Pattern Classification

for
postgraduates

(9 weeks from 21 February to 17 April, 2007)

Akira Imada
Brest State Technical University, Belarus

(last modified on)

December 4, 2007

INDEX

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 ones and male ones. The situation is we have thousands of salmons and our task is classify salmon not yet seen 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 so far we have were 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 so far we have were female and male" are denoted as $P(\omega_1)$ and $P(\omega_2)$, respectively, and called *prior probability.*

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.

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 acount that the risk of misclassify x into ω_2 while x is ω_2 is important[1].

In the example of Fig. 1.3, you might imagine *a classification fo 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.

[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.

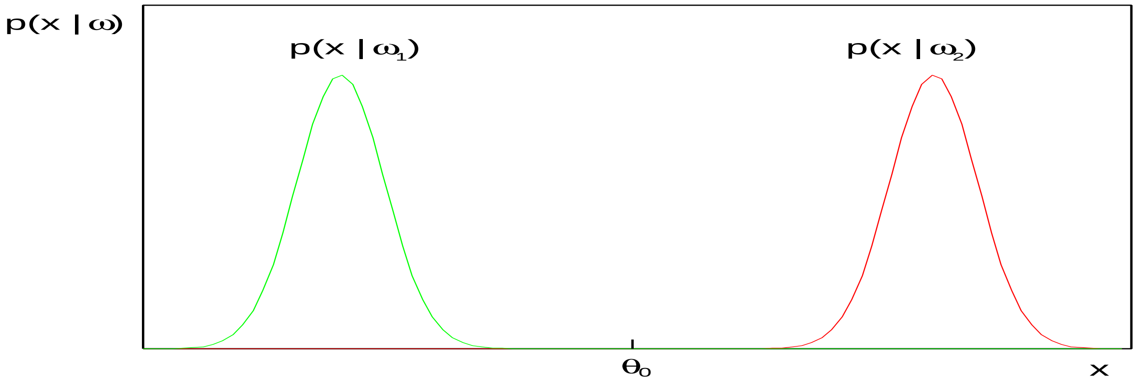


Figure 1: An example of two distribution of the observation x w.r.t. each class of ω_1 and ω_2. Here reader might image that distribution of length of female salmon and male sarmon.

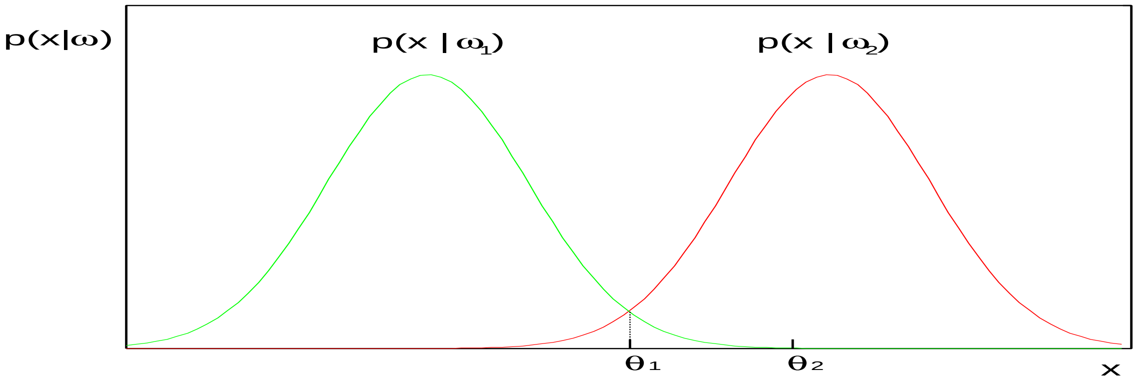


Figure 2: Yet another example where two distributions are closer than the first example. Two threshold are shown, considering a risk of miss-classifcation.

2 What is Bayesian Classification?

Assume we should guess the class that will come next with the knowledge of

(1) Prior probabilities of class-1 ω_1 and class-2 ω_2, that is, $P(\omega_1)$ and $P(\omega_2)$.

(2) Class conditional probability of the observatin of x of each of the two classes, that is, $p(x|\omega_1)$ and $p(x|\omega_2)$.

Then we can calculate $P(\omega_i)|x)$ — the probability that class is ω_i under the condition where x is observed as:

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)} \quad (i = 1, 2), \tag{1}$$

where

$$p(x) = p(x|\omega_1)P(\omega_1) + p(x|\omega_2)P(\omega_2) \tag{2}$$

which is sometimes called *evidence*. Then our rule is:

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

2.1 Bayesian Decision Theory

Assume now we have multiple *features*:

$$x_1, x_2, \cdots, x_d$$

mulitple *classes*:

$$\omega_1, \omega_2, \cdots, \omega_c$$

also *actions*:

$$\alpha_1, \alpha_2, \cdots, \alpha_a$$

instead of guessing which class. Then loss function $\lambda(\alpha_i|\omega_j)$ is defined as the loss when we take an action $alpha_i$ when the situation is ω_j).

If we take an action α_i when we observe $\mathbf{x}$ while the true situation is ω_j, the conditional *risk* is defiened as

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)p(\omega_j|\mathbf{x}). \tag{3}$$

Note that we want to take the action so as to minimize this risk.

Under continuous feature $\mathbf{x}$ we have observed, assume c finite categories: $\omega_1, \omega_2, ...\omega_c$ and a finite possible actions $\alpha_1, \alpha_2, ...\alpha_a$. Then a loss function $\lambda(\alpha_i|\omega_j)$ is defiened as the loss caused by taking an action α_i when ω_j is the situation.

We now assume that we take action α_i when we observe $\mathbf{x}$ while the true situation is ω_j the loss function $\lambda(\alpha_i|\omega_j)$ will be defiend. Then expected loss will be

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}) \tag{4}$$

which called a *risk*. Recall here

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})} \tag{5}$$

where

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j). \tag{6}$$

2.2 Two-category classification

Let's simplify as $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$, that is,

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}, \tag{7}$$

Then

$$R(\alpha_1|\mathbf{x}) = \lambda_{11} P(\omega_1|\mathbf{x}) + \lambda_{12} P(\omega_2|\mathbf{x})$$

$$R(\alpha_2|\mathbf{x}) = \lambda_{21} P(\omega_1|\mathbf{x}) + \lambda_{22} P(\omega_2|\mathbf{x}) \tag{8}$$

Here we have:

Rule 3 (Rule for classification) *If $R(\alpha_1|\mathbf{x}) < R(\alpha_1|\mathbf{x})$ then classify $\mathbf{x}$to ω_1 otherwise to ω_2.*

2.2.1 likelihood ratio

If we further assume that we have one feature x instead of multiple $\mathbf{x}$, the if part of the rule above will be

$$(\lambda_{21} - \lambda_{11}) P(\omega_1|x) > (\lambda_{12} - \lambda_{22}) P(\omega_2|x). \tag{9}$$

And usually we can assume by definition that $\lambda_{21} > \lambda_{21} >$ and $\lambda_{21} > \lambda_{21} >$, the rule will be:

Rule 4 (Rule for classification) *If the following holds*

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} \tag{10}$$

then classify x to ω_1 otherwise to ω_2.

Note that the left-hand side is a function of x and is called a *likelihood ratio* and the right-hand side is a quantity and is a *threshold* to discriminate two classes.

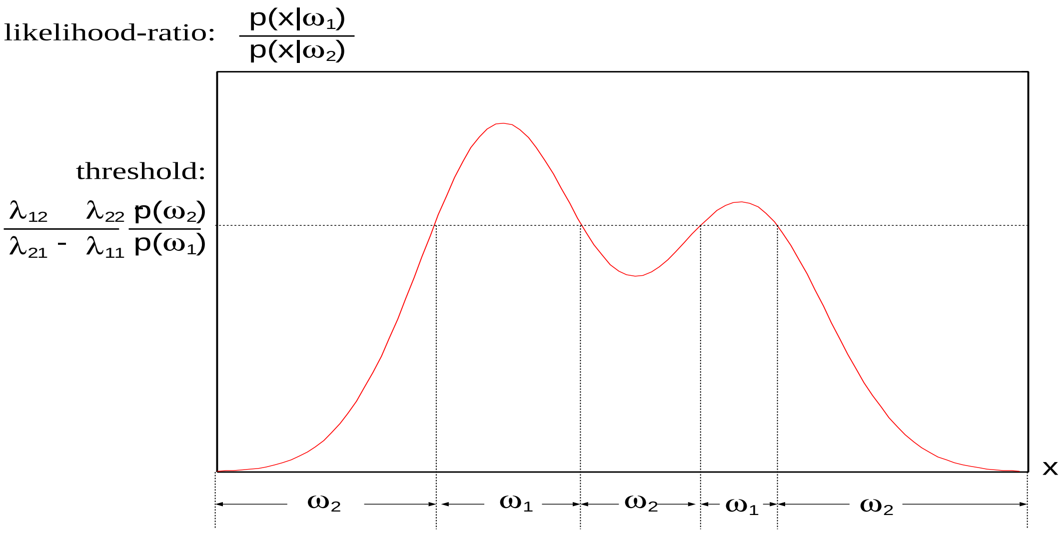


Figure 3: A fictitious plot of likelihood-ratio vs. x. Try to notice which class we should assign according to the value of x with a smaller risk.

2.3 Discriminant function

Let's define a function, not probability or not density but just a function for the purpose of discriminate the space in to categories.

3 Examples of Decision Surface

3.1 1-D Gaussian case.

We now assume

$$p(x|\omega_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\{-\frac{1}{2}\frac{(x-\mu_1)^2}{{\sigma_1}^2}\}$$

and

$$p(x|\omega_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\{-\frac{1}{2}\frac{(x-\mu_2)^2}{{\sigma_2}^2}\}.$$

Hence, one-dimensional version of our Rule-2 is, *If $P(\omega_1|x) > P(\omega_2|x)$ then classify x to ω_1 otherwise ω_2.*

So, recalling

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{P(x)}$$

and $P(x)$ is common for both $i = 1$ and $i = 2$, the *IF-part* of the above rule is

$$\frac{1}{\sqrt{2\pi}\sigma_1}\exp\{-\frac{1}{2}\frac{(x-\mu_1)^2}{{\sigma_1}^2}\}P(\omega_1) > \frac{1}{\sqrt{2\pi}\sigma_2}\exp\{-\frac{1}{2}\frac{(x-\mu_2)^2}{{\sigma_2}^2}\}P(\omega_2)$$

Now that we see all the terms in the equation are positive, the comparison holds true if we take logarithm based on e on both-hands of the equation.[2] Therefore

$$\ln\frac{1}{\sqrt{2\pi}} + \ln\frac{1}{\sigma_1} - \frac{1}{2}\frac{(x-\mu_1)^2}{{\sigma_1}^2} + \ln P(\omega_1) > \ln\frac{1}{\sqrt{2\pi}} + \ln\frac{1}{\sigma_2} - \frac{1}{2}\frac{(x-\mu_2)^2}{{\sigma_2}^2} + \ln P(\omega_2)$$

Excersize 1 *Obtain the decision boundary x_0 when the two classes follow the Gaussian distributions with $N(1, \frac{1}{2})$ and $N(3, \frac{1}{2})$ respectively.*

3.2 2-D Gaussian case

Recalling Baysian formula Eq. (1)

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(x)} \quad (i = 1, 2),$$

where $p(\mathbf{x}) = p(\mathbf{x}|\omega_1)P(\omega_1) + p(\mathbf{x}|\omega_2)P(\omega_2)$ is just a term to normalize $P(\omega_i|\mathbf{x})$ to 1. Further, we assume $P(\omega_1) = P(\omega_1) = 1/2$ just for a sake of simplicity here. So our decision rule to determine which class of ω_1 or ω_2 should be assigned to $\mathbf{x}$:

"If $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$ then $\mathbf{x}$ is ω_1, otherwise ω_2"

just depends on $p(\mathbf{x}|\omega_i)$ which we assume to be Gaussian p.d.f. That is

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t{\Sigma_i}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\} \tag{11}$$

---

[2] That is,

$$A > B \Leftrightarrow \ln(A) > \ln(B).$$

3.2.1 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 Eq. (11) 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 discriminat 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 $g_1(\mathbf{x}) - g_2(\mathbf{x}) = 0$.

$$(\frac{1}{a_1} - \frac{1}{a_2})x^2 + \frac{2}{a_2}x + (\frac{1}{b_1} - \frac{1}{b_2})y^2 = \frac{1}{a_2} + \ln\frac{a_1 b_1}{a_2 b_2} \tag{12}$$

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.

3.2.2 Examples

Let's try some calculations under the above assumptions, that is,

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

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

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

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

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

The example the below is somewhat tricky. I wanted an example in which the right-hand side of the equation (12 becames 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 is 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) $5(x+1)^2 + (8/3)y^2 = 10 - \ln(10/3)$

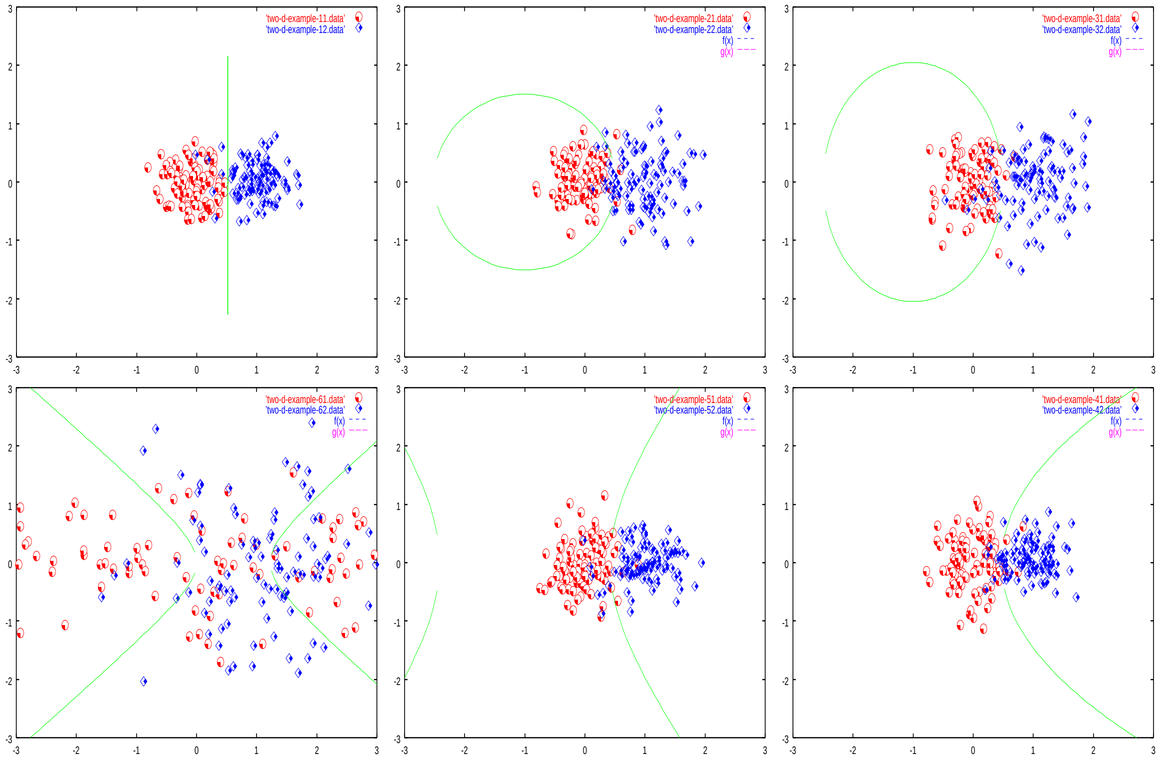


Figure 4: A cloud of 100 points each extracted from a set of two classes and border of the two classes calculated on six different conditions. (Results of (5) and (6) are still fishy and under another trial.)

(4) $5(x+1)^2 - (10/3)y^2 = 10$

(5) $20x - 5y^2 = 10 - \ln 2$

(6) $(1 - 1/2e)x^2 - x - y^2 = 0$

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 $\boldsymbol{\mu}_2 = (1,1)$, and we both classes share the same Σ:

$$(7) \quad \Sigma_1 = \begin{pmatrix} 2.0 & 0.5 \\ 0.5 & 2.0 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 5.0 & 4.0 \\ 4.0 & 5.0 \end{pmatrix}$$

3.3 3-D Gaussian case

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

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

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

$$g_i(\mathbf{x}) \;=\; -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \tag{13}$$

leads to?

Since we calculate (see APPENDIX III.)

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

Now our discriminat 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 caluculation 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 cancelled 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.

3.4 A Higher order Gaussian case

The Equation

$$g_i(\mathbf{x}) = \ln(p(\mathbf{x}|\omega_i) + \ln P(\omega_i)) \tag{14}$$

still holds, of course. Now let's recall that the Gaussoian pdf is

$$p(\mathbf{x}|\omega) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\} \tag{15}$$

and as such

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}|\Sigma_i| + \ln P(\omega_i) \tag{16}$$

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 hypershpers. 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

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i) \tag{17}$$

or, equivalently

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i) + \ln P(\omega_i) \tag{18}$$

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

$$g_i(\mathbf{x}) = \frac{1}{\sigma^2}\boldsymbol{\mu}_i^t\mathbf{x} - \frac{1}{2\sigma_i^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i) \tag{19}$$

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

$$\frac{1}{\sigma^2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i)^t\mathbf{x} - \frac{1}{2\sigma_i^2}(\|\boldsymbol{\mu}_i\|^2 - \|\boldsymbol{\mu}_j\|^2) + \ln\frac{P(\omega_i)}{P(\omega_j)} = 0 \tag{20}$$

In this case our classification rule will be

Rule 5 (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 $\mathbf{x}$.*

If we carefully modify Eq. (18) we will obtain

$$\mathbf{w}\cdot(\mathbf{x} - \mathbf{x}_0) \tag{21}$$

Excersize 2 *Derive the Eq. (21) specifying $\mathbf{w}$ and $\mathbf{x}_0$.*

Eq. (21) is the equation which can be interpret as

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

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

This condition implies that the patterns in each of both classes distribute like hyper-elipsoid. Now that our discriminat 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

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{22}$$

and

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln P(\omega_i)/P(\omega_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \tag{23}$$

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

Excersize 3 *Derive* $\mathbf{w}$ *and* $\mathbf{x}_0$ *above.*

So we modify the avove rule to

Rule 6 (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

$$(\mathbf{a}-\mathbf{b})^t\Sigma^{-1}(\mathbf{a}-\mathbf{b}) \tag{24}$$

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:

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

- **When 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

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

Then, we get the following renewed discriminant function

$$g_i(\mathbf{x}) = \mathbf{x}^tW_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + w_{i0} \tag{25}$$

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) yyper-ellipsoid; (v) hyper paraboloid; (vi) hyper-hyperboloid.

3.4.1 2-D cases revisit

4 Parameter Estimation

Goal is to know $p(\omega_i|\mathbf{x})$.
Assume now that we know the density functin for some reason to believe. What we don't know is the parameter of the density function.

To get a bird's eye view, we name a few of 1-D examples, besides Gaussian, of such density functions which need to be specified by some parameters.

- Uiform-distribution,
  - $p(x) = 1/\theta$ ... if $x > 0$ otherwise 0
- Exponential
  - $f(x) = \theta \exp(-\theta x)$ ... if $x > 0$ otherwise 0
- Rayleigh (See Figure below.)
  - $f(x) = 2\theta x \exp(-\theta x^2)$ ... if $x > 0$ otherwise 0
- Poisson
  - $f(x) = (\theta^x/x!) \exp(-\theta)$ ... for $x = 0, 1, 2, \cdots$
- Bernoulli
  - $f(x) = \theta^x(1-\theta)^{1-x}$ ... for $x = 0, 1$
- Binomial
  - $f(x) = (m!/x!(m-x)! \cdot \theta^x(1-\theta)^{m-x}$ ... for $x = 0, 1, \cdots, m$

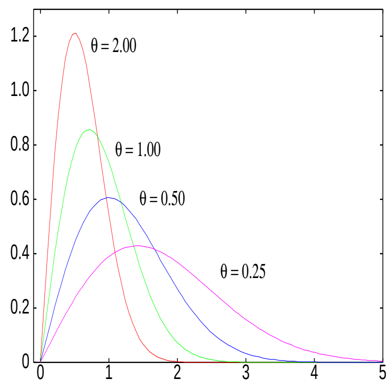


Figure 5: Rayleigh distribution function with four different values of θ.

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.

4.1 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 *0 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* $x_1 = 7$*,* $x_2 = 1$*,* $x_3 = 4$ *and* $x_4 = 2$*. Guess what distribution function of these 4 data follows?*

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

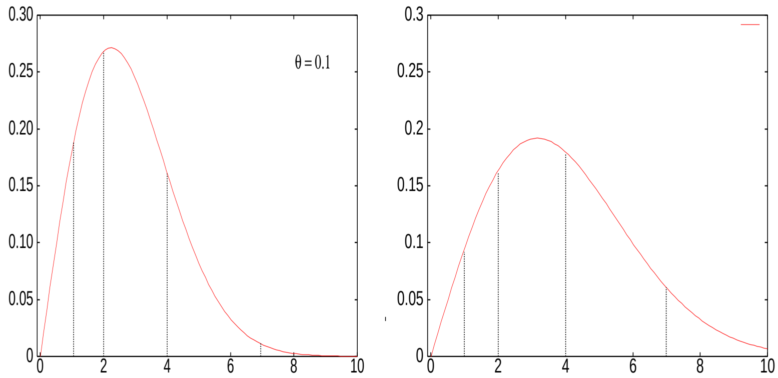


Figure 6: A cloud of 200 points each with different three different $N[\mu, \Sigma]$.

Excersize 4 *(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 randamly from* $N(1, 3)$*, and plot* $p(D|\mu)$ *as a function of* μ *in each of the three cases.*

Rule 7 (An assamption of independence of data) *Take θ which maximizes*

$$P(D|\theta) = \prod_{k=1}^{n} P(x_k|\theta) \tag{26}$$

In short, choose θ which maximizes

$$P(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta) \tag{27}$$

Then let's calculate our previous example of one-dimensional four data $D = \{2, 3, 5, 8\}$ by changign θ from 0.00 to 1.00 with an interval of 0.01. The results are shown the left-most graph of the Fig. 4.1. 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. Important thing is is

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

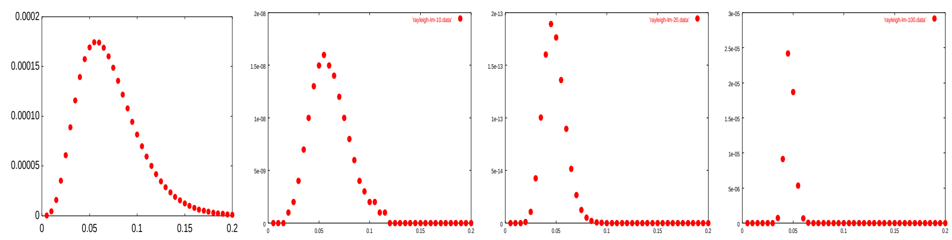


Figure 7: A cloud of 200 points each with different three different $N[\mu, \Sigma]$.

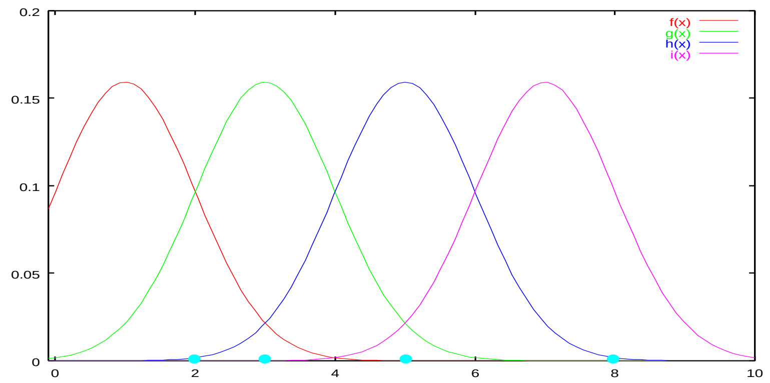


Figure 8: A cloud of 200 points each with different three different $N[\mu, \Sigma]$.

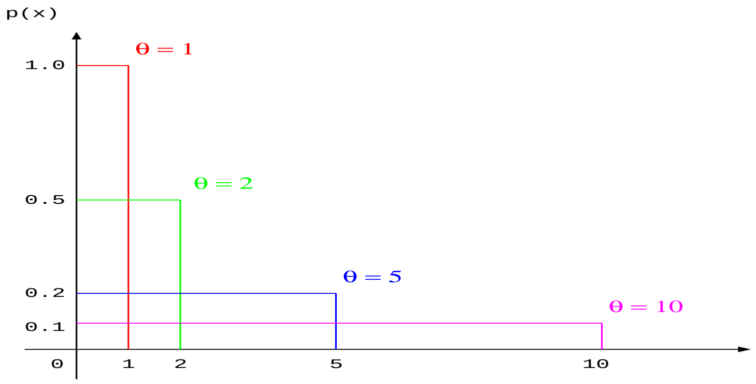


Figure 9: A cloud of 200 points each with different three different $N[\mu, \Sigma]$.

4.2 Bayesian parameter estimation

Since we view here the parameter $\boldsymbol{\theta}$ as random variables with an information of prior distribution $p(\boldsymbol{\theta})$, we view $P(\boldsymbol{\theta}|D)$ as a function of $\boldsymbol{\theta}$. As before, we assume

$$P(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta) \tag{28}$$

where we denote the training data

$$D_n = \{x_1, x_2, \cdots, x_n\}$$

which implies

$$P(D_n|\theta) = P(x_n|\theta)P(D_{n-1}|\theta)$$

Therefore the Baysean Formula

$$p(\theta|D) \approx P(D|\theta)P(\theta). \tag{29}$$

leads us like

$$\begin{aligned} P(\theta|D_n) &= P(D_n)|\theta)P(\theta) \\ &= P(x_n|\theta)P(D_{n-1})|\theta)P(\theta) \\ &= P(x_n|\theta)P(\theta|D_{n-1}) \end{aligned}$$

5 Non-paremetric density estimation

Our assumption: so far are

(1) density function $p(\mathbf{x}|\omega)$ is known

(2) high dimensional density function can be represented as the product of one-dimensional function

Now let's make a more realistic assamption, "We don't know the distribution of samples, i.e., we don't know the form of the density function. Let's estimate $p(\mathbf{x}|\omega)$ from our sample patterns. Or, let's estimate $P(\omega|\mathbf{x})$ directly from our sample patterns. The typical example is *Nearest neighbor rule.*

6 Unknown Density function Estimation

One of the most elementary formula in probability theory tells us the probability that $\mathbf{x}$ will fall in region R will be

$$P = \int_R p(\mathbf{x})d\mathbf{x} \tag{30}$$

Recall, if it's in 1-dimensional case, the probability that x will fall in the region $[x_1, x_2]$ will be

$$P = \int_{x_1}^{x_2} p(x)dx$$

The probability of k samples out of n samples of $\{x_1, x_2, \cdots, x_n\}$ will fall in the region R is

$$\binom{n}{k} P^k (1-P)^{n-k} \tag{31}$$

So, we can assume

$$k = nP \tag{32}$$

holds. On the other hand, Eq. (30) can be

$$P = p(\mathbf{x}) = V \tag{33}$$

where V is the volume of region R. This is true on condition that V is small enough. If you want to think of in 1-dimensional space this is

$$p(x) \cdot (x_2 - x_1)$$

on condition that $x_2 - x_1 \approx 0$. Hence, we now have a relation

$$p(x) = \frac{k/n}{V}. \tag{34}$$

Excersize 5 *Calculate the Eq (31)above in each case of, say,* $n = 4, 20, 100$ *and plot the probability as a function of* k/n *assuming real* P *is, say, 0.6. Make sure the larger the value of* n *the sharper the peak.*

Now in order to estimate the density $p(\mathbf{x})$ at $\mathbf{x}$, let's creat a sequence of region $R_1, R_2, R_3, \cdots$ each of which is a region where estimation is made with n samples.

Now our fundamental Equation becomes

$$p_n(\mathbf{x}) = \frac{k_n/n}{V_n} \tag{35}$$

The condition for this to converge is

$$\lim_{n\to\infty} k_n = \infty \tag{36}$$

$$\lim_{n\to\infty} V_n = 0 \tag{37}$$

$$\lim_{n\to\infty} k_n/n = 0 \tag{38}$$

To realize these conditons we have basically two approaches. In the one approach, we shring $V - n$ as n, the number of training samples, increases, for example, according to $V_n = V_0/\sqrt{n}$. In the other approach, we enlarge V_n as n increases, for example $k_n = k_0\sqrt{n}$. The former is called *Parzen-window* approach and the latter k_n *Nearest Neighbor* approach, both of which we are going to exploring in the next two subsections.

6.1 Parzen-Window classifier

First of all, we define new function

$$\varphi(\mathbf{u}) = \begin{cases} 1 & \text{if} \quad -\frac{1}{2} < u < \frac{1}{2} \\ 0 & \text{if} \quad \textit{otherwise} \end{cases} \tag{39}$$

Then we get

$$\varphi(\mathbf{x} - \mathbf{x}_i/h_n) = \begin{cases} 1 & \text{if} \quad x_i \text{ falls within the hypercube} \\ 0 & \text{if} \quad \textit{otherwise} \end{cases} \tag{40}$$

where the hypercube is locate such that its center at $\mathbf{x}$ and the size of all hedges is h_n. Here we can put k_n as

$$k_n = \sum_{i=1}^{n} \varphi(\mathbf{x} - \mathbf{x}_i/h_n) \tag{41}$$

It's important to note that this is not a function of $\mathbf{x}$ at this moment. Now recalling our goal is to obtain the density $p(\mathbf{x})$ at $\mathbf{x}$, we are happy to obtain the following equation.

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}) \tag{42}$$

Excersize 6 *Assumming $p(x) = U(0,1)$, try to find what happens if wu use $\pi_n(x)$ as Parzen window function. Draw $p_n(x)$ in each case of $h_n = 1, 1/4, 1/16$.*

6.1.1 A Neural Network Installation

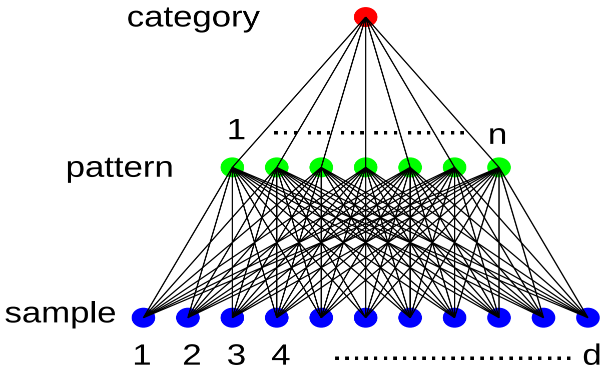


Figure 10: A cloud of 200 points each with different three different $N[\mu, \Sigma]$.

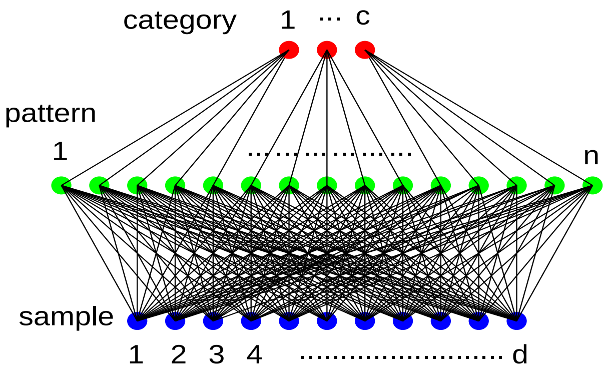


Figure 11: A cloud of 200 points each with different three different $N[\mu, \Sigma]$.

6.2 Nearest-Neighbor classifier

7 Baesian Blief Network

We now assume we have N variables like our previous example of salmon.

Salmonfemale, male
Season spring, summer, autumn, winter
Locationsea, river, lake
Lengthshort, midium, long

Or in more general form

$$A(a_1, a_2, \cdots)$$

$$B(b_1, b_2, \cdots)$$

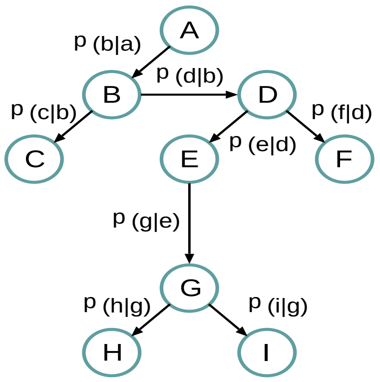
$$C(c_1, c_2, \cdots)$$

and we also know, for example,

$$p(\mathbf{a}|\mathbf{b})$$

$$p(\mathbf{b}|\mathbf{c})$$

etc... then we can fepresent this by a kind of network like:

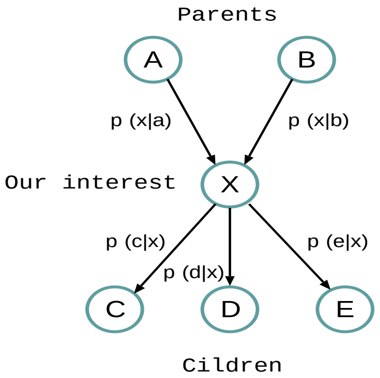


We call this network *Bayesian Blief Network*. Assumption here is we know all ...
The aim is

- To classify a new data $\mathbf{x}$.
- To know the probability of an unknown variable from other known variables.

We now take $\mathbf{x} = (x_1, x_2, \cdots)$ as a *belief* of $x_1, x_2, \cdots$ under *evidences* $\mathbf{e} = (e_1, e_2, \cdots)$ where some of the e_i belong to parents variable and the others to children. Then the elementary calculation is

$$p(\mathbf{x}|\mathbf{e}_{\text{all}}) = p(\mathbf{e}_{\text{children}}|\mathbf{x}) \cdot p(\mathbf{x}|\mathbf{e}_{\text{parents}}) \tag{43}$$



where

$$p(\mathbf{e}_{\text{children}}|\mathbf{x}) = p(e_{c_1} \cdots e_{c_n}|\mathbf{x}) = p(e_{c_1}|\mathbf{x}) \cdots p(e_{c_1}|\mathbf{x}) = \Pi_i p(e_{c_i}|\mathbf{x}) \tag{44}$$

and

$$p(\mathbf{x}|\mathbf{e}_{\text{parents}}) = p(\mathbf{x}|e_{p_1}, e_{p_2}, \cdots) = \sum p(\mathbf{x}|p_1, p_2, \cdots)p(p_1, p_2, \cdots|e_{p_1}, e_{p_2}, \cdots) \tag{45}$$

where

$$p(p_1, p_2, \cdots|e_{p_1}, e_{p_2}, \cdots) = p(p_1|e_1) \cdot (p_2|e_2) \cdots \tag{46}$$

Like

$$p(x_1) = p(x_1|a_1 b_1)p(a_1)p(a_2) + \cdots$$

For example, in our example of salmon, assume we know all

	female	*male*
winter	0.9	0.1
spring	0.3	0.7
summer	0.4	0.6
autumn	0.8	0.2

(47)

Example 3

(1) A new salmon was caught in winter, at the lake. The size is long and very thin. Then, what is the probability that it is a female salmon?

(2) A new salmon we have is short, thick, and cought in a river. Then what is the most likely season when it was caught?

(3) We have a long, thin salmon caught in winter. What is the probability that it is from sea.

(4) It is caught between autumn and winter from a lake, and we have no information as for the length of the fish, but this is medium thickness. Is the salmon male?

(5) Short and thick which season you guess that the salmon was caught?

(6) Long, thick from a lake what season is mostlikely?

8 Hidden Markov Model

8.1 Markov Model

We now assume that pattern is a sequence of states $x(t)$, something like

$$x(1), x(2), x(3), \cdots, x(T)$$

Here we assume time is descrete like $1, 2, 3, \cdots, T$. Let's denote a sequece of states whose length is T as $\mathbf{x}^T$ Imagine

$$\mathbf{x}^7 = x_1, x_6, x_3, x_3, x_1, x_2, x_4$$

An example of states is a whether. Think of x_1 is fine whether, x_2 is cloudy day, x_3 is a rainy day, and so on. Assume we know the transition probability from any state to any state, like

$$P(x_j(t+1)|x_i(t)) = a_{ij}. \tag{48}$$

We call a full set of these a_{ij} *model*. When we denote this full set of a_{ij} as $\boldsymbol{\theta}$, the probability of above example of sequence $\mathbf{x}^7$ under the model $\boldsymbol{\theta}$ is expressed as

$$P(\mathbf{x}^7|\boldsymbol{\theta}) = a_{16} \cdot a_{63} \cdot a_{33} \cdot a_{31} \cdot a_{12} \cdot a_{24}. \tag{49}$$

Our goal here is to find a model which best explains training patterns of a known class, and later, a test pattern is classified by the model that has the highest posterior probability $P(\omega|\mathbf{x})$.

8.2 Hidden Markov Model

Assume now the state $x(t)$ is not observable, but emits visible symbol $y(t)$. For example,

$$y_5, y_3, y_2, y_2, y_7, y_2, y_3 \tag{50}$$

If we know

$$P(y_k(t)|x_j(t)) = b_{jk} \tag{51}$$

Where $\sum_j a_{ij} = 1$ for all i and $\sum_j b_{jk} = 1$ for all j. This is called a *Hidden Markov Model* because a sequence $x(t)$ is hidden to observers.

Then we have three different problems.

- Evaluation Problem
  - Evaluation of probabilitiy of a particular sequence of visible state.
- Decoding Problem
  - Assuming we know all of a_{ij} and b_{jk}, determinatin of the most likely sequence of hidden states for a given observation of visible sequence.
- Learning Problem
  - Estimation of a_{ij} and b_{jk} from a given set of training obserbations of visible sequences.

9 Classification by Fuzzy Logic

9.1 Basics

Fuzzy logic is an extention of classical *set theory.* While in the classical set theory an *element* is either *belong* to a *set* or not, in fuzzy logic we use a concept of *degree of belongingness* with *degree of membership value* ranging from 0 to 1. For example, how we are sure that x is A is expressed with a function called *membership function* $\mu_A(x)$.

9.1.1 Examples of membership function

Do you like cold beer? However *"cold"* is not a very clear idea. Is 10°C is cald? What about −5°C?

In Fuzzy logic, the idea of "Beer is cald," is expressed with a membership function like

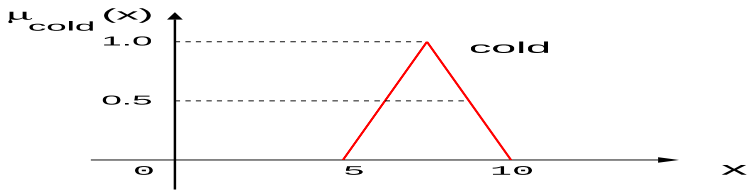


Figure 12: A membership function which expresses Beer is cold.

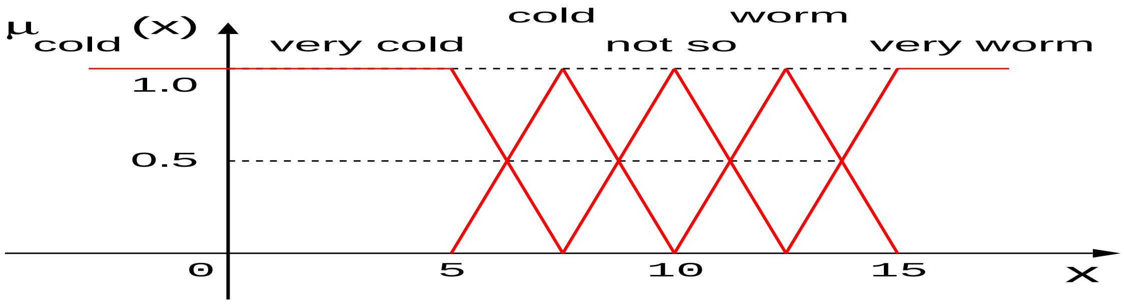


Figure 13: Very cold, cold, not-so cold, worm

Yet another form of membership function is Cauchy Bell-like one which is given by

$$\mu(x) = \frac{1}{1 + (x - 12)^2} \tag{52}$$

as shown in the figure below.

For example this could be a membership function for

All real number close to 12.

The other option will be, of course Gaussian membership function.

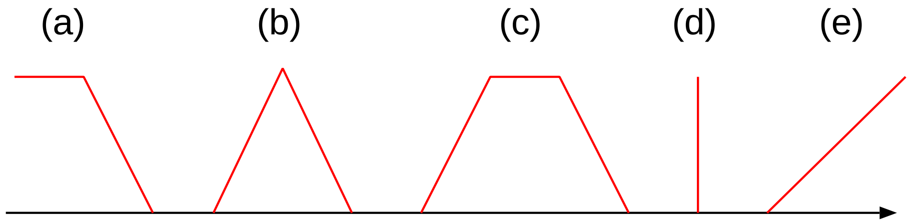


Figure 14: Trapezoid on corner, trapezoid, triangle singleton, traiangle on corner

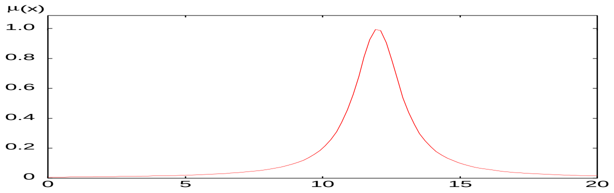


Figure 15: Yet another membership function – Cauchy's belllike function.

9.2 Basic Arethmetic

AuB AcupB Abar

9.2.1 AND & OR

When two membership function have a same domain (which is sometimes called *universe of dicourse*) An example (comparison with traditional set theory) close to 5 and close to 7 close to 5 or close to 7 not close to 5 1) $A \cup B$ 2) $A \cap B$ 3)

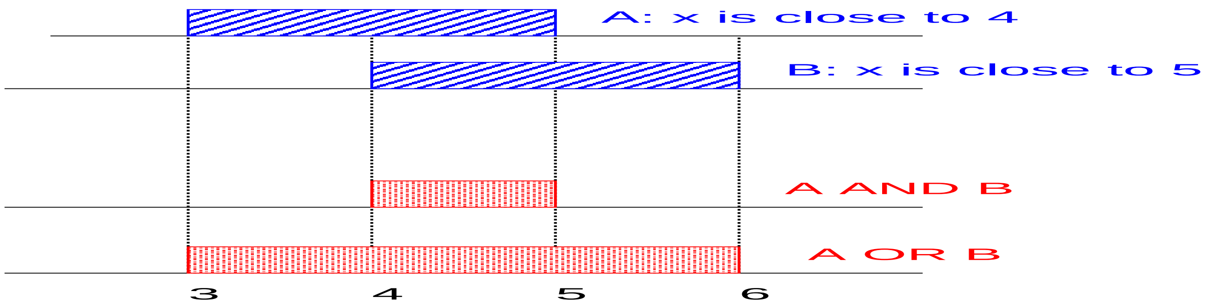


Figure 16:

9.3 when x and y are defiened on different domain

young and tall given membership of each 15-35, 170-180 e.g.

how we represent 1) 3-d 2) gfray scale 3) descretize -¿ matrix 175 180 185 190 15 20 25 30 35

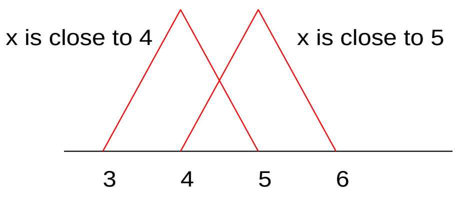


Figure 17:

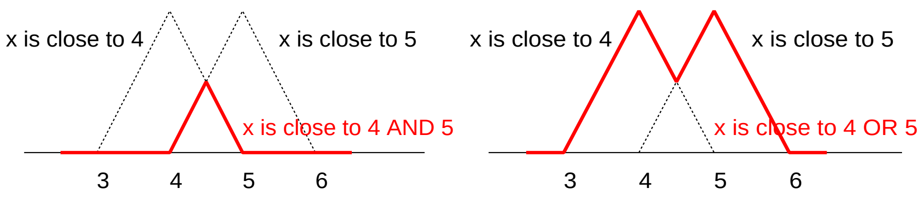


Figure 18:

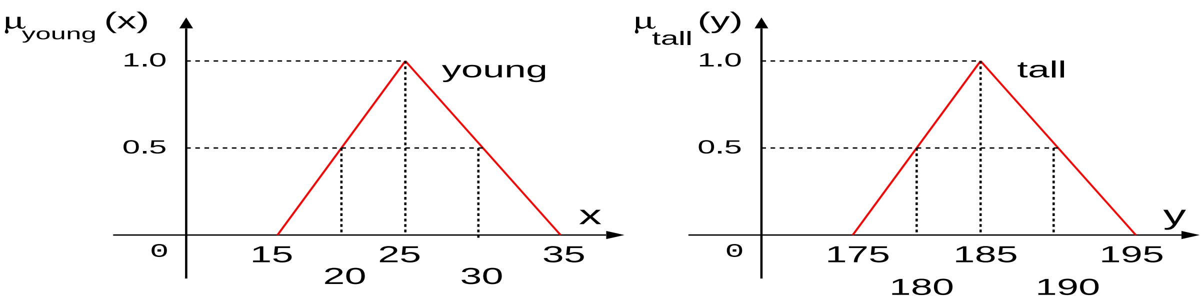


Figure 19: Yet another membership function – Cauchy's belllike function.

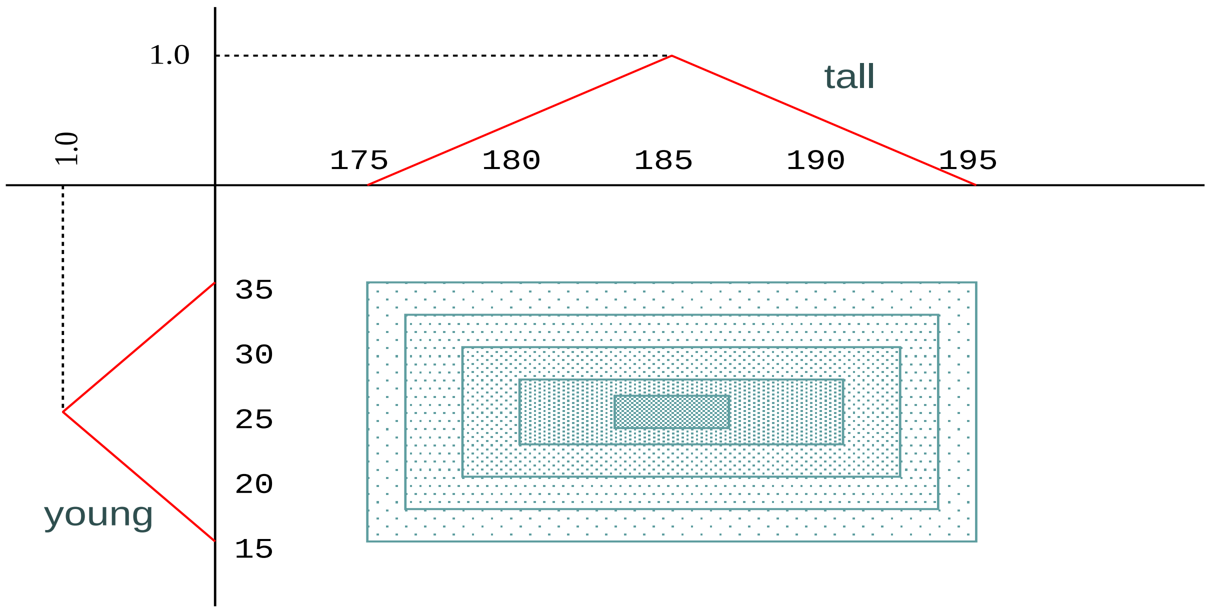


Figure 20:

Standard definition

$$\mu_{AandB} = \min(\mu_A, \mu_B) \tag{53}$$

$$\mu_{AorB} = \max(\mu_A, \mu_B) \tag{54}$$

	A and B	A or B
Yet another standard	$\mu_A \cdot \mu_B$	$\mu_A + \mu_B - \mu_A \cdot \mu_B$
Lukasievicz's	$\max(0, \mu_A + \mu_B - 1)$	$\min(1, \mu_A + \mu_B)$
Einstein's	$\mu_A \cdot \mu_B/(2 - (\mu_A + \mu_B) + \mu_A \cdot \mu_B)$	$(\mu_A + \mu_B)/(1 + \mu_A \cdot \mu_B)$

9.4 Fuzzy expression of IF-THEN rule

It's important to express IF-THEN relation in Fuzzy logic, because one of the goals of fuzzy logic is , in short, to obtain $y = f(x)$ from a set of fuzzy rule. Here we see two such expression.

Mamdani's definition

$$\mu_{A \Rightarrow B} = \min(\mu_A, \mu_B) \tag{55}$$

Larsen's definition

$$\mu_{A\Rightarrow B} = \mu_A \cdot \mu_B \tag{56}$$

9.5 Let's visualize IF-THEN rules

One precondition one rule – simplest case

Let's see an example

$$\text{IF } x = low \text{ THEN } y = weak.$$

Just imagine here, for example, x is a speed of the care I now am driving, and y is strentht of break. We now assume membership function of as follows.

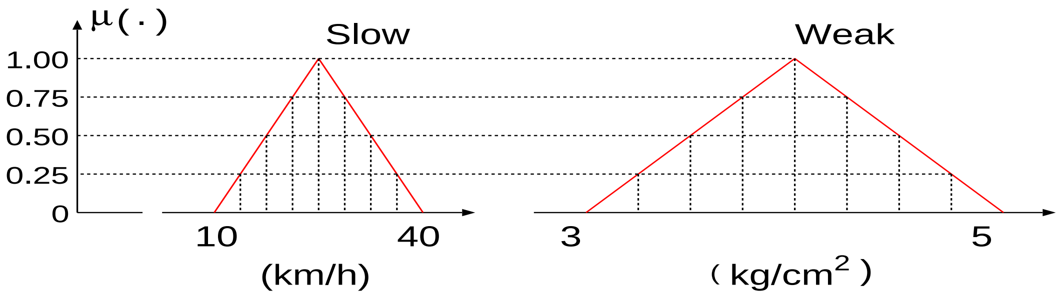


Figure 21:

Excersize 7 *By using Mamdani's expression*

$$\mu_{Rule}(x1, x2) = \min(\mu_{Low}(x1), \mu_{Weak}(x2))$$

Express the membersip fuction $\mu_{Rule}(x1, x2)$.

9.5.1 When precondition is multiple

Let's take an example

$$\text{IF } x_1 = low \text{ and } x_2 = long \text{ THEN } y = midium.$$

In this example, imagine, for example, x_1 is a speed of the care I now am driving, x_2 is a distance between my car and the car in front, and y is strength of break as before. Now $\mu_R(x1, x2) = \min(\mu_{low}(x1), mu_{long}(x2), mu_{medium}(y))$ would be defiend on 3-D space $(x1, x2, y)$ Hence we connot observe the shape of the function. Only when we give x1, and x2 a specific value $\mu(x1, x2, y)$ vs y canbe plotted on 1-D axis of y.

When we have two rules

Example will be

$$\text{Rule 1: IF } x = \textit{low} \text{ THEN } y = \textit{weak}.$$
$$\text{Rule 2: IF } x = \textit{midium} \text{ THEN } y = \textit{midium}.$$

Again, all what we can do is only to plot μ by giving specific value of x Then $\mu_R 1(y)$, and $\mu_R 2(y)$ can be seen as before. By interpreting we have to follow

Rule 1 or Rule 2

and using Mamdani' definition the membership function is

$$\mu_{\text{Rule-1 or Rule-2}} = \max(\mu_{\text{Rule-1}}(x_1), \mu_{\text{Rule-2}}(x_2))$$

Now we plott $\mu(y)$ on y axis.

But why this makes us happy?

9.6 Fuzzy Classification

9.6.1 With a set of fixed membership functions

Assume our rules are like

$$\text{IF } \{x_1 = (S_1 \vee S_5)\} \wedge \{x_2 = S_3\} \vee \cdots \vee \{x_4 = (S_2 \vee S_4)\} \text{ THEN class is } \omega_j.$$

We now think of evolving these rules with chromosom being like

$$((10010)(11000)(00001)(00110))$$

9.6.2 Takagi Sugeno Model

Assume we have H rules. The k-th rule out of these H is like

$$R_k : \text{IF } (x_1 \text{ is } A_1^k) \text{ AND } (x_2 \text{ is } A_2^k) \text{ AND } \cdots \text{ AND } (x_n \text{ is } A_n^k)$$
$$\text{THEN } y_1 \text{ is } \nu_1^k, \cdots, y_m \text{ is } \nu_m^k.$$

where A_i^k is *fuzzy linguistic expression* of i-th attribute like *medium slow* and the membership function is

$$\mu_i^k(x_i) = \exp\{-\frac{(x_i - w_i^k)^2}{\sigma_i^k}\}. \tag{57}$$

Then if a new sample $\mathbf{x}^0$ is given, y_j^0 is known to be given as

$$y_j^0 = \frac{\sum_{k=1}^{H} \mu_k(\mathbf{x}^0) \cdot \nu_j^k}{\sum_{k=1}^{H} \mu_k(\mathbf{x}^0)} \tag{58}$$

where

$$\mu_k(\mathbf{x}^0) = \prod_{i=1}^{n} \mu_i^k(x_i^0). \tag{59}$$

APPENDIX

I. Bayes 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 $p(A) = 1/3$.

You agree that

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

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!

- Now it's clear that
  - $p(A)$ and so on are to be called
    - *a priori* probabilities;
  - and $p(A|D)$ and so on to be
    - *a posteriori* probabilities.

- In the same way,
  - $p(\omega_i)$ is called
    - *a priori* probability[3];
  - $p(\omega_i|\mathbf{x})$
    - *a posteriori* probabilities.

  Furthermore
  - $p(\mathbf{x})$ is called
    - p.d.f. of $\mathbf{x}$
  - $p(\mathbf{x}|\omega_i)$
    - class-conditional p.d.f.[4]
      - which describes the distribution of the feature vectors $\mathbf{x}$ in each of the classes ω_i.

[3]Usually given, but if unknown, it can be estimated as N_i/N where N_i is the number of training samples which belong to class ω_i, and N is the total number of training samples

[4]This is also estimated from training data which will be explained later more in detail.

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 $\mu = (0, 0)$.

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

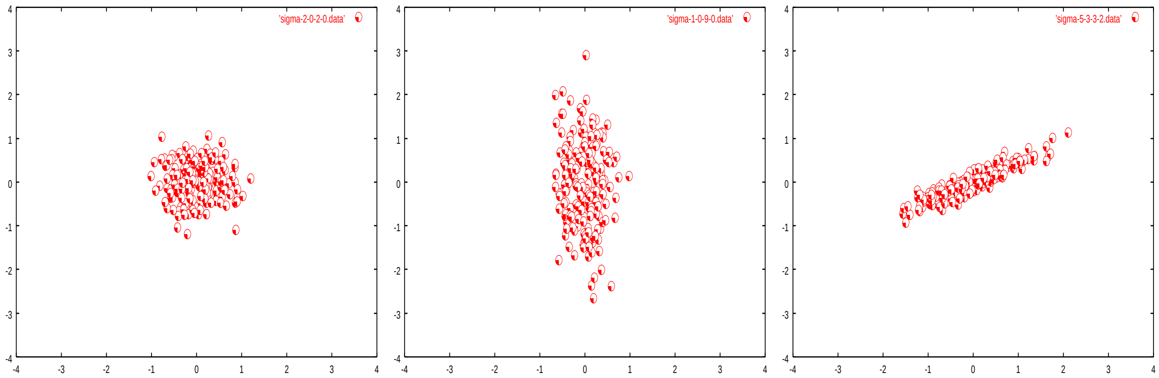


Figure 22: A cloud of 200 points each with different three different $N[\mu, \Sigma]$.

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

We now try to calculate the invers of the following 3-D matrix A which appeared in the subsection 3.3.

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

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, devide the {*1st-raw*} by 2.4, devide the {*2nd-raw*} by 2.4, and devide 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 invers 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

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