Solution Manual For Pattern Recognition and Machine Learning

Edited By

ZHENGQI GAO

Information Science and Technology School
Fudan University

Nov.2017

0.1 Introduction

Problem 1.1 Solution

We let the derivative of *error function* E with respect to vector $\mathbf{w}$ equals to $\mathbf{0}$, (i.e. $\frac{\partial E}{\partial \mathbf{w}} = 0$), and this will be the solution of $\mathbf{w} = \{w_i\}$ which minimizes *error function* E. To solve this problem, we will calculate the derivative of E with respect to every w_i , and let them equal to 0 instead. Based on (1.1) and (1.2) we can obtain :

=>
$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\} x_n^i = 0$$

=>
$$\sum_{n=1}^{N} y(x_n, \mathbf{w}) x_n^i = \sum_{n=1}^{N} x_n^i t_n$$

=>
$$\sum_{n=1}^{N} (\sum_{j=0}^{M} w_j x_n^j) x_n^i = \sum_{n=1}^{N} x_n^i t_n$$

=>
$$\sum_{n=1}^{N} \sum_{j=0}^{M} w_j x_n^{(j+i)} = \sum_{n=1}^{N} x_n^i t_n$$

=>
$$\sum_{j=0}^{M} \sum_{n=1}^{N} x_n^{(j+i)} w_j = \sum_{n=1}^{N} x_n^i t_n$$

If we denote $A_{ij} = \sum_{n=1}^{N} x_n^{i+j}$ and $T_i = \sum_{n=1}^{N} {x_n}^i t_n$, the equation above can be written exactly as (1.222), Therefore the problem is solved.

Problem 1.2 Solution

This problem is similar to Prob.1.1, and the only difference is the last term on the right side of (1.4), the penalty term. So we will do the same thing as in Prob.1.1 :

=>
$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\} x_n^i + \lambda w_i = 0$$

=>
$$\sum_{j=0}^{M} \sum_{n=1}^{N} x_n^{(j+i)} w_j + \lambda w_i = \sum_{n=1}^{N} x_n^i t_n$$

=>
$$\sum_{j=0}^{M} \{\sum_{n=1}^{N} x_n^{(j+i)} + \delta_{ji}\lambda\} w_j = \sum_{n=1}^{N} x_n^i t_n$$

where

$$\delta_{ji} \begin{cases} 0 & \text{j} \neq \text{i} \\ 1 & \text{j} = \text{i} \end{cases}$$

Problem 1.3 Solution

This problem can be solved by *Bayes' theorem*. The probability of selecting an apple $P(a)$:

$$P(a) = P(a|r)P(r)+P(a|b)P(b)+P(a|g)P(g) = \frac{3}{10}\times 0.2+\frac{1}{2}\times 0.2+\frac{3}{10}\times 0.6 = 0.34$$

Based on *Bayes' theorem*, the probability of an selected orange coming from the green box $P(g|o)$:

$$P(g|o) = \frac{P(o|g)P(g)}{P(o)}$$

We calculate the probability of selecting an orange $P(o)$ first :

$$P(o) = P(o|r)P(r)+P(o|b)P(b)+P(o|g)P(g) = \frac{4}{10}\times 0.2+\frac{1}{2}\times 0.2+\frac{3}{10}\times 0.6 = 0.36$$

Therefore we can get :

$$P(g|o) = \frac{P(o|g)P(g)}{P(o)} = \frac{\frac{3}{10}\times 0.6}{0.36} = 0.5$$

Problem 1.4 Solution

This problem needs knowledge about *calculus*, especially about *Chain rule*. We calculate the derivative of $P_y(y)$ with respect to y, according to (1.27) :

$$\frac{dp_y(y)}{dy} = \frac{d(p_x(g(y))|g^{`}(y)|)}{dy} = \frac{dp_x(g(y))}{dy}|g^{`}(y)| + p_x(g(y))\frac{d|g^{`}(y)|}{dy} \qquad (*)$$

The first term in the above equation can be further simplified:

$$\frac{dp_x(g(y))}{dy}|g^{`}(y)| = \frac{dp_x(g(y))}{dg(y)}\frac{dg(y)}{dy}|g^{`}(y)| \qquad (**)$$

If $\hat{x}$ is the maximum of density over x, we can obtain :

$$\frac{dp_x(x)}{dx}\Big|_{\hat{x}} = 0$$

Therefore, when $y = \hat{y}, s.t. \hat{x} = g(\hat{y})$, the first term on the right side of $(**)$ will be 0, leading the first term in $(*)$ equals to 0, however because of the existence of the second term in $(*)$, the derivative may not equal to 0. But

when linear transformation is applied, the second term in $(*)$ will vanish, (e.g. $x = ay + b$). A simple example can be shown by :

$$p_x(x) = 2x, \quad x \in [0,1] \quad => \quad \hat{x} = 1$$

And given that:

$$x = sin(y)$$

Therefore, $p_y(y) = 2\sin(y)|\cos(y)|$, $y \in [0, \frac{\pi}{2}]$, which can be simplified :

$$p_y(y) = \sin(2y), \quad y \in [0, \frac{\pi}{2}] \quad => \quad \hat{y} = \frac{\pi}{4}$$

However, it is quite obvious :

$$\hat{x} \neq \sin(\hat{y})$$

Problem 1.5 Solution

This problem takes advantage of the property of expectation:

$$\begin{aligned} var[f] &= \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \\ &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2 \\ => var[f] &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \end{aligned}$$

Problem 1.6 Solution

Based on (1.41), we only need to prove when x and y is independent, $\mathbb{E}_{x,y}[xy] = \mathbb{E}[x]\mathbb{E}[y]$. Because x and y is independent, we have :

$$p(x,y) = p_x(x)\,p_y(y)$$

Therefore:

$$\begin{aligned} \int\int xyp(x,y)dxdy &= \int\int xyp_x(x)p_y(y)dxdy \\ &= (\int xp_x(x)dx)(\int yp_y(y)dy) \\ => \mathbb{E}_{x,y}[xy] &= \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

Problem 1.7 Solution

This problem should take advantage of *Integration by substitution*.

$$\begin{aligned} I^2 &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} exp(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2)dxdy \\ &= \int_0^{2\pi}\int_0^{+\infty} exp(-\frac{1}{2\sigma^2}r^2)r\,dr\,d\theta \end{aligned}$$

Here we utilize :

$$x = r\cos\theta, \quad y = r\sin\theta$$

Based on the fact :

$$\int_0^{+\infty} exp(-\frac{1}{2\sigma^2}) r\,dr = -\sigma^2 exp(-\frac{r^2}{2\sigma^2})\Big|_0^{+\infty} = -\sigma^2(0-(-1)) = \sigma^2$$

Therefore, I can be solved :

$$I^2 = \int_0^{2\pi} \sigma^2\,d\theta = 2\pi\sigma^2, \quad => I = \sqrt{2\pi}\sigma$$

And next,we will show that Gaussian distribution $\mathcal{N}(x|\mu,\sigma^2)$ is normalized, (i.e. $\int_{-\infty}^{+\infty}\mathcal{N}(x|\mu,\sigma^2)dx = 1$) :

$$\begin{aligned}
\int_{-\infty}^{+\infty}\mathcal{N}(x|\mu,\sigma^2)dx &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}dx \\
&= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}y^2\}dy \quad (y = x-\mu) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{+\infty}exp\{-\frac{1}{2\sigma^2}y^2\}dy \\
&= 1
\end{aligned}$$

Problem 1.8 Solution

The first question will need the result of Prob.1.7 :

$$\begin{aligned}
\int_{-\infty}^{+\infty}\mathcal{N}(x|\mu,\sigma^2)x\,dx &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}x\,dx \\
&= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}y^2\}(y+\mu)dy \quad (y = x-\mu) \\
&= \mu\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}y^2\}dy + \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}y^2\}y\,dy \\
&= \mu + 0 = \mu
\end{aligned}$$

The second problem has already be given hint in the description. Given that :

$$\frac{d(fg)}{dx} = f\frac{dg}{dx} + g\frac{df}{dx}$$

We differentiate both side of (1.127) with respect to σ^2, we will obtain :

$$\int_{-\infty}^{+\infty}(-\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4})\mathcal{N}(x|\mu,\sigma^2)dx = 0$$

Provided the fact that $\sigma \neq 0$, we can get:

$$\int_{-\infty}^{+\infty} (x-\mu)^2 \mathcal{N}(x|\mu,\sigma^2)\,dx = \int_{-\infty}^{+\infty} \sigma^2 \mathcal{N}(x|\mu,\sigma^2)\,dx = \sigma^2$$

So the equation above has actually proven (1.51), according to the definition:

$$var[x] = \int_{-\infty}^{+\infty} (x-\mathbb{E}[x])^2 \mathcal{N}(x|\mu,\sigma^2)\,dx$$

Where $\mathbb{E}[x] = \mu$ has already been proved. Therefore :

$$var[x] = \sigma^2$$

Finally,

$$\mathbb{E}[x^2] = var[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$$

Problem 1.9 Solution

Here we only focus on (1.52), because (1.52) is the general form of (1.42). Based on the definition : The maximum of distribution is known as its mode and (1.52), we can obtain :

$$\begin{aligned} \frac{\partial \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \mathbf{x}} &= -\frac{1}{2}[\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^T](\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ &= -\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \end{aligned}$$

Where we take advantage of :

$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A}+\mathbf{A}^T)\mathbf{x} \quad \text{and} \quad (\boldsymbol{\Sigma}^{-1})^T = \boldsymbol{\Sigma}^{-1}$$

Therefore,

$$\text{only when} \quad \mathbf{x} = \boldsymbol{\mu}, \ \frac{\partial \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \mathbf{x}} = 0$$

Note: You may also need to calculate *Hessian Matrix* to prove that it is maximum. However, here we find that the first derivative only has one root. Based on the description in the problem, this point should be maximum point.

Problem 1.10 Solution

We will solve this problem based on the definition of *expectation*, *variation*

and *independence*.

$$\begin{aligned}
\mathbb{E}[x+z] &= \int\int (x+z)p(x,z)\,dx\,dz \\
&= \int\int (x+z)p(x)p(z)\,dx\,dz \\
&= \int\int xp(x)p(z)\,dx\,dz + \int\int zp(x)p(z)\,dx\,dz \\
&= \int (\int p(z)\,dz)xp(x)\,dx + \int (\int p(x)\,dx)zp(z)\,dz \\
&= \int xp(x)\,dx + \int zp(z)\,dz \\
&= \mathbb{E}[x] + \mathbb{E}[z]
\end{aligned}$$

$$\begin{aligned}
var[x+z] &= \int\int (x+z-\mathbb{E}[x+z])^2 p(x,z)\,dx\,dz \\
&= \int\int \{(x+z)^2 - 2(x+z)\mathbb{E}[x+z]) + \mathbb{E}^2[x+z]\}p(x,z)\,dx\,dz \\
&= \int\int (x+z)^2 p(x,z)\,dx\,dz - 2\mathbb{E}[x+z]\int\int (x+z)p(x,z)\,dx\,dz + \mathbb{E}^2[x+z] \\
&= \int\int (x+z)^2 p(x,z)\,dx\,dz - \mathbb{E}^2[x+z] \\
&= \int\int (x^2+2xz+z^2)p(x)p(z)\,dx\,dz - \mathbb{E}^2[x+z] \\
&= \int (\int p(z)dz)x^2 p(x)\,dx + \int\int 2xz\,p(x)p(z)\,dx\,dz + \int (\int p(x)dx)z^2 p(z)\,dz - \mathbb{E}^2[x+z] \\
&= \mathbb{E}[x^2]+\mathbb{E}[z^2]-\mathbb{E}^2[x+z]+\int\int 2xz\,p(x)p(z)\,dx\,dz \\
&= \mathbb{E}[x^2]+\mathbb{E}[z^2]-(\mathbb{E}[x]+\mathbb{E}[z])^2+\int\int 2xz\,p(x)p(z)\,dx\,dz \\
&= \mathbb{E}[x^2]-\mathbb{E}^2[x]+\mathbb{E}[z^2]-\mathbb{E}^2[z]-2\mathbb{E}[x]\mathbb{E}[z]+2\int\int xz\,p(x)p(z)\,dx\,dz \\
&= var[x]+var[z]-2\mathbb{E}[x]\mathbb{E}[z]+2(\int xp(x)\,dx)(\int zp(z)\,dz) \\
&= var[x]+var[z]
\end{aligned}$$

Problem 1.11 Solution

Based on prior knowledge that μ_{ML} and σ^2_{ML} will decouple. We will first calculate μ_{ML} :

$$\frac{d(ln\,p(\mathbf{x}|\mu,\sigma^2))}{d\mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)$$

We let :

$$\frac{d(ln\,p(\mathbf{x}|\mu,\sigma^2))}{d\mu} = 0$$

Therefore :

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

And because:

$$\frac{d(ln\, p(\mathbf{x}|\mu,\sigma^2))}{d\sigma^2} = \frac{1}{2\sigma^4}(\sum_{n=1}^{N}(x_n-\mu)^2 - N\sigma^2)$$

We let :

$$\frac{d(ln\, p(\mathbf{x}|\mu,\sigma^2))}{d\sigma^2} = 0$$

Therefore :

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$$

Problem 1.12 Solution

It is quite straightforward for $\mathbb{E}[\mu_{ML}]$, with the prior knowledge that x_n is i.i.d. and it also obeys Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$.

$$\mathbb{E}[\mu_{ML}] = \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N} x_n] = \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N} x_n] = \mathbb{E}[x_n] = \mu$$

For $\mathbb{E}[\sigma^2_{ML}]$, we need to take advantage of (1.56) and what has been given in the problem :

$$\begin{aligned}
\mathbb{E}[\sigma^2_{ML}] &= \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2] \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}(x_n-\mu_{ML})^2] \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}(x_n^2 - 2x_n\mu_{ML}+\mu^2_{ML})] \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}x_n^2] - \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}2x_n\mu_{ML}] + \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}\mu^2_{ML}] \\
&= \mu^2+\sigma^2 - \frac{2}{N}\mathbb{E}[\sum_{n=1}^{N}x_n(\frac{1}{N}\sum_{n=1}^{N}x_n)] + \mathbb{E}[\mu^2_{ML}] \\
&= \mu^2+\sigma^2 - \frac{2}{N^2}\mathbb{E}[\sum_{n=1}^{N}x_n(\sum_{n=1}^{N}x_n)] + \mathbb{E}[(\frac{1}{N}\sum_{n=1}^{N}x_n)^2] \\
&= \mu^2+\sigma^2 - \frac{2}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] + \frac{1}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] \\
&= \mu^2+\sigma^2 - \frac{1}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] \\
&= \mu^2+\sigma^2 - \frac{1}{N^2}[N(N\mu^2+\sigma^2)]
\end{aligned}$$

Therefore we have:

$$\mathbb{E}[\sigma^2_{ML}] = (\frac{N-1}{N})\sigma^2$$

Problem 1.13 Solution

This problem can be solved in the same method used in Prob.1.12 :

$$\begin{aligned}
\mathbb{E}[\sigma^2_{ML}] &= \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2] \quad \text{(Because here we use } \mu \text{ to replace } \mu_{ML}) \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}(x_n-\mu)^2] \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}(x_n^2-2x_n\mu+\mu^2)] \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}x_n^2] - \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}2x_n\mu] + \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}\mu^2] \\
&= \mu^2+\sigma^2-\frac{2\mu}{N}\mathbb{E}[\sum_{n=1}^{N}x_n]+\mu^2 \\
&= \mu^2+\sigma^2-2\mu^2+\mu^2 \\
&= \sigma^2
\end{aligned}$$

Note: The biggest difference between Prob.1.12 and Prob.1.13 is that the mean of Gaussian Distribution is known previously (in Prob.1.13) or not (in Prob.1.12). In other words, the difference can be shown by the following equations:

$$\mathbb{E}[\mu^2] = \mu^2 \quad (\mu \text{ is determined, i.e. its } \textit{expectation} \text{ is itself, also true for } \mu^2)$$

$$\mathbb{E}[\mu^2_{ML}] = \mathbb{E}[(\frac{1}{N}\sum_{n=1}^{N}x_n)^2] = \frac{1}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] = \frac{1}{N^2}N(N\mu^2+\sigma^2) = \mu^2+\frac{\sigma^2}{N}$$

Problem 1.14 Solution

This problem is quite similar to the fact that *any function $f(x)$ can be written into the sum of an odd function and an even function*. If we let:

$$w^S_{ij} = \frac{w_{ij}+w_{ji}}{2} \quad \text{and} \quad w^A_{ij} = \frac{w_{ij}-w_{ji}}{2}$$

It is obvious that they satisfy the constraints described in the problem, which are :

$$w_{ij} = w^S_{ij} + w^A_{ij}, \quad w^S_{ij} = w^S_{ji}, \quad w^A_{ij} = -w^A_{ji}$$

To prove (1.132), we only need to simplify it :

$$\begin{aligned}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j &= \sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}^S + w_{ij}^A)x_i x_j \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^A x_i x_j\end{aligned}$$

Therefore, we only need to prove that the second term equals to 0, and here we use a simple trick: we will prove twice of the second term equals to 0 instead.

$$\begin{aligned}2\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^A x_i x_j &= \sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}^A + w_{ij}^A)x_i x_j \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}^A - w_{ji}^A)x_i x_j \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^A x_i x_j - \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ji}^A x_i x_j \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^A x_i x_j - \sum_{j=1}^{D}\sum_{i=1}^{D} w_{ji}^A x_j x_i \\ &= 0\end{aligned}$$

Therefore, we choose the coefficient matrix to be symmetric as described in the problem. Considering about the symmetry, we can see that if and only if for $i = 1, 2, ..., D$ and $i \leq j$, w_{ij} is given, the whole matrix will be determined. Hence, the number of independent parameters are given by :

$$D + D - 1 + ... + 1 = \frac{D(D+1)}{2}$$

Note: You can view this intuitively by considering if the upper triangular part of a symmetric matrix is given, the whole matrix will be determined.

Problem 1.15 Solution

This problem is a more general form of Prob.1.14, so the method can also be used here: we will find a way to use $w_{i_1 i_2 ... i_M}$ to represent $\widetilde{w}_{i_1 i_2 ... i_M}$.

We begin by introducing a mapping function:

$$F(x_{i1}x_{i2}...x_{iM}) = x_{j1}x_{j2}..., x_{jM}$$

$$s.t. \bigcup_{k=1}^{M} x_{ik} = \bigcup_{k=1}^{M} x_{jk}, \quad \text{and} \quad x_{j1} \geq x_{j2} \geq x_{j3}... \geq x_{jM}$$

It is complexed to write F in mathematical form. Actually this function does a simple work: it rearranges the element in a decreasing order based on its subindex. Several examples are given below, when $D = 5$, $M = 4$:

$$F(x_5x_2x_3x_2) = x_5x_3x_2x_2$$

$$F(x_1x_3x_3x_2) = x_3x_3x_2x_1$$

$$F(x_1x_4x_2x_3) = x_4x_3x_2x_1$$

$$F(x_1x_1x_5x_2) = x_5x_2x_1x_1$$

After introducing F, the solution will be very simple, based on the fact that F will not change the value of the term, but only rearrange it.

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{D}\cdots\sum_{i_M=1}^{D} w_{i_1i_2\ldots i_M}x_{i1}x_{i2}\ldots x_{iM} = \sum_{j_1=1}^{D}\sum_{j_2=1}^{j_1}\cdots\sum_{j_M=1}^{j_{M-1}} \widetilde{w}_{j_1j_2\ldots j_M}x_{j1}x_{j2}\ldots x_{jM}$$

where $$\widetilde{w}_{j_1j_2\ldots j_M} = \sum_{w\in\Omega} w$$

$$\Omega = \{w_{i_1i_2\ldots i_M} \,\big|\, F(x_{i1}x_{i2}\ldots x_{iM}) = x_{j1}x_{j2}\ldots x_{jM},\ \forall x_{i1}x_{i2}\ldots x_{iM}\}$$

By far, we have already proven (1.134). *Mathematical induction* will be used to prove (1.135) and we will begin by proving $D = 1$, i.e. $n(1,M) = n(1,M-1)$. When $D = 1$, (1.134) will degenerate into $\widetilde{w}x_1^M$, i.e., it only has one term, whose coefficient is govern by $\widetilde{w}$ regardless the value of M.

Therefore, we have proven when $D = 1$, $n(D,M) = 1$. Suppose (1.135) holds for D, let's prove it will also hold for $D+1$, and then (1.135) will be proved based on *Mathematical induction*.

Let's begin based on (1.134):

$$\sum_{i_1=1}^{D+1}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} \widetilde{w}_{i_1i_2\ldots i_M}x_{i1}x_{i2}\ldots x_{iM} \qquad (*)$$

We divide $(*)$ into two parts based on the first summation: the first part is made up of $i_i = 1,2,\ldots,D$ and the second part $i_1 = D+1$. After division, the first part corresponds to $n(D,M)$, and the second part corresponds to $n(D+1,M-1)$. Therefore we obtain:

$$n(D+1,M) = n(D,M) + n(D+1,M-1) \qquad (**)$$

And given the fact that (1.135) holds for D :

$$n(D,M) = \sum_{i=1}^{D} n(i,M-1)$$

Therefore,we substitute it into $(**)$

$$n(D+1,M) = \sum_{i=1}^{D} n(i,M-1) + n(D+1,M-1) = \sum_{i=1}^{D+1} n(i,M-1)$$

We will prove (1.136) in a different but simple way. We rewrite (1.136) in *Permutation and Combination* view:

$$\sum_{i=1}^{D} C_{i+M-2}^{M-1} = C_{D+M-1}^{M}$$

Firstly, We expand the summation.

$$C_{M-1}^{M-1} + C_{M}^{M-1} + ... C_{D+M-2}^{M-1} = C_{D+M-1}^{M}$$

Secondly, we rewrite the first term on the left side to C_M^M, because $C_{M-1}^{M-1} = C_M^M = 1$. In other words, we only need to prove:

$$C_{M}^{M} + C_{M}^{M-1} + ... C_{D+M-2}^{M-1} = C_{D+M-1}^{M}$$

Thirdly, we take advantage of the property : $C_N^r = C_{N-1}^r + C_{N-1}^{r-1}$. So we can recursively combine the first term and the second term on the left side, and it will ultimately equal to the right side.

(1.137) gives the mathematical form of $n(D,M)$, and we need all the conclusions above to prove it.

Let's give some intuitive concepts by illustrating $M = 0,1,2$. When $M = 0$, (1.134) will consist of only a constant term, which means $n(D,0) = 1$. When $M = 1$,it is obvious $n(D,1) = D$, because in this case (1.134) will only have D terms if we expand it. When $M = 2$, it degenerates to Prob.1.14, so $n(D,2) = \frac{D(D+1)}{2}$ is also obvious. Suppose (1.137) holds for $M-1$, let's prove it will also hold for M.

$$\begin{aligned}
n(D,M) &= \sum_{i=1}^{D} n(i,M-1) \quad \text{(based on (1.135))} \\
&= \sum_{i=1}^{D} C_{i+M-2}^{M-1} \quad \text{(based on (1.137) holds for } M-1 \text{)} \\
&= C_{M-1}^{M-1} + C_{M}^{M-1} + C_{M+1}^{M-1} ... + C_{D+M-2}^{M-1} \\
&= (C_{M}^{M} + C_{M}^{M-1}) + C_{M+1}^{M-1} ... + C_{D+M-2}^{M-1} \\
&= (C_{M+1}^{M} + C_{M+1}^{M-1}) ... + C_{D+M-2}^{M-1} \\
&= C_{M+2}^{M} ... + C_{D+M-2}^{M-1} \\
&\quad ... \\
&= C_{D+M-1}^{M}
\end{aligned}$$

By far, all have been proven.

Problem 1.16 Solution

This problem can be solved in the same way as the one in Prob.1.15. Firstly, we should write the expression consisted of all the independent terms up to Mth order corresponding to $N(D,M)$. By adding a summation regarding to M on the left side of (1.134), we obtain:

$$\sum_{m=0}^{M}\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_m=1}^{i_{m-1}}\widetilde{w}_{i_1 i_2 \ldots i_m}x_{i1}x_{i2}\ldots x_{im} \qquad (*)$$

(1.138) is quite obvious if we view m as an looping variable, iterating through all the possible orders less equal than M, and for every possible oder m, the independent parameters are given by $n(D,m)$.

Let's prove (1.138) in a formal way by using *Mathematical Induction*. When $M = 1$,$(*)$ will degenerate to two terms: $m = 0$, corresponding to $n(D,0)$ and $m = 1$, corresponding to $n(D,1)$. Therefore $N(D,1) = n(D,0) + n(D,1)$. Suppose (1.138) holds for M, we will see that it will also hold for $M+1$. Let's begin by writing all the independent terms based on $(*)$:

$$\sum_{m=0}^{M+1}\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_m=1}^{i_{m-1}}\widetilde{w}_{i_1 i_2 \ldots i_m}x_{i1}x_{i2}\ldots x_{im} \qquad (**)$$

Using the same technique as in Prob.1.15, we divide $(**)$ to two parts based on the summation regarding to m: the first part consisted of $m = 0,1,\ldots,M$ and the second part $m = M+1$. Hence, the first part will correspond to $N(D,M)$ and the second part will correspond to $n(D,M+1)$. So we obtain:

$$N(D,M+1) = N(D,M) + n(D,M+1)$$

Then we substitute (1.138) into the equation above :

$$\begin{aligned} N(D,M+1) &= \sum_{m=0}^{M} n(D,m) + n(D,M+1) \\ &= \sum_{m=0}^{M+1} n(D,m) \end{aligned}$$

To prove (1.139), we will also use the same technique in Prob.1.15 instead of *Mathematical Induction*. We begin based on already proved (1.138):

$$N(D,M) = \sum_{m=0}^{M} n(D,M)$$

We then take advantage of (1.137):

$$\begin{aligned} N(D,M) &= \sum_{m=0}^{M} C_{D+m-1}^{m} \\ &= C_{D-1}^{0} + C_{D}^{1} + C_{D+1}^{2} + ... + C_{D+M-1}^{M} \\ &= (C_{D}^{0} + C_{D}^{1}) + C_{D+1}^{2} + ... + C_{D+M-1}^{M} \\ &= (C_{D+1}^{1} + C_{D+1}^{2}) + ... + C_{D+M-1}^{M} \\ &= ... \\ &= C_{D+M}^{M} \end{aligned}$$

Here as asked by the problem, we will view the growing speed of $N(D,M)$. We should see that in $n(D,M)$, D and M are symmetric, meaning that we only need to prove when $D \gg M$, it will grow like D^M, and then the situation of $M \gg D$ will be solved by symmetry.

$$\begin{aligned} N(D,M) &= \frac{(D+M)!}{D!M!} \approx \frac{(D+M)^{D+M}}{D^D M^M} \\ &= \frac{1}{M^M}(\frac{D+M}{D})^D (D+M)^M \\ &= \frac{1}{M^M}[(1+\frac{M}{D})^{\frac{D}{M}}]^M (D+M)^M \\ &\approx (\frac{e}{M})^M (D+M)^M \\ &= \frac{e^M}{M^M}(1+\frac{M}{D})^M D^M \\ &= \frac{e^M}{M^M}[(1+\frac{M}{D})^{\frac{D}{M}}]^{\frac{M^2}{D}} D^M \\ &\approx \frac{e^{M+\frac{M^2}{D}}}{M^M} D^M \approx \frac{e^M}{M^M} D^M \end{aligned}$$

Where we use *Stirling's approximation*, $\lim_{n\to+\infty}(1+\frac{1}{n})^n = e$ and $e^{\frac{M^2}{D}} \approx e^0 = 1$. According to the description in the problem, When $D \gg M$, we can actually view $\frac{e^M}{M^M}$ as a constant, so $N(D,M)$ will grow like D^M in this case. And by symmetry, $N(D,M)$ will grow like M^D, when $M \gg D$.

Finally, we are asked to calculate $N(10,3)$ and $N(100,3)$:

$$N(10,3) = C_{13}^{3} = 286$$

$$N(100,3) = C_{103}^{3} = 176851$$

Problem 1.17 Solution

$$\begin{aligned}\Gamma(x+1) &= \int_0^{+\infty} u^x e^{-u} du \\ &= \int_0^{+\infty} -u^x de^{-u} \\ &= -u^x e^{-u}\Big|_0^{+\infty} - \int_0^{+\infty} e^{-u} d(-u^x) \\ &= -u^x e^{-u}\Big|_0^{+\infty} + x\int_0^{+\infty} e^{-u} u^{x-1} du \\ &= -u^x e^{-u}\Big|_0^{+\infty} + x\Gamma(x)\end{aligned}$$

Where we have taken advantage of *Integration by parts* and according to the equation above, we only need to prove the first term equals to 0. Given *L'Hospital's Rule*:

$$\lim_{u\to+\infty} -\frac{u^x}{e^u} = \lim_{u\to+\infty} -\frac{x!}{e^u} = 0$$

And also when $u = 0, -u^x e^u = 0$, so we have proved $\Gamma(x+1) = x\Gamma(x)$. Based on the definition of $\Gamma(x)$, we can write:

$$\Gamma(1) = \int_0^{+\infty} e^{-u} du = -e^{-u}\Big|_0^{+\infty} = -(0-1) = 1$$

Therefore when x is an integer:

$$\Gamma(x) = (x-1)\Gamma(x-1) = (x-1)(x-2)\Gamma(x-2) = \ldots = x!\,\Gamma(1) = x!$$

Problem 1.18 Solution

Based on (1.124) and (1.126) and by substituting x to $\sqrt{2}\sigma y$, it is quite obvious to obtain :

$$\int_{-\infty}^{+\infty} e^{-x_i^2} dx_i = \sqrt{\pi}$$

Therefore, the left side of (1.42) will equal to $\pi^{\frac{D}{2}}$. For the right side of (1.42):

$$\begin{aligned}S_D \int_0^{+\infty} e^{-r^2} r^{D-1} dr &= S_D \int_0^{+\infty} e^{-u} u^{\frac{D-1}{2}} d\sqrt{u} \quad (u = r^2) \\ &= \frac{S_D}{2}\int_0^{+\infty} e^{-u} u^{\frac{D}{2}-1} du \\ &= \frac{S_D}{2}\Gamma(\frac{D}{2})\end{aligned}$$

Hence, we obtain:

$$\pi^{\frac{D}{2}} = \frac{S_D}{2}\Gamma(\frac{D}{2}) \quad => \quad S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2})}$$

S_D has given the expression of the surface area with radius 1 in dimension D, we can further expand the conclusion: the surface area with radius r in dimension D will equal to $S_D \cdot r^{D-1}$, and when $r = 1$, it will reduce to S_D. This conclusion is naive, if you find that the surface area of different sphere in dimension D is proportion to the $D-1$th power of radius, i.e. r^{D-1}. Considering the relationship between V and S of a sphere with arbitrary radius in dimension D: $\frac{dV}{dr} = S$, we can obtain :

$$V = \int S\,dr = \int S_D r^{D-1}\,dr = \frac{S_D}{D} r^D$$

The equation above gives the expression of the volume of a sphere with radius r in dimension D, so we let $r = 1$:

$$V_D = \frac{S_D}{D}$$

For $D = 2$ and $D = 3$:

$$V_2 = \frac{S_2}{2} = \frac{1}{2} \cdot \frac{2\pi}{\Gamma(1)} = \pi$$

$$V_3 = \frac{S_3}{3} = \frac{1}{3} \cdot \frac{2\pi^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} = \frac{1}{3} \cdot \frac{2\pi^{\frac{3}{2}}}{\frac{\sqrt{\pi}}{2}} = \frac{4}{3}\pi$$

Problem 1.19 Solution

We have already given a hint in the solution of Prob.1.18, and here we will make it more clearly: the volume of a sphere with radius r is $V_D \cdot r^D$. This is quite similar with the conclusion we obtained in Prob.1.18 about the surface area except that it is proportion to Dth power of its radius, i.e. r^D not r^{D-1}.

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{V_D a^D}{(2a)^D} = \frac{S_D}{2^D D} = \frac{\pi^{\frac{D}{2}}}{2^{D-1} D\,\Gamma(\frac{D}{2})} \qquad (*)$$

Where we have used the result of (1.143). And when $D \to +\infty$, we will use a simple method to show that $(*)$ will converge to 0. We rewrite it :

$$(*) = \frac{2}{D} \cdot (\frac{\pi}{4})^{\frac{D}{2}} \cdot \frac{1}{\Gamma(\frac{D}{2})}$$

Hence, it is now quite obvious, all the three terms will converge to 0 when $D \to +\infty$. Therefore their product will also converge to 0. The last problem is quite simple :

$$\frac{\text{center to one corner}}{\text{center to one side}} = \frac{\sqrt{a^2 \cdot D}}{a} = \sqrt{D} \quad \text{and} \quad \lim_{D \to +\infty} \sqrt{D} = +\infty$$

Problem 1.20 Solution

The density of probability in a thin shell with radius r and thickness ϵ can be viewed as a constant. And considering that a sphere in dimension D with radius r has surface area $S_D r^{D-1}$, which has already been proved in Prob.1.19 :

$$\int_{shell} p(\mathbf{x})d\mathbf{x} = p(\mathbf{x})\int_{shell} d\mathbf{x} = \frac{exp(-\frac{r^2}{2\sigma^2})}{(2\pi\sigma^2)^{\frac{D}{2}}} \cdot V(\text{shell}) = \frac{exp(-\frac{r^2}{2\sigma^2})}{(2\pi\sigma^2)^{\frac{D}{2}}} S_D r^{D-1}\epsilon$$

Thus we denote :

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{\frac{D}{2}}} exp(-\frac{r^2}{2\sigma^2})$$

We calculate the derivative of (1.148) with respect to r :

$$\frac{dp(r)}{dr} = \frac{S_D}{(2\pi\sigma^2)^{\frac{D}{2}}} r^{D-2} exp(-\frac{r^2}{2\sigma^2})(D-1-\frac{r^2}{\sigma^2}) \qquad (*)$$

We let the derivative equal to 0, we will obtain its unique root(stationary point) $\hat{r} = \sqrt{D-1}\sigma$, because $r \in [0, +\infty]$. When $r < \hat{r}$, the derivative is large than 0, $p(r)$ will increase as $r \uparrow$, and when $r > \hat{r}$, the derivative is less than 0, $p(r)$ will decrease as $r \uparrow$. Therefore $\hat{r}$ will be the only maximum point. And it is obvious when $D \gg 1$, $\hat{r} \approx \sqrt{D}\sigma$.

$$\begin{aligned}
\frac{p(\hat{r}+\epsilon)}{p(\hat{r})} &= \frac{(\hat{r}+\epsilon)^{D-1} exp(-\frac{(\hat{r}+\epsilon)^2}{2\sigma^2})}{\hat{r}^{D-1} exp(-\frac{\hat{r}^2}{2\sigma^2})} \\
&= (1+\frac{\epsilon}{\hat{r}})^{D-1} exp(-\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2}) \\
&= exp(-\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + (D-1)ln(1+\frac{\epsilon}{\hat{r}}))
\end{aligned}$$

We process for the exponential term by using *Taylor Expansion*.

$$\begin{aligned}
-\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + (D-1)ln(1+\frac{\epsilon}{\hat{r}}) &\approx -\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + (D-1)(\frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}) \\
&= -\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + \frac{2\hat{r}\epsilon-\epsilon^2}{2\sigma^2} \\
&= -\frac{\epsilon^2}{\sigma^2}
\end{aligned}$$

Therefore, $p(\hat{r}+\epsilon) = p(\hat{r})exp(-\frac{\epsilon^2}{\sigma^2})$. **Note: Here I draw a different conclusion compared with (1.149),** but I do not think there is any mistake in my deduction.

Finally, we see from (1.147) :

$$p(\mathbf{x})\Big|_{\mathbf{x}=0} = \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}}$$

$$p(\mathbf{x})\Big|_{||\mathbf{x}||^2=\hat{r}^2} = \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} exp(-\frac{\hat{r}^2}{2\sigma^2}) \approx \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} exp(-\frac{D}{2})$$

Problem 1.21 Solution

The first question is rather simple :

$$(ab)^{\frac{1}{2}} - a = a^{\frac{1}{2}}(b^{\frac{1}{2}} - a^{\frac{1}{2}}) \geq 0$$

Where we have taken advantage of $b \geq a \geq 0$. And based on (1.78):

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in R_1, C_2) + p(\mathbf{x} \in R_2, C_1) \\ &= \int_{R_1} p(\mathbf{x}, C_2) dx + \int_{R_2} p(\mathbf{x}, C_1) dx \end{aligned}$$

Recall that the decision rule which can minimize misclassification is that if $p(\mathbf{x}, C_1) > p(\mathbf{x}, C_2)$, for a given value of $\mathbf{x}$, we will assign that $\mathbf{x}$ to class C_1. We can see that in decision area R_1, it should satisfy $p(\mathbf{x}, C_1) > p(\mathbf{x}, C_2)$. Therefore, using what we have proved, we can obtain :

$$\int_{R_1} p(\mathbf{x}, C_2) dx \leq \int_{R_1} \{p(\mathbf{x}, C_1) p(\mathbf{x}, C_2)\}^{\frac{1}{2}} dx$$

It is the same for decision area R_2. Therefore we can obtain:

$$p(\text{mistake}) \leq \int \{p(\mathbf{x}, C_1) p(\mathbf{x}, C_2)\}^{\frac{1}{2}} dx$$