Solutions Manual to Pattern Recognition and Machine Learning

Hiromichi Inawashiro

June 14, 2024

1 Introduction

1.1

To minimise

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n,\mathbf{w}) - t_n\right)^2, \tag{1.1}$$

setting its derivative to zero gives

$$\mathbf{0} = \sum_{n=1}^{N}\frac{\partial y(x_n,\mathbf{w})}{\partial \mathbf{w}}\left(y(x_n,\mathbf{w}) - t_n\right). \tag{1.2}$$

Substituting

$$y(x_n,\mathbf{w}) = \sum_{j=0}^{M} w_j x_n^j \tag{1.3}$$

gives

$$0 = \sum_{n=1}^{N} x_n^i \left(\sum_{j=0}^{M} w_j x_n^j - t_n\right). \tag{1.4}$$

Therefore,

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{1.5}$$

where

$$\begin{aligned} A_{ij} &= \sum_{n=1}^{N} x_n^{i+j}, \\ T_i &= \sum_{n=1}^{N} x_n^i t_n. \end{aligned} \tag{1.6}$$

1.2

To minimise

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n,\mathbf{w}) - t_n\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2, \tag{1.7}$$

setting its derivative to zero gives

$$\mathbf{0} = \sum_{n=1}^{N} \frac{\partial y(x_n, \mathbf{w})}{\partial \mathbf{w}} \left(y(x_n, \mathbf{w}) - t_n\right) + \lambda \mathbf{w}. \tag{1.8}$$

Substituting

$$y(x_n, \mathbf{w}) = \sum_{j=0}^{M} w_j x_n^j \tag{1.9}$$

gives

$$0 = \sum_{n=1}^{N} x_n^i \left(\sum_{j=0}^{M} w_j x_n^j - t_n \right) + \lambda w_i. \tag{1.10}$$

Therefore,

$$\sum_{j=0}^{M} \tilde{A}_{ij} w_j = T_i \tag{1.11}$$

where

$$\begin{aligned} \tilde{A}_{ij} &= \sum_{n=1}^{N} x_n^{i+j} + \lambda \delta_{ij}, \\ T_i &= \sum_{n=1}^{N} x_n^i t_n. \end{aligned} \tag{1.12}$$

1.3

Let a, o and l be the events where an apple, orange and lime are selected respectively. The probability that an apple is selected is given by

$$p(a) = p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g). \tag{1.13}$$

Substituting $p(a|r) = \frac{3}{10}$, $p(r) = \frac{1}{5}$, $p(a|g) = \frac{1}{2}$, $p(r) = \frac{1}{5}$, $p(a|g) = \frac{3}{10}$ and $p(g) = \frac{3}{5}$ gives

$$p(a) = \frac{17}{50}. \tag{1.14}$$

If an orange is selected, the probability that it came from the geen box is given by

$$p(g|o) = \frac{p(g, o)}{p(o)}. \tag{1.15}$$

Here,

$$\begin{aligned} p(g,o) &= p(o|g)p(g), \\ p(o) &= p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g). \end{aligned} \tag{1.16}$$

Substituting $p(o|r) = \frac{2}{5}$, $p(r) = \frac{1}{5}$, $p(o|b) = \frac{1}{2}$, $p(b) = \frac{1}{5}$, $p(o|g) = \frac{3}{10}$ and $p(g) = \frac{3}{5}$ gives $p(g,o) = \frac{9}{50}$ and $p(o) = \frac{9}{25}$. Therefore,

$$p(g|o) = \frac{1}{2}. \tag{1.17}$$

1.4

Let

$$x = g(y) \tag{1.18}$$

and $\hat{x}$ and $\hat{y}$ be the locations of the maximum of $p_x(x)$ and $p_y(y)$ respcetively. Let us assume that there exists $\epsilon > 0$ such that $g'(y) \neq 0$ for $|y - \hat{y}| < \epsilon$. Then, differentiating both sides of the transoformation

$$p_y(y) = p_x\left(g(y)\right)|g'(y)| \tag{1.19}$$

and substituting $y = \hat{y}$ gives

$$0 = g'(\hat{y})p'_x\left(g\left(\hat{y}\right)\right) + p_x\left(g\left(\hat{y}\right)\right)g''\left(\hat{y}\right). \tag{1.20}$$

Therefore, in general,

$$\hat{x} \neq g\left(\hat{y}\right). \tag{1.21}$$

Here, let us assume that

$$g(y) = ay + b. \tag{1.22}$$

Then, differentiating both sides of the transformation and substituting $y = \hat{y}$ gives

$$0 = p'_x\left(g\left(\hat{y}\right)\right). \tag{1.23}$$

Therefore,

$$\hat{x} = g\left(\hat{y}\right). \tag{1.24}$$

1.5

By the definition,

$$\mathrm{var} f(x) = \mathrm{E}\left(f(x) - \mathrm{E}f(x)\right)^2 . \tag{1.25}$$

The right hand side can be written as

$$\mathrm{E}\left((f(x))^2 - 2f(x)\mathrm{E}f(x) + (\mathrm{E}f(x))^2\right) = \mathrm{E}\left(f(x)\right)^2 - \left(\mathrm{E}f(x)\right)^2 . \tag{1.26}$$

Therefore,

$$\mathrm{var} f(x) = \mathrm{E}\left(f(x)\right)^2 - \left(\mathrm{E}f(x)\right)^2 . \tag{1.27}$$

1.6

By the definition,

$$\mathrm{cov}(x, y) = \mathrm{E}\left((x - \mathrm{E}x)\left(y - \mathrm{E}y\right)\right) . \tag{1.28}$$

The right hand side can be written as

$$\mathrm{E}xy - \mathrm{E}\left(x\mathrm{E}y\right) - \mathrm{E}\left(y\mathrm{E}x\right) + \mathrm{E}\left(\mathrm{E}x\mathrm{E}y\right) = \mathrm{E}xy - \mathrm{E}x\mathrm{E}y. \tag{1.29}$$

The right hand side can be written as

$$\int xyp(x, y)dxdy - \int xp(x)dx \int yp(y)dy. \tag{1.30}$$

If x and y are independent, by the definition,

$$f(x, y) = f(x)f(y). \tag{1.31}$$

Then,

$$\int xyp(x, y)dxdy = \int p(x)dx \int p(y)dy. \tag{1.32}$$

Therefore,

$$\mathrm{cov}(x, y) = 0. \tag{1.33}$$

1.7

Let

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx. \tag{1.34}$$

Then

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}\left(x^2+y^2\right)\right) dxdy. \tag{1.35}$$

By the transformation from Cartesian coordinates (x, y) to polar coordinates (r, θ), the right hand side can be written as

$$\int_0^{\infty}\int_0^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} drd\theta = 2\pi \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) rdr. \tag{1.36}$$

By the transformation $s = \frac{r}{\sigma}$, the right hand side can be written as

$$2\pi\sigma^2 \int_0^{\infty} \exp\left(-\frac{1}{2}s^2\right) sds = 2\pi\sigma^2 \left[-\exp\left(-\frac{1}{2}s^2\right)\right]_0^{\infty}. \tag{1.37}$$

Therefore,

$$I = \left(2\pi\sigma^2\right)^{\frac{1}{2}}. \tag{1.38}$$

By the definition,

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1.39}$$

Then

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) dx = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx. \tag{1.40}$$

By the transformation $t = x - \mu$, the right hand side can be written as

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}t^2\right) dt = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} I. \tag{1.41}$$

Therefore,

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) dx = 1. \tag{1.42}$$

1.8

If x is under the Gaussian distribution with mean μ and variance σ^2, then

$$\mathrm{E}x = \int_{-\infty}^{\infty} x\mathcal{N}\left(x|\mu, \sigma^2\right) dx. \tag{1.43}$$

By the definition, the right hand side can be written as

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx. \tag{1.44}$$

By the transformation $y = x - \mu$, it can be written as

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} (y+\mu) \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy. \tag{1.45}$$

Since

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} y \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy = 0, \tag{1.46}$$

and

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \mu \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy = \mu \int_{-\infty}^{\infty} \mathcal{N}\left(y|\mu, \sigma^2\right) dy, \tag{1.47}$$

we have

$$\mathrm{E}x = \mu. \tag{1.48}$$

By the definition,

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) dx = 1 \tag{1.49}$$

can be written as

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = 1. \tag{1.50}$$

Differentiating both sides with respect to σ^2 gives

$$\begin{aligned} &(2\pi)^{-\frac{1}{2}} \left(-\frac{1}{2}\right) \left(\sigma^2\right)^{-\frac{3}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\ &+ \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{2} \left(\sigma^2\right)^{-2} (x-\mu)^2 \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = 0. \end{aligned} \tag{1.51}$$

The left hand side can be written as

$$-\frac{1}{2}\left(\sigma^2\right)^{-1}\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)dx+\frac{1}{2}\left(\sigma^2\right)^{-2}\int_{-\infty}^{\infty}(x-\mu)^2\mathcal{N}\left(x|\mu,\sigma^2\right)dx$$
$$=-\frac{1}{2}\left(\sigma^2\right)^{-1}+\frac{1}{2}\left(\sigma^2\right)^{-2}\mathrm{var}x. \tag{1.52}$$

Therefore,

$$\mathrm{var}x=\sigma^2. \tag{1.53}$$

1.9

By the definition,

$$\mathcal{N}\left(x|\mu,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1.54}$$

Setting its derivative to zero gives

$$0=\left(2\pi\sigma^2\right)^{-\frac{1}{2}}\left(-\frac{1}{\sigma^2}(x-\mu)\right)\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1.55}$$

Therefore, the mode is given by μ.

Similarly, by the definition,

$$\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)=(2\pi)^{-\frac{D}{2}}\left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{1.56}$$

Setting its derivative to zero gives

$$\mathbf{0}=-\left(2\pi\right)^{-\frac{D}{2}}\left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}^{-1}+\left(\boldsymbol{\Sigma}^{-1}\right)^\intercal\right)(\mathbf{x}-\boldsymbol{\mu})\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{1.57}$$

Therefore, the mode is given by $\boldsymbol{\mu}$.

1.10

By the definition,

$$\mathrm{E}(x+y)=\int\int(x+y)p(x,y)dxdy. \tag{1.58}$$

The right hand side can be written as

$$\int x \left(\int p(x, y) dy \right) dx + \int y \left(\int p(x, y) dx \right) dy = \int xp(x) dx + \int yp(y) dy. \tag{1.59}$$

By the definition, the right hand side can be written as

$$\mathrm{E}x + \mathrm{E}y. \tag{1.60}$$

Therefore,

$$\mathrm{E}(x + y) = \mathrm{E}x + \mathrm{E}y. \tag{1.61}$$

Similarly, by the definition,

$$\mathrm{var}(x + y) = \mathrm{E}\left(x + y - \mathrm{E}(x + y)\right)^2 \tag{1.62}$$

By the result above and the definition, the right hand side can be written as

$$\begin{aligned} \mathrm{E}\left(x - \mathrm{E}x\right)^2 + 2\mathrm{E}\left(\left(x - \mathrm{E}x\right)\left(y - \mathrm{E}y\right)\right) + \mathrm{E}\left(y - \mathrm{E}y\right)^2 \\ = \mathrm{var}x + 2\mathrm{cov}(x, y) + \mathrm{var}y. \end{aligned} \tag{1.63}$$

If x and y are independent, then

$$\mathrm{cov}(x, y) = 0, \tag{1.64}$$

by 1.6. Therefore,

$$\mathrm{var}(x + y) = \mathrm{var}x + \mathrm{var}y. \tag{1.65}$$

1.11

To maximise

$$\ln p\left(\mathbf{x}|\mu, \sigma^2\right) = -\frac{N}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2, \tag{1.66}$$

setting the partial derivatives with respect to μ and σ^2 to zero gives

$$\begin{aligned} 0 &= \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu), \\ 0 &= -\frac{N}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2} \sum_{n=1}^{N} (x_n - \mu)^2. \end{aligned} \tag{1.67}$$

Therefore,

$$\begin{aligned} \mu_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N} x_n, \\ \sigma^2_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2. \end{aligned} \tag{1.68}$$

1.12

If x_m and x_n are independent, then

$$\mathrm{E}x_m x_n = \mathrm{E}x_m \mathrm{E}x_n. \tag{1.69}$$

If they are samples from the Gaussian distribution with mean μ and variance σ^2, the right hand side is given by μ^2. On the other hand, by the definition,

$$\mathrm{E}x_n^2 = \mathrm{var}x_n + (\mathrm{E}x_n)^2. \tag{1.70}$$

If x_n is a sample from the Gaussian distribution with mean μ and variance σ^2, the right hand side is given by $\sigma^2 + \mu^2$. Therefore,

$$\mathrm{E}x_m x_n = \mu^2 + \delta_{mn}\sigma^2. \tag{1.71}$$

Here, since

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n, \tag{1.72}$$

we have

$$\mathrm{E}\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}x_n. \tag{1.73}$$

Therefore,

$$\mathrm{E}\mu_{\mathrm{ML}} = \mu. \tag{1.74}$$

Similarly, since

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2, \tag{1.75}$$

we have

$$\mathrm{E}\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}(x_n - \mu_{\mathrm{ML}})^2. \tag{1.76}$$

The right hand side can be writen as

$$\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\left(x_n^2-2\mu_{\mathrm{ML}}x_n+\mu_{\mathrm{ML}}^2\right)=\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}x_n^2-\frac{2}{N}\mathrm{E}\left(\mu_{\mathrm{ML}}\left(\sum_{n=1}^{N}x_n\right)\right)+\mathrm{E}\mu_{\mathrm{ML}}^2. \tag{1.77}$$

The first term of the right hand side can be written as

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2\right)=\mu^2+\sigma^2, \tag{1.78}$$

while the second and third terms can be writen as

$$-2\mathrm{E}\mu_{\mathrm{ML}}^2+\mathrm{E}\mu_{\mathrm{ML}}^2=-\mathrm{E}\mu_{\mathrm{ML}}^2. \tag{1.79}$$

Here,

$$\mathrm{E}\mu_{\mathrm{ML}}^2=\mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)^2. \tag{1.80}$$

The right hand side can be written as

$$\frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}x_n^2+\frac{2}{N^2}\sum_{1\le m<n\le N}\mathrm{E}x_mx_n=\frac{1}{N}(\mu^2+\sigma^2)+\frac{N-1}{N}\mu^2. \tag{1.81}$$

Therefore,

$$\mathrm{E}\mu_{\mathrm{ML}}^2=\mu^2+\frac{1}{N}\sigma^2. \tag{1.82}$$

Thus,

$$\mathrm{E}\sigma_{\mathrm{ML}}^2=\frac{N-1}{N}\sigma^2. \tag{1.83}$$

1.13

It is clear that

$$\mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2\right)=\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\left(x_n-\mu\right)^2. \tag{1.84}$$

The right hand side can be writen as

$$\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\left(x_n^2-2\mu x_n+\mu^2\right)=\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}x_n^2-\frac{2\mu}{N}\sum_{n=1}^{N}\mathrm{E}x_n+\mu^2. \tag{1.85}$$

The first term of the right hand side can be written as

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2\right)=\mu^2+\sigma^2, \tag{1.86}$$

while the second term can be writen as

$$-\frac{2\mu}{N}\sum_{n=1}^{N}\mu=-2\mu^2. \tag{1.87}$$

Therefore,

$$\mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2\right)=\sigma^2. \tag{1.88}$$

1.14

Let

$$\begin{aligned} w_{ij}^{\mathrm{S}} &= \frac{1}{2}(w_{ij}+w_{ji}), \\ w_{ij}^{\mathrm{A}} &= \frac{1}{2}(w_{ij}-w_{ji}). \end{aligned} \tag{1.89}$$

Then

$$\begin{aligned} w_{ij} &= w_{ij}^{\mathrm{S}}+w_{ij}^{\mathrm{A}}, \\ w_{ij}^{\mathrm{S}} &= w_{ji}^{\mathrm{S}}, \\ w_{ij}^{\mathrm{A}} &= -w_{ji}^{\mathrm{A}}. \end{aligned} \tag{1.90}$$

Here,

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{\mathrm{A}}x_ix_j=\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}-w_{ji})x_ix_j. \tag{1.91}$$

The right hand side can be written as

$$\frac{1}{2}\left(\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}x_ix_j-\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ji}x_ix_j\right)=0. \tag{1.92}$$

Therefore,

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{\mathrm{A}}x_ix_j=0. \tag{1.93}$$

Additionally,

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j = \sum_{i=1}^{D}\sum_{j=1}^{D}\left(w_{ij}^{\mathrm{S}} + w_{ij}^{\mathrm{A}}\right)x_ix_j. \tag{1.94}$$

The right hand side can be written as

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{S}}x_ix_j + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{A}}x_ix_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{S}}x_ix_j, \tag{1.95}$$

where the result above is used. Therefore,

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{S}}x_ix_j. \tag{1.96}$$

Finally, since the matrix w_{ij}^{S} is $D \times D$ symmetric matrix, its number of independent parameters is $\frac{D(D+1)}{2}$.

1.17

By the definition,

$$\Gamma(x) = \int_0^\infty u^{x-1}\exp(-u)du. \tag{1.97}$$

Then

$$\Gamma(x+1) = \int_0^\infty u^{x}\exp(-u)du. \tag{1.98}$$

The right hand side can be written as

$$\left[-u^x\exp(-u)\right]_{u=0}^{u=\infty} + \int_0^\infty xu^{x-1}\exp(-u)du = x\Gamma(x). \tag{1.99}$$

Therefore,

$$\Gamma(x+1) = x\Gamma(x). \tag{1.100}$$

By the definition,

$$\Gamma(1) = \int_0^1 \exp(-u)du. \tag{1.101}$$

The right hand side can be written as

$$[-\exp(-u)]_0^\infty = 1. \tag{1.102}$$

Therefore,

$$\Gamma(1) = 0!. \tag{1.103}$$

For a positive integer x, let us assume that

$$\Gamma(x) = (x-1)!. \tag{1.104}$$

Then,

$$\Gamma(x+1) = x\Gamma(x), \tag{1.105}$$

where the right hand side can be written as

$$x(x-1)! = x!. \tag{1.106}$$

Therefore,

$$\Gamma(x+1) = x!. \tag{1.107}$$

Thus, the assumption is proved by induction on x.

1.18

Let us consider the transformation from Cartesian to polar coordinates

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} \exp\left(-x_i^2\right) dx_i = S_D \int_0^\infty \exp\left(-r^2\right) r^{D-1} dr, \tag{1.108}$$

where S_D is the surface area of a sphere of unit raidus in D dimensions. By 1.7, the left hand side can be written as $\pi^{\frac{D}{2}}$. By the transformation $s = r^2$, the right hand side can be written as

$$\frac{S_D}{2} \int_0^\infty \exp(-s) s^{\frac{D-1}{2}} s^{-\frac{1}{2}} ds = \frac{S_D}{2} \Gamma\left(\frac{D}{2}\right). \tag{1.109}$$

Therefore,

$$S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}\right)}. \tag{1.110}$$

Additionally, the volume of the sphere can can be written as

$$V_D = S_D \int_0^1 r^{D-1} dr. \tag{1.111}$$

The right hand side can be written as

$$S_D \left[\frac{r^D}{D} \right]_{r=0}^{r=1} = \frac{S_D}{D}. \tag{1.112}$$

Therefore,

$$V_D = \frac{S_D}{D}. \tag{1.113}$$

Finally, the results above reduce to

$$\begin{aligned} S_2 &= \frac{2\pi}{\Gamma(1)}, \\ V_2 &= \frac{S_2}{2}. \end{aligned} \tag{1.114}$$

Therefore,

$$\begin{aligned} S_2 &= 2\pi, \\ V_2 &= \pi. \end{aligned} \tag{1.115}$$

Similarly,

$$\begin{aligned} S_3 &= \frac{2\pi^{\frac{3}{2}}}{\Gamma\left(\frac{3}{2}\right)}, \\ V_3 &= \frac{S_3}{3}. \end{aligned} \tag{1.116}$$

Therefore,

$$\begin{aligned} S_3 &= 4\pi, \\ V_3 &= \frac{4}{3}\pi. \end{aligned} \tag{1.117}$$

1.19

The volume of a cube of side 2 in D dimensions is 2^D. Therefore, the ratio of the volume of the cocentric sphere of radius 1 divided by the volume of the cube is given by

$$\frac{V_D}{2^D} = \frac{\pi^{\frac{D}{2}}}{D 2^{D-1} \Gamma\left(\frac{D}{2}\right)}, \tag{1.118}$$

by 1.18.

Additionally, by Stering's formula

$$\Gamma(x+1) \simeq (2\pi)^{\frac{1}{2}} \exp(-x) x^{\frac{x+1}{2}}, \tag{1.119}$$

the ratio can be approximated as

$$\frac{V_D}{2^D} \simeq \frac{\pi^{\frac{D}{2}}}{D 2^{D-1} (2\pi)^{\frac{1}{2}} \exp\left(1 - \frac{D}{2}\right) \left(\frac{D}{2} - 1\right)^{\frac{D}{4}}}. \tag{1.120}$$

The right hand side can be written as

$$\frac{1}{2e(2\pi)^{\frac{1}{2}}} \frac{1}{D} \left(\frac{e^2 \pi^2}{8D - 16}\right)^{\frac{D}{4}}. \tag{1.121}$$

Therefore, the ratio goes to zero as $D \to \infty$.

Finally, the ratio of the distance from the center of the cube to one of the corners divided by the perpendicular distance to one of the sides is given by

$$\frac{\sqrt{\sum_{i=1}^{D} 1^2}}{1} = \sqrt{D}. \tag{1.122}$$

Therefore, the ration goes to ∞ as $D \to \infty$.

1.20

Integrating both sides of a Gaussian distributino in D dimensions

$$p(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right) \tag{1.123}$$

from $\|\mathbf{x}\| = r$ to $\|\mathbf{x}\| = r + \epsilon$ gives

$$\int_{r \le \|\mathbf{x}\| \le r+\epsilon} p(\mathbf{x}) d\mathbf{x} = \int_{r}^{r+\epsilon} \int (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r'^2}{2\sigma^2}\right) J dr' d\boldsymbol{\phi}, \tag{1.124}$$

where $\boldsymbol{\phi}$ is the vector of the angular components of the polar corrdinate and J is the Jacobian of the transformation from the Cartesian to polar coordinate.

For a sufficiently small ϵ, the right hand side can be approximated as

$$\begin{aligned}&(2\pi\sigma^2)^{-\frac{D}{2}}\exp\left(-\frac{r^2}{2\sigma^2}\right)\int_r^{r+\epsilon}\int J dr' d\boldsymbol{\phi}\\&=(2\pi\sigma^2)^{-\frac{D}{2}}\exp\left(-\frac{r^2}{2\sigma^2}\right)\int_{r\le\|\mathbf{x}\|\le r+\epsilon}d\mathbf{x}.\end{aligned}\tag{1.125}$$

Therefore,

$$\int_{r\le\|\mathbf{x}\|\le r+\epsilon}p(\mathbf{x})d\mathbf{x}\simeq p(r)\epsilon,\tag{1.126}$$

where

$$p(r)=(2\pi\sigma^2)^{-\frac{D}{2}}S_D r^{D-1}\exp\left(-\frac{r^2}{2\sigma^2}\right),\tag{1.127}$$

and S_D is the surface area of a unit sphere in D dimensions.

Secondly, to maximise $p(r)$, setting the derivative to zero gives

$$0=(2\pi\sigma^2)^{-\frac{D}{2}}S_D\left((D-1)r^{D-2}-\frac{r^D}{\sigma^2}\right)\exp\left(-\frac{r^2}{2\sigma^2}\right).\tag{1.128}$$

Therefore, $p(r)$ is maximised at a sigle stationary point

$$\hat{r}=\sqrt{D-1}\sigma.\tag{1.129}$$

Thirdly, by the expression of $p(r)$ above,

$$\frac{p\left(\hat{r}+\epsilon\right)}{p\left(\hat{r}\right)}=\left(\frac{\hat{r}+\epsilon}{\hat{r}}\right)^{D-1}\exp\left(-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right).\tag{1.130}$$

Using the expression of $\hat{r}$ above, the right hand side can be written as

$$\begin{aligned}&\exp\left((D-1)\ln\left(1+\frac{\epsilon}{\hat{r}}\right)-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right)\\&=\exp\left(\frac{\hat{r}^2}{\sigma^2}\ln\left(1+\frac{\epsilon}{\hat{r}}\right)-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right).\end{aligned}\tag{1.131}$$

By the Taylor series

$$\ln(1+x)=x-\frac{1}{2}x^2+o\left(x^3\right),\tag{1.132}$$

the right hand side can be approximated as

$$\exp\left(\frac{\hat{r}^2}{\sigma^2}\left(\frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}\right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right) = \exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \tag{1.133}$$

Therefore,

$$p\left(\hat{r} + \epsilon\right) \simeq p\left(\hat{r}\right)\exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \tag{1.134}$$

Finally, let a vector of length $\hat{r}$ be $\hat{\mathbf{r}}$. Then, by the definition of $p(\mathbf{x})$,

$$\frac{p(\mathbf{0})}{p\left(\hat{\mathbf{r}}\right)} = \exp\left(\frac{\hat{r}^2}{2\sigma^2}\right). \tag{1.135}$$

Substituting the expression of $\hat{r}$ above, the right hand side can be written as $\exp\left(\frac{D-1}{2}\right)$. Therefore,

$$\frac{p(\mathbf{0})}{p\left(\hat{\mathbf{r}}\right)} = \exp\left(\frac{D-1}{2}\right). \tag{1.136}$$

1.21

If $0 \leq a \leq b$, then

$$0 \leq a(b-a). \tag{1.137}$$

Therefore,

$$a \leq (ab)^{\frac{1}{2}}. \tag{1.138}$$

For a two-class classification problem of $\mathbf{x}$, let the classes be $\mathcal{C}_1$ and $\mathcal{C}_2$ and let the decision regions be $\mathcal{R}_1$ and $\mathcal{R}_2$. Let us choose the decision regions to minimise the probability of misclassification. Then,

$$p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2) \Rightarrow \mathbf{x} \in \mathcal{C}_1, \tag{1.139}$$

and

$$p(\mathbf{x}, \mathcal{C}_2) > p(\mathbf{x}, \mathcal{C}_1) \Rightarrow \mathbf{x} \in \mathcal{C}_2. \tag{1.140}$$

Then, using the inequality above,

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} \leq \int_{\mathcal{R}_1} \left(p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2)\right)^{\frac{1}{2}} d\mathbf{x}, \tag{1.141}$$

and

$$\int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \leq \int_{\mathcal{R}_2} \left(p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2)\right)^{\frac{1}{2}} d\mathbf{x}. \tag{1.142}$$

Therefore,

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \leq \int \left(p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2)\right)^{\frac{1}{2}} d\mathbf{x}. \tag{1.143}$$

1.22

Let

$$\mathrm{E}L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \tag{1.144}$$

If

$$L_{kj} = 1 - \delta_{kj}, \tag{1.145}$$

then the right hand side can be written as

$$\sum_k \sum_j \int_{\mathcal{R}_j} \left(p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j)\right) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left(\sum_k p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j) \right) d\mathbf{x}. \tag{1.146}$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \left(p(\mathbf{x}) - p(\mathbf{x}, \mathcal{C}_j)\right) d\mathbf{x} = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathbf{x}, \mathcal{C}_j) d\mathbf{x}. \tag{1.147}$$

Therefore,

$$\mathrm{E}L = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathcal{C}_j | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.148}$$

Thus, minimising $\mathrm{E}L$ reduces to choosing the criterion to maximise the posterior probatility $p(\mathcal{C}_j|\mathbf{x})$.

1.23

Let

$$\mathrm{E}L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \tag{1.149}$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left(\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \tag{1.150}$$

Therefore,

$$\mathrm{E}L = \sum_j \int_{\mathcal{R}_j} \left(\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \tag{1.151}$$

Thus, mimising $\mathrm{E}L$ reduces to choosing to minimise $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$.

1.25

Let

$$\mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right) = \int\int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{1.152}$$

Then

$$\delta \mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right) = \int\int \left(\|\mathbf{y}(\mathbf{x}) + \delta\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 - \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2\right) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{1.153}$$

The right hand side can be written as

$$\int\int \left(2\left(\delta\mathbf{y}(\mathbf{x})\right)^\intercal \left(\mathbf{y}(\mathbf{x}) - \mathbf{t}\right) + \left(\delta\mathbf{y}(\mathbf{x})\right)^\intercal \delta\mathbf{y}(\mathbf{x})\right) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{1.154}$$

Therefore,

$$\frac{\delta \mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right)}{\delta\mathbf{y}(\mathbf{x})} = 2\int \left(\mathbf{y}(\mathbf{x}) - \mathbf{t} + \delta\mathbf{y}(\mathbf{x})\right) p(\mathbf{x}, \mathbf{t}) d\mathbf{t}. \tag{1.155}$$

To minimise $\mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right)$, setting the left hand side to zero and assuming that $\delta\mathbf{y}(\mathbf{x})$ is sufficiently small gives

$$\mathbf{0} = \int \left(\mathbf{y}(\mathbf{x}) - \mathbf{t}\right) p(\mathbf{x}, \mathbf{t}) d\mathbf{t}. \tag{1.156}$$

The right hand side can be written as

$$\mathbf{y}(\mathbf{x}) \int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} - \int \mathbf{t} p(\mathbf{x}, \mathbf{t}) d\mathbf{t} = p(\mathbf{x}) \left(\mathbf{y}(\mathbf{x}) - \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}\right). \tag{1.157}$$

Thus,

$$\mathbf{y}(\mathbf{x}) = \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \tag{1.158}$$

Finally, for a single target variable t, it reduces to

$$\mathbf{y}(\mathbf{x}) = \mathrm{E}_t(t|\mathbf{x}). \tag{1.159}$$

1.26

Let

$$\mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right) = \int\int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{1.160}$$

The right hand side can be written as

$$\begin{aligned}\int\int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) + \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x},\mathbf{t}) d\mathbf{x} d\mathbf{t} \\ = \int\int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x},\mathbf{t}) d\mathbf{x} d\mathbf{t} \\ +2\int\int (\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\intercal (\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t})\, p(\mathbf{x},\mathbf{t}) d\mathbf{x} d\mathbf{t} \\ + \int\int \|\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x},\mathbf{t}) d\mathbf{x} d\mathbf{t}.\end{aligned} \tag{1.161}$$

Let us look at each term of the right hand side. The first term can be written as

$$\int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 \left(\int p(\mathbf{x},\mathbf{t}) d\mathbf{t}\right) d\mathbf{x} = \int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x}. \tag{1.162}$$

Additionally, the second term can be written as

$$2\int (\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\intercal \left(\int (\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t})\, p(\mathbf{t}|\mathbf{x}) d\mathbf{t}\right) p(\mathbf{x}) d\mathbf{x}. \tag{1.163}$$

Since

$$\begin{aligned}\int \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} = \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \frac{\int p(\mathbf{x},\mathbf{t}) d\mathbf{t}}{p(\mathbf{x})}, \\ \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} = \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}),\end{aligned} \tag{1.164}$$

the second term is zero. Finally, the third term can be written as

$$\int \left(\int \|\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{t}|\mathbf{x}) d\mathbf{t}\right) p(\mathbf{x}) d\mathbf{x} = \int \mathrm{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.165}$$

Therefore,

$$\mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right) = \int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x} + \int \mathrm{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.166}$$

Thus, $\mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right)$ is mimimised if

$$\mathbf{y}(\mathbf{x}) = \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \tag{1.167}$$