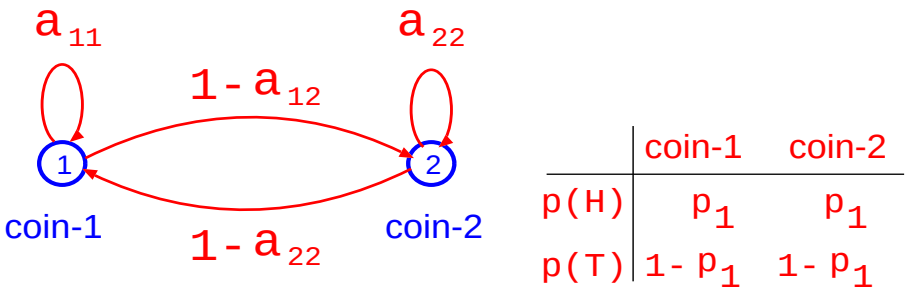

Figure 41: A coin model assuming two coins.

Now our question might be to guess these 15 parameters, that is, $\lambda = \{\pi_i, A, B\}$ so that the observation is most likely, or equivalently, to maximize the probability of the observation:

$$p(Head, Head, Head, Tail, Head, Tail, Head, Head, Head, Tail, Tail, Head).$$

Or, the other question might be, "Which series of hidden states are most likely?"

Three coin model
Yet another possibility is, the dealer has 3 coins, not two. Then the state transition would be:

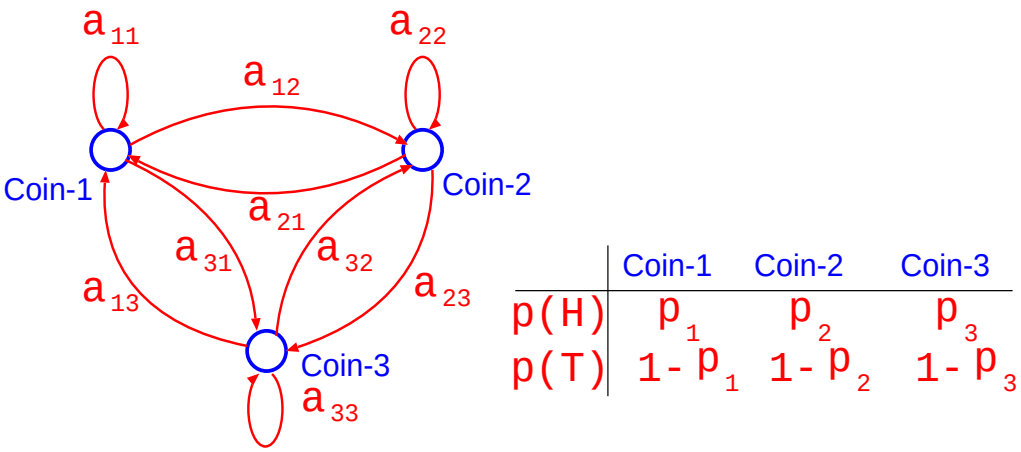

Figure 42: A coin model assuming three coins.

10.2.3 Further example: two urns with colored balls

The final toy example is with two urns each contains a number of balls with red, blue, and purple color. The. The state-1 is the first urn and the state-2 is the second urn. Assume now, the first urn contains 5 red, 3 blue, and 1 purple balls while the second contains 2 red, 4 blue, and 2 purple balls. The dealer, behind the curtain, pick up one ball from either of the urns and tell us the color of the ball. Then the dealer return the ball to the urn he picked up the ball. This is repeated by changing urn to pick up the ball from according to the preset probability table as previous two examples. See the Figure.

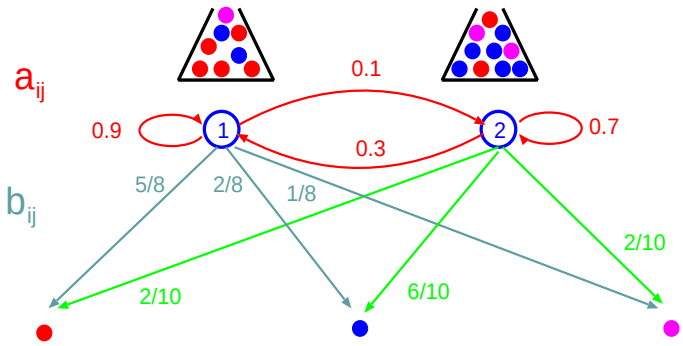

Figure 43: A toy example of thought experiment with two urns each of which contains balls with three different colors.

The question might be, "What is the probability of a series of observations, like in the Figure below.

10.3 Three important problems

As we have seen in examples above, three important questions raise.

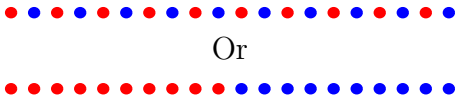

Figure 44: A toy example of thought experiment with two urns each of which contains balls with three different colors.

- Given $\lambda = \{\pi, A, B\}$, estimate "How likely the observation is?" That is:

  $p(o_1, o_2, o_3, \cdots, o_T) =?$

- Given $\lambda = \{\pi, A, B\}$, guess the most likely series of hidden states?

  $(q_1, q_2, q_3, \cdots, q_T)$

  so that we can predict the probability of the next state:

  $p(q_{T+1}|q_1, q_2, q_3, \cdots, q_T)$

- Given a series of observations $(o_1, o_2, o_3, \cdots, o_T)$ modify λ so that it will maximize the $p(o_1, o_2, o_3, \cdots, o_T)$.
  - ⋆ This might be called *TRAINING*, or equivalently, *LEARNING* of the model.

To be more specific:

- The Evaluation Problem
  - ⋆ Given the observation sequence O and a model *lambda*, how do we efficiently evaluate the probability of being produced by the source model */lambda*. That is, what is $p(O|\lambda)$.
- The Decoding Problem
  - ⋆ how to deduce from O the most likely state sequence q in a meaningful manner.
  - ⋆ it is important in many applications to have the knowledge of the most likely state sequence for several reasons. As an example, if we use the states of a word model to represent the distinct sounds in the word, it may be desirable to know the correspondence between the speech segments and the sounds of the word, because the duration of the individual speech segments provides useful information for speech recognition.
- The Estimation Problem
  - ⋆ Given the observation O, how do we solve the inverse problem of estimating the parameters in λ?
  - ⋆ Given an observation sequence (or a set of sequence) O, the estimation problem involves finding the "right" model parameter values that specify a model most likely to produce the given sequence. In speech recognition, this is often called "training,"

11 Speech recognition

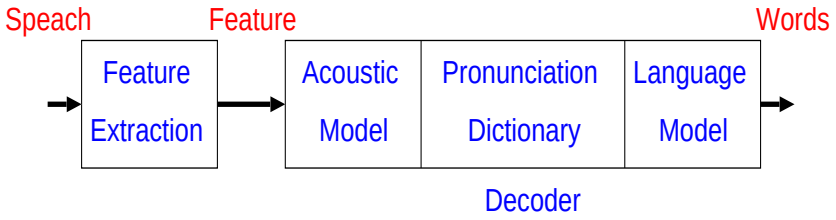

Figure 45: A block diagram of a system for speech recognition.

Usually in Speech recognition hidden states represent *phonemes* such as */k/*, */ae/*, */t/* etc.

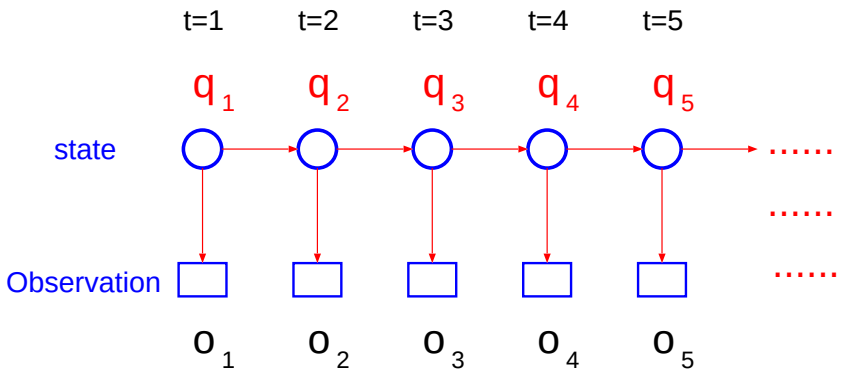

Figure 46: A schematic diagram of a time series of speech recognition by Hidden Markov Model.

$$\hat{w} = \arg\max_w \{p(w|\boldsymbol{y})\}.$$

However, as $p(w|\boldsymbol{y})$ is difficult to be treated, we transform this to

$$\hat{w} = \arg\max_w \{p(\boldsymbol{y}|w)p(w)\}.$$

recalling the Bayes' rule.
Then $p(\boldsymbol{y}|w)$ is determined by an *acoustic model* and the $p(w)$ is determined by a *language model*.
1) Feature analysis: A special and/or temporal analysis of the speech signal is to give observation vectors for training the HMMs which characterize speech sounds 2) Unit matching Typically each such unit is characterized by an HMM whose parameters are estimated from a training set 3) Lexical decoding: Vocabulary must be specified in terms of the basic units for recognition. such a specification can be either deterministic, or stochastic. deterministic =¿ one or more FSA for each word in the vocabulary statistical =¿ p is attached to the arcs in the FSA representation of words. 4) Syntactic analysis:

A grammar can be represented by a deterministic finite state automaton. 5) Semantic analysis: depending on the state of the recognizer, some syntactically correct input strings are eliminated from consideration.

11.1 Speech recognition by Hidden Markov Model

11.2 Speech recognition by Dynamic Bayesian Network

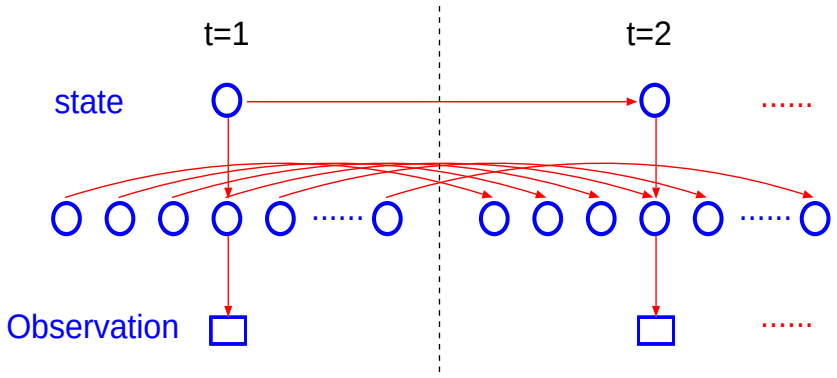

Figure 47: A block diagram of speech recognition by Dynamic Bayesian Network.

11.3 Hidden Markov Model & Dynamic Bayesian Network

Thus far we have learned Hidden Markov models are a particular form of Dynamic Bayesian networks through an application of speech recognition.

12 Stock Market Forecasting

12.1 A simple model - Not hidden

Let's start with a very simple model in which states are not hidden. States are, for example, (1) good financial state also know as Bull State, (2) normal state, and (3) bad state a.k.a. Bear state. Assume state transition probabilities are know. When one state changes to another state, economy also change. For example, either of (i) large rise, (ii), small rise, (iii) no change, (iv) small fall, and (v) large fall. Then the phenomena will be shown as Figure 48 [12].

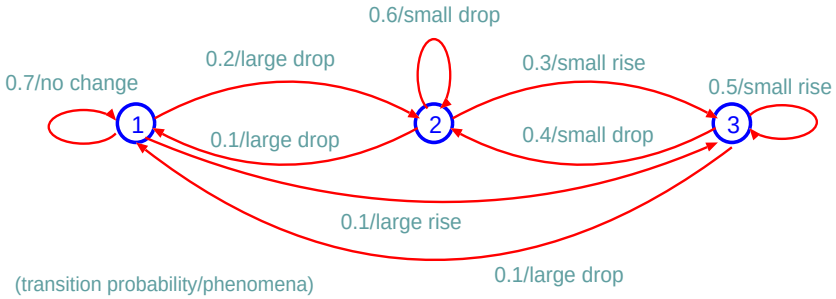

Figure 48: An example of three-state model for stock market forecasting.

Then we can calculate the probability of a series of state changes, like

$$1 \to 1 \to 3 \to 2 \to 3 \to 2 \to 1$$

12.2 Hidden Markov Model for stock market forecasting

States are financial states, as the example in the previous section. Or trends of investors, government's financial strategies, or something else. In any case each state reflects, or equivalently, results in, the stock price. Depending on what we want to forecast, the price is average of all stocks, or specific one stock, such as Microsoft Corporation, I.B.M, or Google Co, etc. Or instead of prices we can think of fluctuation of the price. Anyway those values are not discrete as we have studied so far, but a continuous value. We can digitize the value, but usually we analyse the continuous values. Therefore, the probability of one state to result in an observation should be also continuous. The most popular way is using a weighting sum of Gaussian probability distribution functions. This is called

[12] This example was taken from
Y. Zhang (2001) "Prediction of financial time series with Hidden Markov Models." Msc dissertation submitted to Simon Fraser University.

Gaussian Mixture p.d.f.

$$\sum_{i=1}^{k} N_i(\mu_i, \sigma_i)$$

where

$$N(x, \mu, \sigma) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{x - x_i}{h})^2).$$

For example:

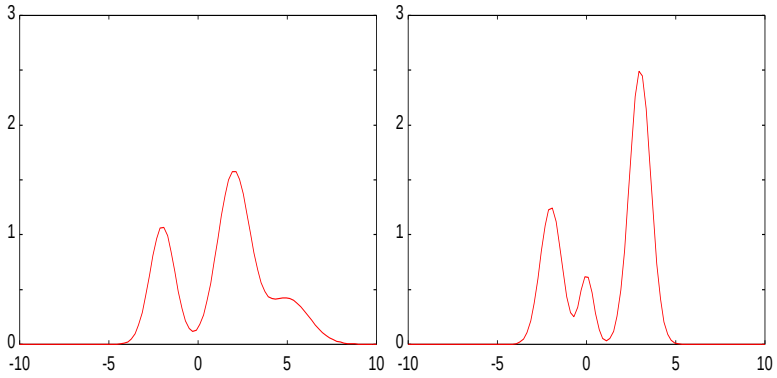

Figure 49: Two Gaussian-mixtures. Left $0.2 * N(5, 1.2) + 0.5 * N(2, 0.9) + 0.3 * N(-2, 0.7)$ and right $0.6 * N(3, 0.6) + 0.1 * N(0, 0.4) + 0.3 * N(-2, 0.6)$

Or, if observation is not one dimensional value such as price, but a high-dimensional vector, like

$$\{\textit{today's price, yesterday's price, price two day's ago }\}$$

then we use a high-dimensional Gaussian p.d.f.

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t {\Sigma_i}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\}$$

as we studied previously in Part I.

Then imagine two-dice-example we saw in Figure 40. We have two states each of which result in observation of a continuous value following the Gaussian mixture p.d.f., instead of 1, 2, 3, 4, 5, or 6 following a probability like 1/6 or 1/10.

Recall a training of a HMM from a dataset D is obtaining the model M by maximizing

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}.$$

When, however, the model M includes a set of continuous parameters θ the probability is

$$p(M|D) = \frac{\int p(D|\theta, M)p(\theta)d\theta p(M)}{p(D)}.$$

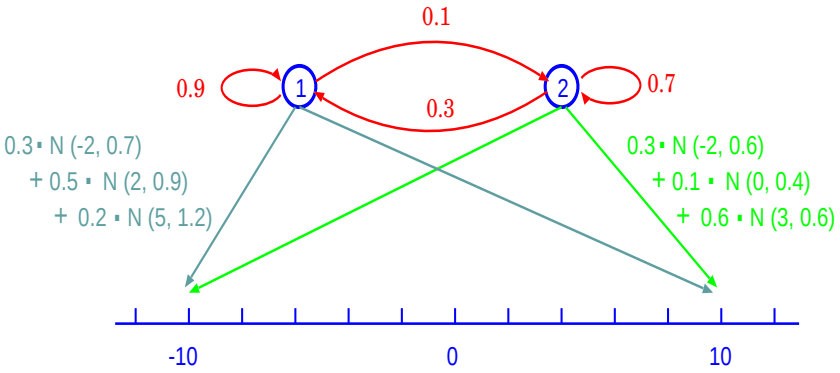

Figure 50: A schematic diagram of two states each of which yields an observation with its Gaussian mixture p.d.f.

In stock market prediction, assume we have a series of observation y_t, that is

$$D = \{y_1, y_2, y_3, \cdots, y_t\}.$$

Then what we want to know is,

$$p(y_{t+1}|D) = \int p(y_{t+1}|\theta, M, D)p(\theta|M, D)p(M|D)d\theta dM.$$

This is called *predictive distribution*.

Now our hidden states are investment strategies or trend that are mixture of the weighted sum of 4 multi dimensional Gaussian pdf.

$$\sum_{i=1}^{4} w_i N_i(\Sigma_i, \mu_i)$$

Then we can calculate how likely the series of state transitions which yields the series of change of prices we have observed, or more importantly, we can calculate the probability of the next state and most likely observations from the predicted state, that is, the price of tomorrow, for example.

Probably most importantly, from a series of observations:

$$O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8, \cdots, O_T$$

you can optimize the parameters such that this series of observations is most likely, then with those parameters you may infer the most likely series of hidden states,

$$S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, \cdots, S_T$$

and S_{T+1} which means a most likely state of tomorrow. Finally, you can infer the most likely O_{T+1} that will most likely result in from S_{T+1}. This is what we might want to know, namely, tomorrow's price.

I wish your luck!

CONCLUDING REMARKS

In the article in The New York Times on 16 March 2012, Steve Lohr wrote:

> *Google search, I.B.M.'s Watson Jeopardy-winning computer, credit-card fraud detection and automated speech recognition. There seems not much in common on that list. But it is a representative sampling of the kinds of modern computing chores that use the ideas and technology developed by Judea Pearl, the winner of this year's Turing Award. The award, often considered the computer science equivalent of a Nobel prize, was announced on Wednesday by the Association for Computing Machinery. "It allowed us to learn from the data rather than write down rules of logic," said Peter Norvig, an artificial intelligence expert and research director at Google. "It really opened things up."*
>
> *Dr. Pearl, with his work, he added, "was influential in getting me, and many others, to adopt this probabilistic point of view." Dr. Pearl, 75, a professor at the University of California, Los Angeles, is being honored for his contributions to the development of artificial intelligence. In the 1970s and 1980s, the dominant approach to artificial intelligence was to try to capture the process of human judgment in rules a computer could use. They were called rules-based expert systems. Dr. Pearl championed a different approach of letting computers calculate probable outcomes and answers. It helped shift the pursuit of artificial intelligence onto more favorable terrain for computing. Dr. Pearl's work on Bayesian networks - named for the 18th-century English mathematician Thomas Bayes - provided "a basic calculus for reasoning with uncertain information, which is everywhere in the real world," said Stuart Russell, a professor of computer science at the University of California, Berkeley. "That was a very big step for artificial intelligence."*
>
> *Dr. Pearl said he was not surprised that his ideas are seen in many computing applications. "The applications are everywhere, because uncertainty is everywhere," Dr. Pearl said. "But I didn't do the applications," he continued. "I provided a way for thinking about your problem, and the formalism and framework for how to do it."*
>
> *(The Turing Award, named for the English mathematician Alan M. Turing, includes a cash prize of $250,000, with financial support from Intel and Google.)*

APPENDIX

1. Three Prisoners two of whom will be executed tomorrow.
2. Quadratic form in 2-dimensional space.
3. How to calculate inverse of 3-dimensional matrix?

I. Bayesian formula in another scenario

- Three prisoners (**A**, **B**, and **C**) are in a prison.
- **A** knows that the two out of the three are to be executed tomorrow, and the rest becomes free.
- **A** thought either one of **B** or **C** is sure to be executed.
- Then, **A** asked a guard "even if you tell me which of **B** and **C** is executed, that will not give me any information as for me. So please tell it to me."
- The guard answers that **C** will. $\Rightarrow$ data D
- Now, **A** knows one of **A** or **B** is sure to be free.

Do you guess probability $p(A|D) = 1/2$?

If this is correct, then the answer of the guard had given an information as for A, since probability $p(A)$ was $1/3$ without the information.

You agree that prior probabilities of being free tomorrow for each of **A**, **B**, and **C** are

$$p(A) = p(B) = p(C) = 1/3.$$

Then, try to apply Bayesian rule, i.e., obtain the conditional probability of the data "C will be executed" under the condition that "A will be free tomorrow" And in the same way for B and C. They are:

$$\begin{aligned} p(D|A) &= 1/2. \\ p(D|B) &= 1. \\ p(D|C) &= 0. \end{aligned}$$

In conclusion:

$$p(A|D) = \frac{p(D|A)p(A)}{p(D|A)p(A) + p(D|B)p(B) + p(D|C)p(C)} = 1/3.$$

This shows probability did not change after the information!

II. Quadratic form in 2-dimensional space

You might be interested, first of all, in how points scattered are influenced by values in Σ, that is, σ_1^2, σ_2^2, and $\sigma_{12} = \sigma_{21}$. Let's observe here three different cases of Σ when $\mu = (0, 0)$.

$$(1)\ \Sigma_1 = \begin{pmatrix} 0.20 & 0 \\ 0 & 0.20 \end{pmatrix} \quad (2)\ \Sigma_2 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.9 \end{pmatrix} \quad (3)\ \Sigma_3 = \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.2 \end{pmatrix}$$

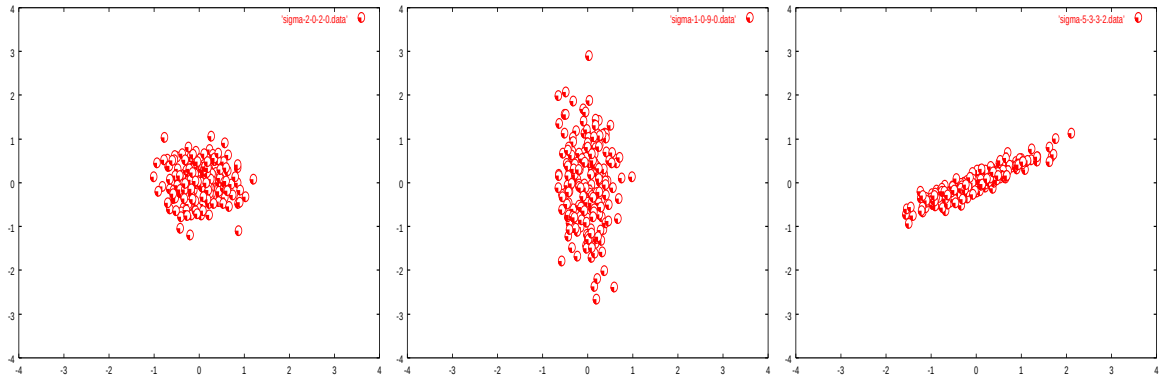

Figure 51: A cloud of 200 Gaussian random points with three different three Σ.

III. How to calculate inverse of 3-dimensional matrix.

We now try to calculate the inverse of the following 3-D matrix A which appeared in the subsection 3.2.

$$A = \begin{pmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.3 & -0.1 \\ 0.1 & -0.1 & 0.3 \end{pmatrix}$$

We use a relation $A\mathbf{x} = I$ where $\mathbf{x} = (x, y, z)^T$ and I is *identity matrix*, i.e.,

$$\begin{pmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.3 & -0.1 \\ 0.1 & -0.1 & 0.3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

It remains identical if we multiply {*2nd-raw*} by 3 and subtract the {*1st-raw*}, i.e.,

$$\begin{pmatrix} 0.3 & 0.1 & 0.1 \\ 0 & 0.8 & -0.4 \\ 0.1 & -0.1 & 0.3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

In the same way, but this time, we multiply {*3rd-raw*} by 3 and subtract the {*1st-raw*}.

$$\begin{pmatrix} 0.3 & 0.1 & 0.1 \\ 0 & 0.8 & -0.4 \\ 0 & -0.4 & 0.8 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 3 & 0 \\ -1 & 0 & 3 \end{pmatrix}$$

Then, e.g., multiply the {*1st-raw*} by 8 and then subtract the {*2nd-raw*}:

$$\begin{pmatrix} 2.4 & 0 & 1.2 \\ 0 & 0.8 & -0.4 \\ 0 & -0.4 & 0.8 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 9 & -3 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

Multiply the {*3rd-raw*} by 2 and then add the {*2nd-raw*}:

$$\begin{pmatrix} 2.4 & 0 & 1.2 \\ 0 & 0.8 & -0.4 \\ 0 & 0 & 1.2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 9 & -3 & 0 \\ -1 & 3 & 0 \\ -1 & 3 & 6 \end{pmatrix}$$

Subtract {*3rd-raw*} from the {*1st-raw*}:

$$\begin{pmatrix} 2.4 & 0 & 0 \\ 0 & 0.8 & -0.4 \\ 0 & 0 & 1.2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 10 & -6 & -6 \\ -1 & 3 & 0 \\ -1 & 3 & 6 \end{pmatrix}$$

Multiply the {*2nd-raw*} by 3 then add the {*3rd-raw*}:

$$\begin{pmatrix} 2.4 & 0 & 0 \\ 0 & 2.4 & 0 \\ 0 & 0 & 1.2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 10 & -6 & -6 \\ -4 & 12 & 6 \\ -1 & 3 & 6 \end{pmatrix}$$

Finally, divide the {*1st-raw*} by 2.4, divide the {*2nd-raw*} by 2.4, and divide the {*3rd-raw*} by 1.2, we obtain,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \frac{1}{30} \begin{pmatrix} 5 & -3 & -3 \\ -2 & 6 & 3 \\ -1 & 3 & 6 \end{pmatrix}$$

Now we know the right-hand-side is the inverse of A because the equation implies $I\mathbf{x} = B$ and it holds $AI\mathbf{x} = AB$, that is, $A\mathbf{x} = AB$. Hence $AB = I$ which means $B = A^{-1}$.

To make it sure, calculate and find

$$\begin{pmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.3 & -0.1 \\ 0.1 & -0.1 & 0.3 \end{pmatrix} \times \frac{1}{30} \begin{pmatrix} 5 & -3 & -3 \\ -2 & 6 & 3 \\ -1 & 3 & 6 \end{pmatrix}$$

$$= \frac{1}{6} \begin{pmatrix} 25 & -15 & -15 \\ -10 & 30 & 15 \\ -5 & 15 & 30 \end{pmatrix} \times \begin{pmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.3 & -0.1 \\ 0.1 & -0.1 & 0.3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Therefore

$$A^{-1} = \frac{1}{30} \begin{pmatrix} 5 & -3 & -3 \\ -2 & 6 & 3 \\ -1 & 3 & 6 \end{pmatrix}$$