Bayesian Decision Making

Lecture notes for special course in 2012

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.

PART I
BAYESIAN CLASSIFICATION

1 Bayesian Rule

The Bayesian rule is a rule to calculate the probability of a hypothesis h under the condition on some evidence e.

$$p(h|e) = \frac{p(e|h)p(h)}{p(e)}.$$

where $p(e)$ is for normalization so that the sum of probabilities $p(h|e)$ of all hypotheses is one, that is,

$$p(e) = \sum_{e} p(e|h)p(h).$$

When hypothesis is just TRUE or FALSE, then

$$p(h|e) = \frac{p(e|h)p(h)}{p(e|h)p(h) + p(e|\neg h)p(\neg h)}.$$

1.1 Examples 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 one ball. The color of the ball was red. Then was the bag estimated to be R or B, and how probable the estimate is?

Let's denote the event of picking red ball as r and blue ball as b then the probability of the bag is R under the condition is the ball picked up was red is:

$$p(R|r) = \frac{p(r|R)p(R)}{p(r|R)p(R) + p(r|B)p(B)} = \frac{(70/100)(1/2)}{(70/100)(1/2) + (30/100)(1/2)} = 0.7$$

while the probability of the bag is B under the condition is the ball picked up was red is:

$$p(B|r) = \frac{p(r|B)p(B)}{p(r|B)p(B) + p(r|R)p(R)} = \frac{(30/100)(1/2)}{(30/100)(1/2) + (70/100)(1/2)} = 0.3$$

Therefore the bag was, in conclusion, more likely to be R.

• Example-2
Then what if we bick up 5 balls, instead of just one ball, and 4 out of them were red?

Now the date is 3 balls are red and 2 balls are blue. The probability of the bag is R is:

$$p(R|data) = \frac{p(data|R)p(R)}{p(data|R)p(R) + p(data|B)p(B)}$$

where

$$p(data|R) = \begin{pmatrix} 5 \\ 3 \end{pmatrix} (70/100)^3(30/100)^2 = 0.3087$$

and

$$p(data|B) = \begin{pmatrix} 5 \\ 2 \end{pmatrix} (30/100)^3(70/100)^2 = 0.1323$$

Therefore

$$p(R|data) = \frac{0.3087}{0.3087 + 0.1323} = \frac{0.3087}{0.4410} = 0.61$$

• Example-3
After winning a race, an Olympic runner is tested for the presence of steroids. The test comes up positive, and the athlete is accused of doping. Suppose it is known that 5% of all victorious Olympic runners do use performance-enhancing drugs. For this particular test, the probability of a positive finding given that drugs are used is 95The probability of a false positive is 2%. What is the (posterior) probability that the athlete did in fact use steroids, given the positive outcome of the test?

• Example-4
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), which means that the probability of a positive test given HIV is 90%. Now suppose a guy is declared that his test was positive. What is the probability that he has HIV?

Now our hypothesis is $H =$ "He has HIV," while evidence is $E =$ "test was positive." So,

$$p(H|E) = \frac{p(E|H)p(H)}{p(E|H)p(H) + p(E|\neg H)p(\neg H)}.$$

Now that $p(H) = 1/100$, $p(\neg H) = 99/100$, $p(E|H) = 1 - 0.1$ (this is called *true positive*), and $p(E|\neg H) = 0.2$ according to false positive rate. Therefore:

$$p(H|E) = \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.2 \times 0.99} = \frac{0.009}{0.009 + 0.198} = \frac{9}{207} \approx 0.043.$$

The value is much less than you'd expected, isn't it?

• Example-5 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
- 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

This example is Taken from the book "Bayesian Artificial Intelligence" by Kevin B. Korb & Ann E. Nicholson (2004).

Note here that $p(e|\neg h)$ is not $\prod_i p(e_i\ \neg h)$, but anyway accept this is very small, say 1/3000. Also note that $p(h|e)$ is not $1 - p(e|\neg h)$ but instead

$$p(h|e) = \frac{p(e|h)p(h)}{p(e|h)p(h) + p(e|\neg h)p(\neg h)}.$$

Now if the couple in question were guilty, what is the probability of evidences? This is difficult to assess but assume it's 1 as prosecution claims. So $p(e|h) = 1$ The last question is $p(h)$ – the prior probability of a random couple being guilty.

The authors proposed an estimation of $p(h|e) = 1/1625000$ from the population of Los Angeles, and then concluded:

$$p(h|e) \approx 0.002.$$

That is, 99.8% chance of innocence.

This is what really happened in 1968 in Los Angeles. Collins and his wife were accused of robbery. Collins was a black mane with a beard and his wife was a blond white woman.

• Example-6
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 "**C** will," which is data, and we denote it **D**. Now, **A** knows one of **A** or **B** is sure to be free.

Now let's change the expression of the Bayes formula to:

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

The question is, "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!

2 Bayesian Classification

2.1 1-dimensional Gaussian

Assume, for simplicity, we now classify an object whose feature is x into either of the two classes ω_1 or ω_2. Then the probability of the object belongs to the class ω_1 given the feature x is, using our Bayesian formula:

$$p(\omega_1|x) = \frac{p(x|\omega_1)p(\omega_1)}{p(x|\omega_1)p(\omega_1) + p(x|\omega_2)p(\omega_2)}.$$

Similar calculation holds for $p(\omega_2|x)$. Then

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

Note that $p(x|\omega)$ is no more a probability value but a probability distribution function (pdf), like the Gaussian distribution function which we can apply to many cases. When we assume the Gaussian pdf, we can describe $p(x|\omega)$ as:

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

where μ is mean value and σ is standard deviation of the distribution.

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

Then what if we have multiple number of features? We should use a high dimensional pdf.

2.2 2-dimensional Gaussian

The form of the 2D Gaussian pdf is similar to the 1D Gaussian pdf, but now mean is not scalar value but a vector, and the standard deviation is not scalar either but a matrix. So, let's represent them $\boldsymbol{\mu}$ and Σ instead of μ and σ.

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

where $\boldsymbol{\mu}$ is called a mean but a vector which is made up of mean value of each feature, and Σ is called still standard deviation but a matrix.

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

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

2.2.1 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.

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) $(1 - 1/2e)x^2 - x - y^2 = 0$

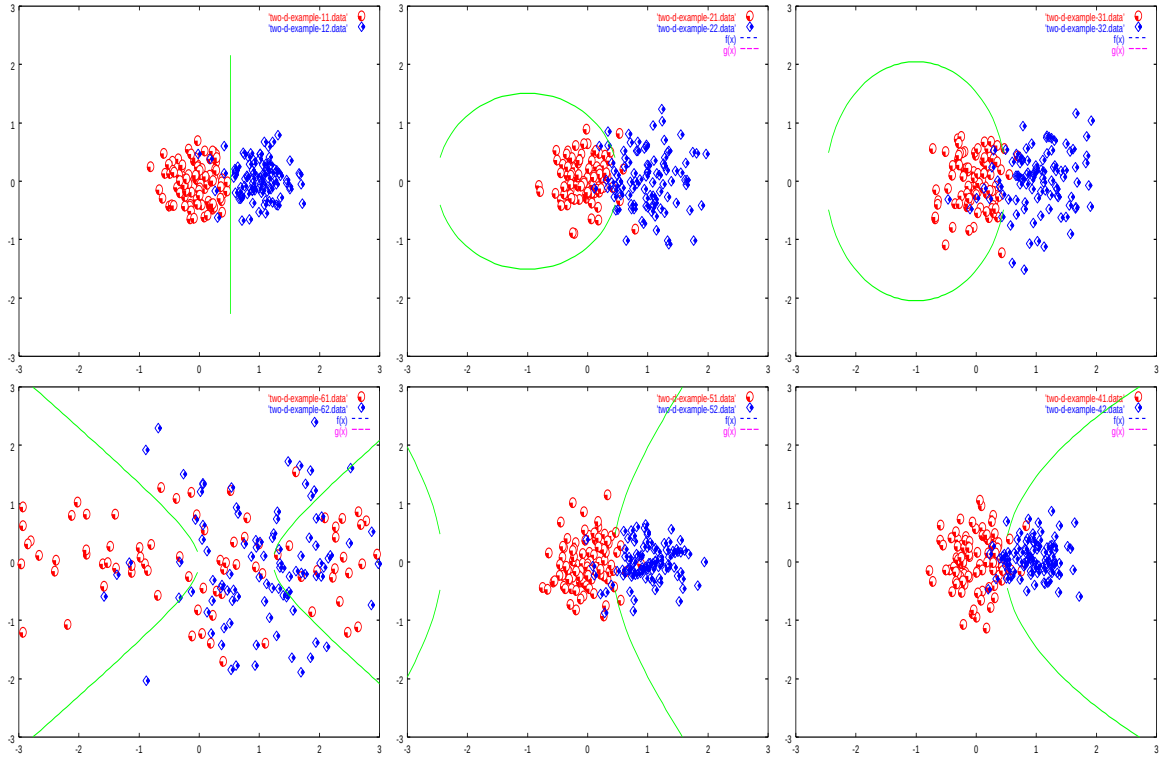


Figure 1: A cloud of 100 points each extracted from a set of two classes and border of the two classes calculated on six different conditions. (Results of (5) and (6) are still fishy and under another trial.)

⋆ **When all Σ_i's are arbitrary**

The final example in this sub-section is a 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:

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

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

$$g_i(\mathbf{x}) = \mathbf{x}^t W_i \mathbf{x} + \mathbf{w}_i^t\mathbf{x} + w_{i0} \tag{2}$$

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.

2.3 A Higher order Gaussian case

The Equation

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

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 1 (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 2 (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

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

• 3-D Gaussian case as an example

⋆ 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

$$g_i(\mathbf{x}) \quad = \quad -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \tag{14}$$

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.

PART II
BAYESIAN NETWORK

3 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."

3.1 Examples

3.1.1 Flu & Temperature

This example is taken from Korb et al.[1]

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 2.

Exercise 5 *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?*

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

Or conversely,

Exercise 6 *We have an evidence that one guy is flu then how much is a belief of this guy has a high-temperature?*

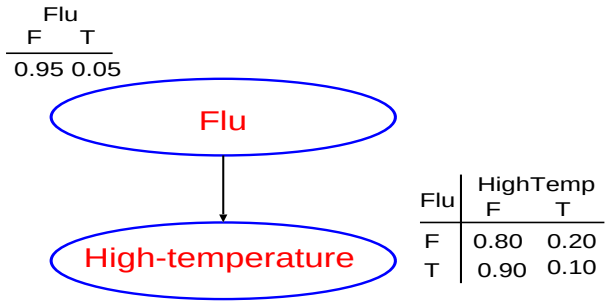


Figure 2: Flu causes high-temperature. Redrawn from Korb et al. (Sorry but without permission.)

3.1.2 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 3.

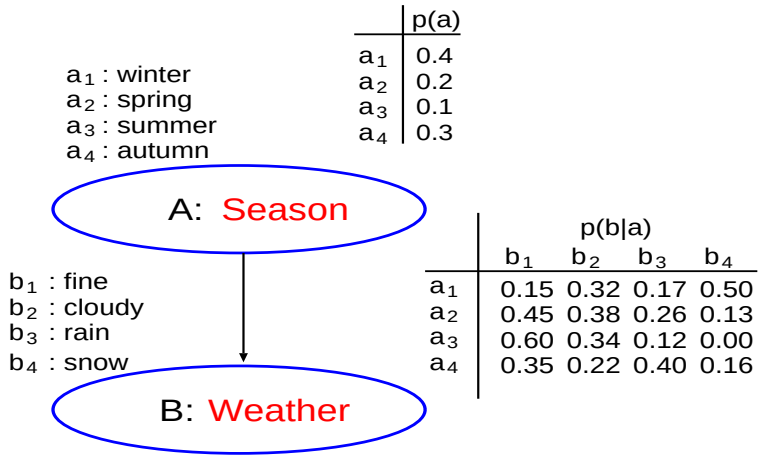


Figure 3: Season & Weather

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

Exercise 7 *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 8 *Now it's snow, then how much is the probability of being autumn now?*

3.1.3 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,

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

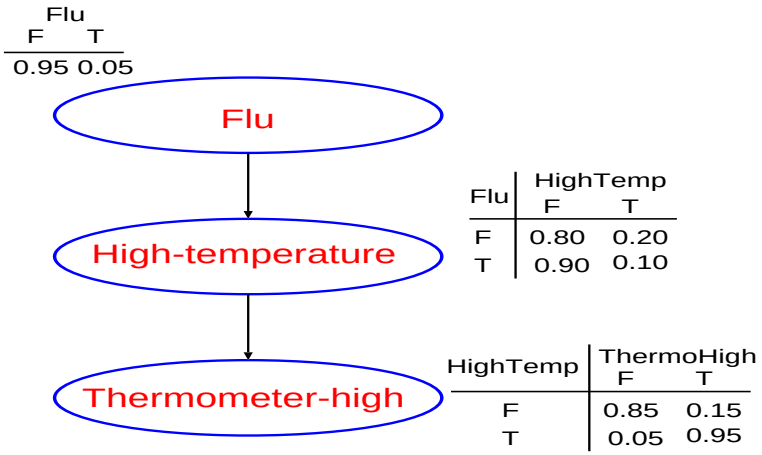


Figure 4: Flu → HighTemp → ThermoHigh. Redrawn from Korb et al. (Sorry but without permission.)

Exercise 9 *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 10 *Now thermometer suggests HighTmp, then how much is the probability of his being Flu?*

3.1.4 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.[2]
The situation is described in the page as:

> *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.*

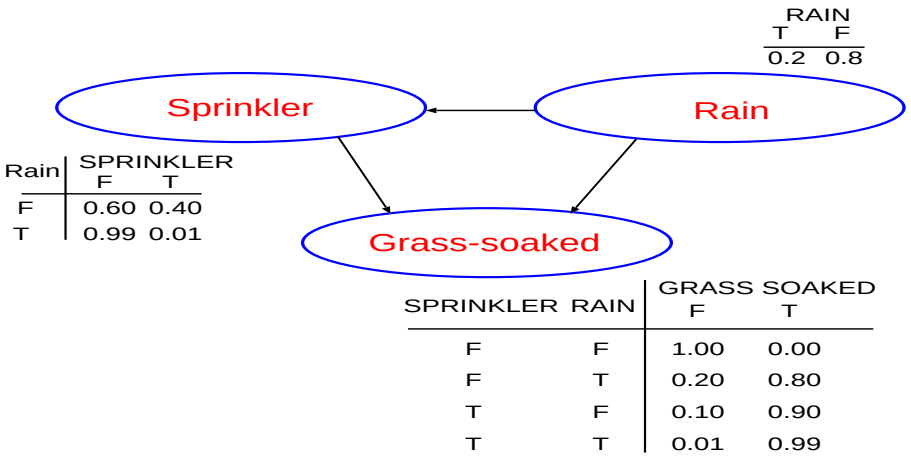


Figure 5: Grass are soaked because it's rain and/or sprinkler?

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

3.1.5 Pearl's earthquake Bayesian Network

This is a very popular example to show how Bayesian network looks like by Pearl.[3]

> *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.*

[2] at *http : //en.wikipedia.org/wiki/Bayesian_network*
[3] Pearl, J. (1988) "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference." San Mateo, Morgan Kaufmann.

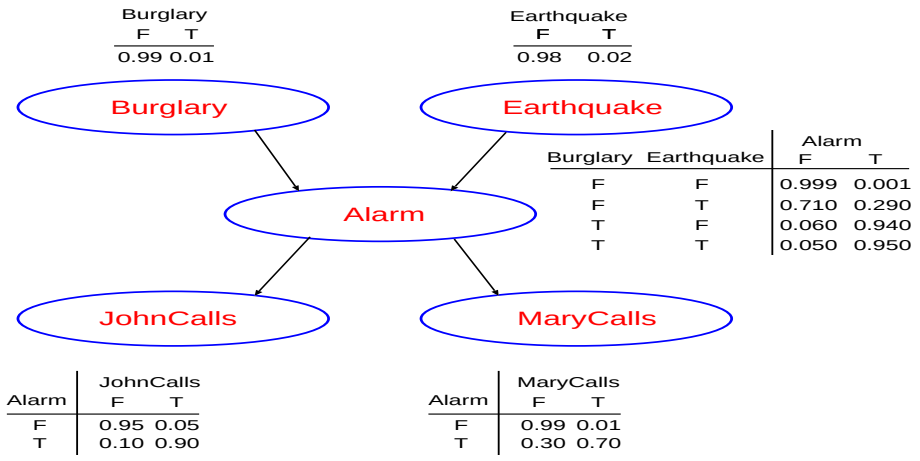


Figure 6: Pearl's Earthquake

3.1.6 Salmon or Sea-bass?

The example of this subsection including three exercises is totally taken, with minor modifications, from Duda et al. [4]

Try to think of the following Bayesian network shown in Figure 7

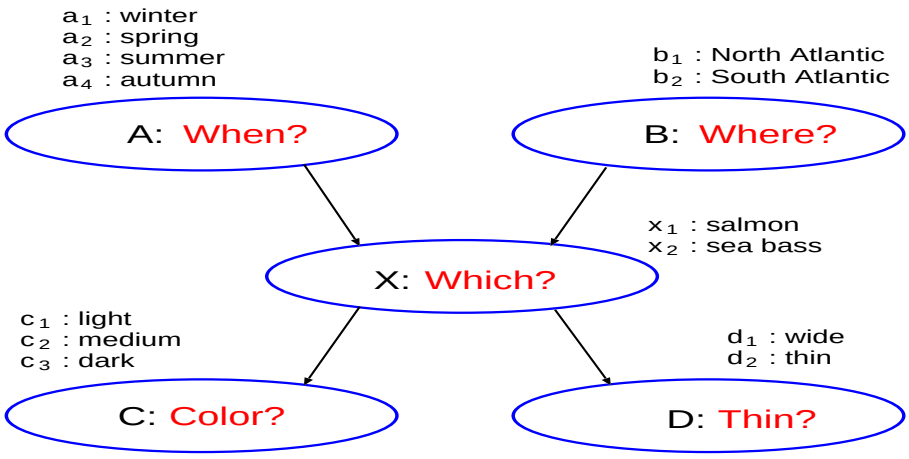


Figure 7: A Bayesian network as to classify the fish caught to Salmon or Sea-bass.

Now the arc of the network, that is, the conditional probability of each of the arcs are given in Figure8.

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

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

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 $p(b_2|e_B) = 0.8$.

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 11 *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 12 *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 13 *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?*

3.2 A formula for inference

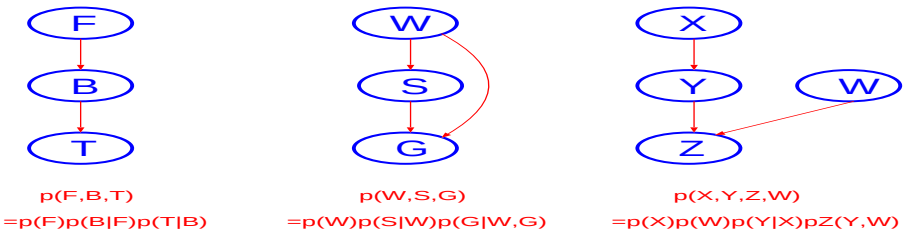
Recall a basic formula in probability theory named joint probability

Rule 3 (Joint Probability) *Assuming we have n nodes of variables $X_1, X_2, \cdots, X_n$. Then we can calculate the joint probability of $X_1, X_2, \cdots, X_n$.*

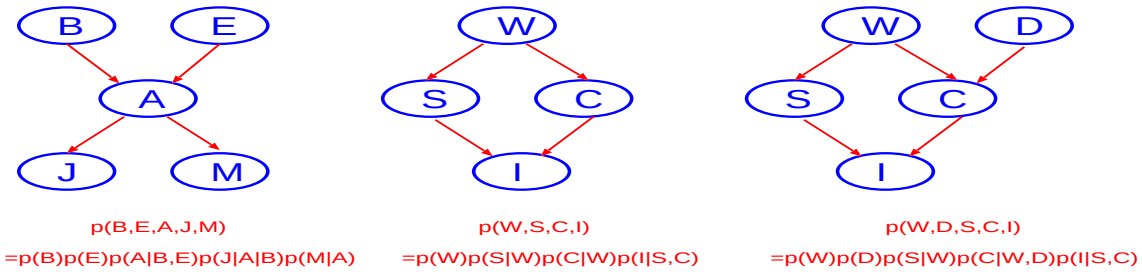
$$p(X_1, X_2, \cdots, X_n) = \Pi_{i=1}^{n} p(X_i | parent(X_i))$$

where $Parent(X)$ means parent node of X.

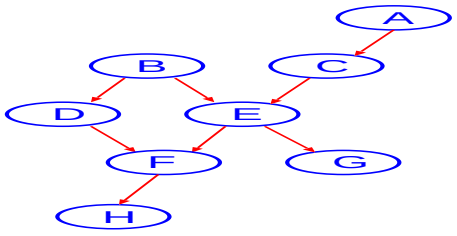
See the following examples.



Let's try more challenging examples.



Exercise 14 *Then what about the following example?*



How to calculate the probability of hypotheses given evidences?

Now question is, wow we calculate the probability of hypotheses given evidences? Assume we have 5 variables A, B, C, D, and E, of which $D = d$ is hypotheses, and $A = a$, and $E = e$ are evidences, just an example. Then what we want to calculate is:

$$p(D = d|A = a, E = e).$$

A basic formula of probability tells us:

$$p(X, Y) = p(X|Y)p(Y).^5$$

So,

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

Hence

$$p(D = d|A = a, E = e) = \frac{p(A = a, D = d, E = e)}{p(A = a, E = e)}.$$

Then,

$$p(A = a, D = d, E = e) = \sum_{B,C} p(A, B, C, D, E).$$

As we already obtained $p(A, B, C, D, E)$ we can calculate this. And similarly,

$$p(A = a, E = e) = \sum_{B,C,D} p(A, B, C, D, E),$$

where, for example,

$$\sum_{B,C,D} p(A, B, C, D, E)$$

means sum over all possible value of B, C, and D while remain $A = a$ and $E = e$. Now let us take a further concrete example *Weather-Sprinkler-GrassWet* where we have already learned

$$p(R, S, G) = p(R)p(S|R)p(G|S, R).$$

Assume now we want to know the probability of hypotheses "It's rain" under the evidence of "Grass is wet," for example. Then

$$p(R = true|G = true) = \frac{p(R = true, G = true)}{p(G = true)} = \frac{\sum_S p(R = true, S, G = true)}{\sum_{R,S} p(R, S, G = true)}$$

For the sake of simplicity, let's denote $p(R = true, S = true, G = true) = TTT$, $p(R = true, S = true, G = false) = TTF$, $p(R = false, S = false, G = true) = FFT$, and so on, then the above equation can be described as

$$= \frac{TTT + TFT}{TTT + TTF + TFT + TFF}$$

Now we can calculate the probability value, can we not?

[5] or, $p(X, Y) = p(Y|X)p(X)$ depending on the situation.

4 Bayesian network for decision making

4.1 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.

4.1.1 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.

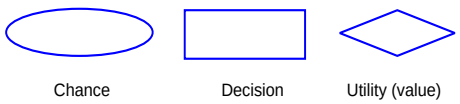


Figure 9: Symbol to express BDN – Chance, Decision, and Utility

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

Clares football team, Melbourne, is going to play her friend Johns team, Carlton. John offers Clare a friendly bet: whoevers 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 teams 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.

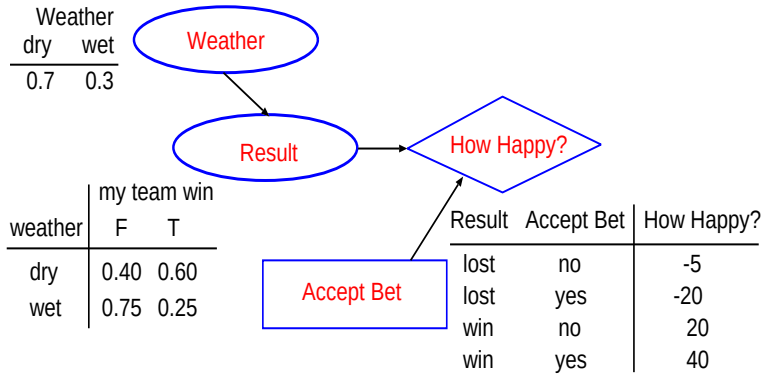


Figure 10: To bet or not to my football team?

Algorithm 1 *Decision network evaluation with a single decision node:*

1. *For each action value in the decision node:*
   (a) *Set the decision node to that value;*
   (b) *Calculate the probability for the parent nodes of the utility node;*
   (c) *Estimate the utility for the action by selecting variables one by one from these parent nodes: such that if the variable is evidence then select it, otherwise the highest probability variable in the current node. From the utility table, We can decide the most favorable action for these selected variables of the parent nodes.*
2. *Return the action.*

4.2.1 Information links

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

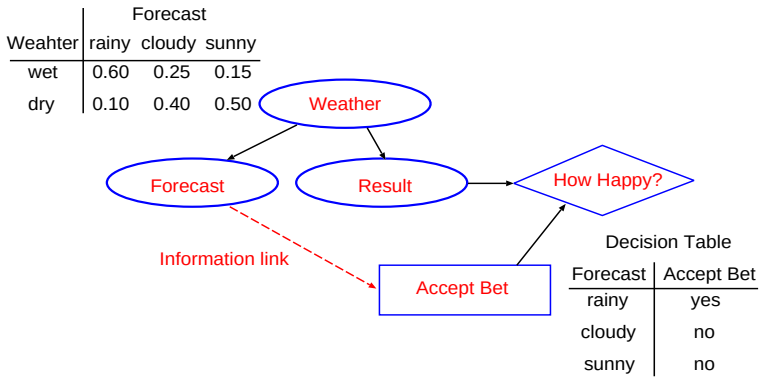


Figure 11: An example of information link.

Algorithm 2 *Decision network evaluation with multiple decision nodes:*

1. *For each combination of values of the parents of decision node:*
   - (a) *For each action value in the decision node:*
     - i *Set the decision node to that action.*
     - ii *Calculate the posterior probabilities for each of the parent nodes of the utility node from one parent node to the next.*
     - iii *Record the utility value corresponding to the combination of the highest probability set of parent nodes of the utility node calculated in (ii).*
   - (b) *Record the action with the highest utility value for the action in the decision table.*
3. *Return the decision table.*

4.2.2 An example of decision making table

(Allow me to skip this.)

4.3 Sequential decision making

4.3.1 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.

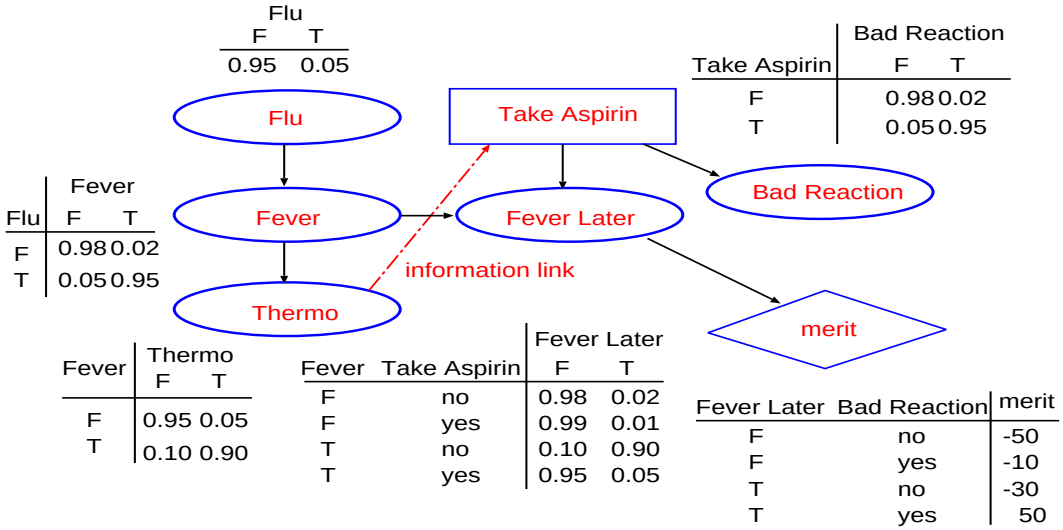


Figure 12:

4.3.2 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.

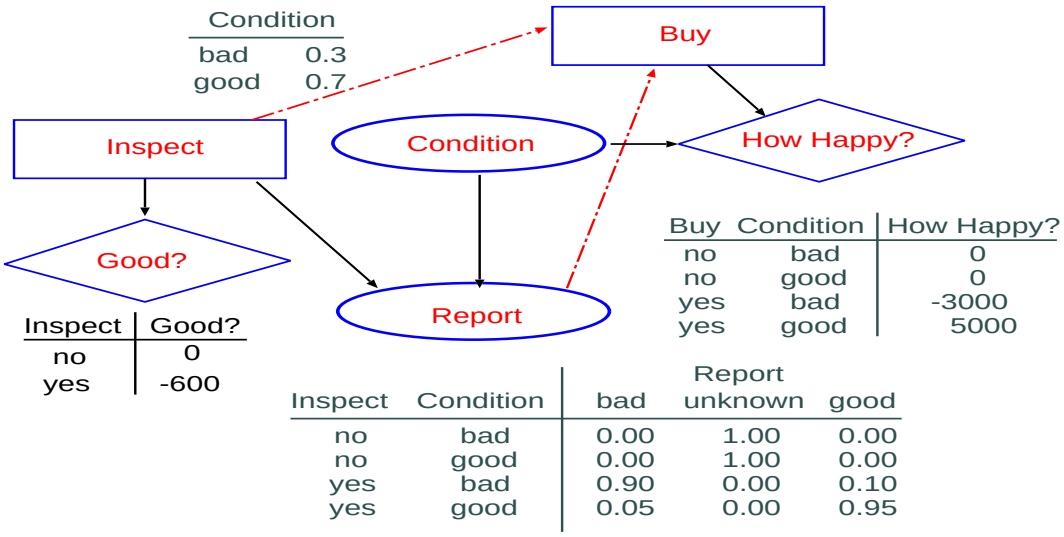


Figure 13: Revised real-estate example.

4.4 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 $X_i(T) \rightarrow X_j(t+1)$.

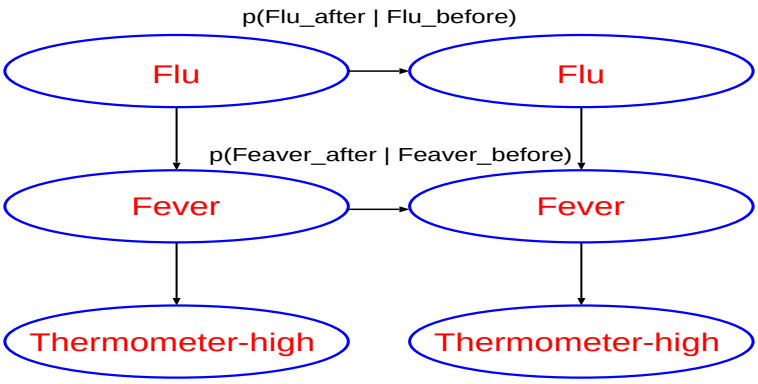


Figure 14: Revised Flu example as a simple case of Dynamic Bayesian Network.

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*.

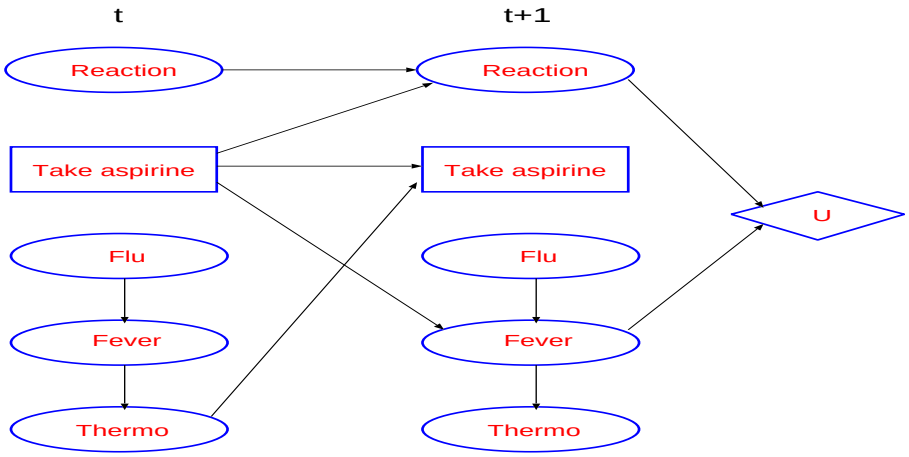


Figure 15: Still simple but more realistic Dynamic Bayesian Network.

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*.

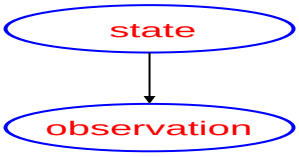


Figure 16: Node "*state*" and node "*observation*"

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.

4.4.1 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.

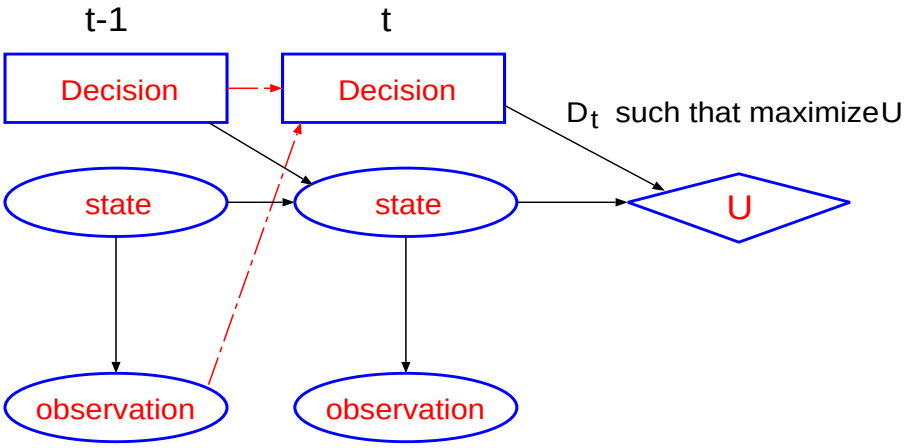


Figure 17: A Dynamic Bayesian Network for decision making.

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.

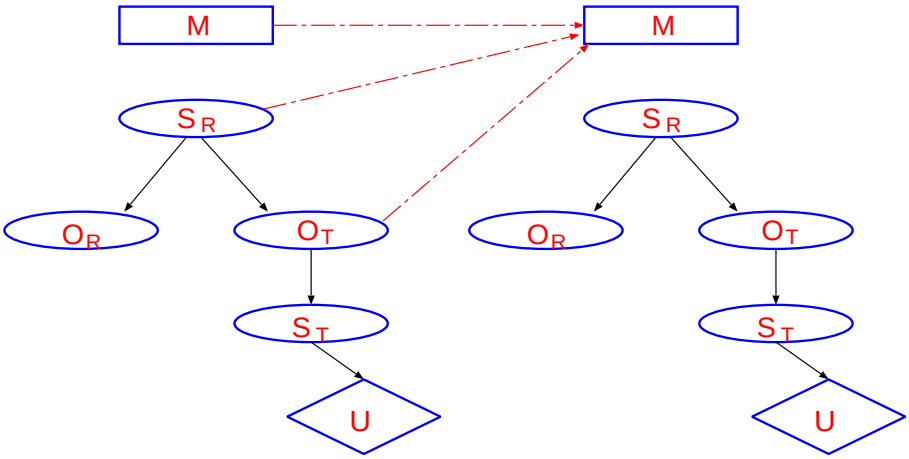


Figure 18: Dynamic Bayesian Network for a mobile robot.

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

I. 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 $\mu = (0, 0)$.

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

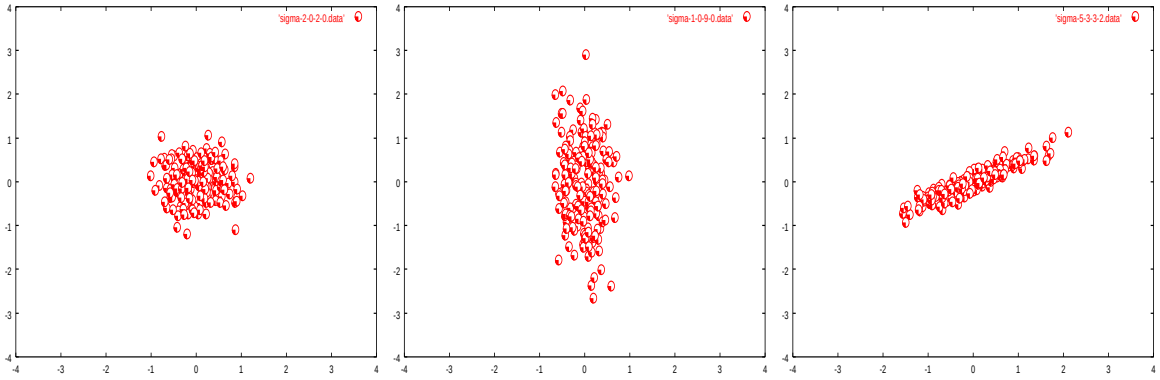


Figure 19: A cloud of 200 Gaussian random points with three different three Σ.

II. 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 **??**.

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

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

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