# Derivations of Equations for The Elements of Statistical Learning

by

Soong Lee

August, 2024

This document is a collection of derivations of non-trivial equations and statements from ESL (Second Edition). I did not include the equations that were assigned as exercises, since the solutions of them are available from the resources in the internet.

1. ESL Solutions by Yuhang Zhou (github/YuhangZhou88)
2. A Solution Manual and Notes for ESL by J. L. Weatherwax and D. Epstein
3. A Guide and Solution Manual to ESL by J. C. Ma

I used the same mathematical notation as in ESL. However, on a few occasions I used boldface lower characters for vector notation, like when referring Wikipedia or Matrix Cookbook.

References:

1. The Matrix Cookbook (Nov 2012) by K. B. Peterson and M. S. Pederson (https://www2.imm.dtu.dk/pubdb/pubs/3274-full.html)
2. A Solution Manual and Notes for ESL (October 2021) by J. L. Weatherwax and D. Epstein
3. Pattern Recognition and Machine Learning (February 2006) by C. M. Bishop

# Chapter 2. Overview of Supervised Learning

## Eq 2.16: (ESL p.19)

$$\beta = [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(XY)$$

**Proof** :

Utilizing Matrix Cookbook Eq (78):

$$\frac{\partial(\mathbf{Bx}+\mathbf{b})^T\mathbf{C}(\mathbf{Dx}+\mathbf{d})}{\partial\mathbf{x}} = \mathbf{B}^T\mathbf{C}(\mathbf{D}x+\mathbf{d}) + \mathbf{D}^T\mathbf{C}^T(\mathbf{Bx}+\mathbf{b})$$

$$\begin{aligned}\mathrm{EPE}(f) &= \mathrm{E}(Y-f(X))^2\\ &= \mathrm{E}[(Y-X^T\beta)^T(Y-X^T\beta)]\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\beta}[(Y-X^T\beta)^T(Y-X^T\beta)] &= (-X^T)^T(-X^T\beta+Y) + (-X^T)^T(Y-X^T\beta)\\ &= -2X(Y-X^T\beta) = 0\end{aligned}$$

$$\Rightarrow \quad \mathrm{E}(XY) = \mathrm{E}(XX^T)\beta$$

$$\therefore \ \ \beta = [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(XY)$$

## Eq 2.22: (ESL p.21)

$$\hat{G}(x) = \mathrm{argmin}_{g\in\mathcal{G}}[1-\Pr(g|X=x)]$$

**Proof** :

$$\hat{G}(x) = \mathrm{argmin}_{g\in\mathcal{G}}\sum_{k=1}^{K} L(\mathcal{G}_k, g)\Pr(\mathcal{G}_k|X=x) \tag{2.21}$$

0-1 loss function means that

$$L(\mathcal{G}_k, g) = \begin{cases} 1 & \text{when } \mathcal{G}_k \neq g \\ 0 & \text{else}\end{cases}$$

$$\sum_{k=1}^{K} L(\mathcal{G}_k, g)\Pr(\mathcal{G}_k|X=x) = \Pr(\mathcal{G}_1|X=x) + \Pr(\mathcal{G}_2|X=x) + \cdots + 0 \cdot \Pr(g|X=x)$$
$$+ \Pr(\mathcal{G}_{g+1}|X=x) + \cdots + \Pr(\mathcal{G}_k|X=x)$$

Since
$$\sum_{k=1}^{K} \Pr(\mathcal{G}_k|X=x) = 1$$
$$\therefore \quad \sum_{k=1}^{K} L(\mathcal{G}_k, g)\Pr(\mathcal{G}_k|X=x) = 1 - \Pr(g|X=x)$$

## Eq 2.25: (ESL p.24)

$$\begin{aligned}\mathrm{MSE}(x_0) &= \mathrm{E}_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 \\ &= \mathrm{E}_{\mathcal{T}}[\hat{y}_0 - \mathrm{E}_{\mathcal{T}}(\hat{y}_0)]^2 + [\mathrm{E}_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2\end{aligned}$$

**Proof** :

$$\begin{aligned}\mathrm{E}_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 &= \mathrm{E}_{\mathcal{T}}[f(x_0) - \hat{y}_0 + \mathrm{E}_{\mathcal{T}}(\hat{y}_0) - \mathrm{E}_{\mathcal{T}}(\hat{y}_0)]^2 \\ &= \mathrm{E}_{\mathcal{T}}[(\mathrm{E}_{\mathcal{T}}(\hat{y}_0) - \hat{y}_0) + (f(x_0) - \mathrm{E}_{\mathcal{T}}(\hat{y}_0))]^2 \\ &= \mathrm{E}_{\mathcal{T}}[\hat{y}_0 - \mathrm{E}_{\mathcal{T}}(\hat{y}_0)]^2 + \mathrm{E}_{\mathcal{T}}[\mathrm{E}_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2 \\ &\quad + 2\mathrm{E}_{\mathcal{T}}[\mathrm{E}_{\mathcal{T}}(\hat{y}_0) - \hat{y}_0] \cdot \mathrm{E}_{\mathcal{T}}[f(x_0) - \mathrm{E}_{\mathcal{T}}(\hat{y}_0)] \\ (\text{Since } \mathrm{E}_{\mathcal{T}}[\mathrm{E}_{\mathcal{T}}(\hat{y}_0) - \hat{y}_0] &= \mathrm{E}_{\mathcal{T}}(\hat{y}_0) - \mathrm{E}_{\mathcal{T}}(\hat{y}_0) = 0) \\ &= \mathrm{E}_{\mathcal{T}}[\hat{y}_0 - \mathrm{E}_{\mathcal{T}}(\hat{y}_0)]^2 + \mathrm{E}_{\mathcal{T}}[\mathrm{E}_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2\end{aligned}$$

## ESL p.24

"For an arbitray test point $x_0$, we have $\hat{y}_0 = x_0^T\hat{\beta}$, which can be written as $\hat{y}_0 = x_0^T\beta + \sum_{i=1}^{N} l_i(x_0)\epsilon_i$, where $l_i(x_0)$ is the i-th element of $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0$."

**Proof** :

$$\hat{y}_0 = x_0^T\beta$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad \leftarrow \text{ from the method of least squares} \tag{2.6}$$

$$Y = X\beta + \epsilon \tag{2.26}$$

Since $X^T = [X_1 X_2, \cdots, X_p]$ $\quad$ ($\leftarrow$ see ESL p.10.)

$$[X_1, X_2, \cdots, X_p]\beta = X^T\beta \quad \Rightarrow \quad Y = X^T\beta$$

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & & & \\ X_{N1} & X_{N2} & \cdots & X_{Np} \end{bmatrix} \quad \Rightarrow \quad \mathbf{y} = \mathbf{X}\beta$$

$\beta$ becomes,

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon) = \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

$$\hat{y}_0 = x_0^T\hat{\beta} = x_0^T\beta + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

To match the dimension compared to the sum notation, the second term is transposed,

$$[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0$$

# Chapter 3. Linear Methods for Regression

## Eq 3.12: (ESL p.48)

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

**Proof** :

$$z_j = \frac{\hat{\beta}_j - 0}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} \qquad (\leftarrow \; z = \frac{\overline{x} - m}{\sigma/\sqrt{n}})$$

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \tag{3.8}$$

where $\sigma^2$ is the variance of the observations $y_i$'s.

$$\Rightarrow \quad z_j = \frac{\hat{\beta}_j}{\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}\hat{\sigma}^2}} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}}$$

where

$$\hat{\sigma} = \frac{1}{N-P-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad \leftarrow \text{ estimate of } \sigma^2$$

$v_j$ = j-th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$

$$\therefore \;\; z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

## ESL p.53:

"Then it is easy to check that the multiple least squares estimates $\hat{\beta}_j$ are equal to $< \mathbf{x}_j, \mathbf{y} > / < \mathbf{x}_j, \mathbf{x}_j >$ - the univariate estimates."

**Proof** :

X: N×p matrix

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.6}$$

Since $\mathbf{X}^T\mathbf{X}$ is p×p dimension, and $\mathbf{X}^T\mathbf{y}$ is p×1, so $\hat{\beta}$ is p×1.

$$\mathbf{x}_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{N1} \end{bmatrix} \qquad : \quad \text{N data collection of first component of } \mathbf{x}$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_1 & \cdots & \mathbf{x}_p \end{bmatrix} \qquad : \quad \text{N×p}$$

$$\mathbf{X}^T = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_p^T \end{bmatrix} \qquad : \quad \text{p×N}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_p^T \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_1 & \cdots & \mathbf{x}_p \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{x}_2 & \cdots & \mathbf{x}_1^T\mathbf{x}_p \\ \mathbf{x}_2^T\mathbf{x}_1 & \mathbf{x}_2^T\mathbf{x}_2 & \cdots & \mathbf{x}_2^T\mathbf{x}_p \\ \vdots \\ \mathbf{x}_p^T\mathbf{x}_1 & \mathbf{x}_p^T\mathbf{x}_2 & \cdots & \mathbf{x}_p^T\mathbf{x}_p \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_p^T \end{bmatrix} \mathbf{y} \ = \begin{bmatrix} \mathbf{x}_1^T\mathbf{y} \\ \mathbf{x}_2^T\mathbf{y} \\ \vdots \\ \mathbf{x}_p^T\mathbf{y} \end{bmatrix}$$

If $< \mathbf{x}_j, \mathbf{x}_k >= 0$ for $j \neq k$,

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T\mathbf{x}_1 & & & \\ & \mathbf{x}_2^T\mathbf{x}_2 & & \\ & & \ddots & \\ & & & \mathbf{x}_p^T\mathbf{x}_p \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} (\mathbf{x}_1^T\mathbf{x}_1)^{-1} & & & \\ & (\mathbf{x}_2^T\mathbf{x}_2)^{-1} & & \\ & & \ddots & \\ & & & (\mathbf{x}_p^T\mathbf{x}_p)^{-1} \end{bmatrix}$$

Now $\hat{\beta}$ can be calculated using the above equations,

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \begin{bmatrix} \frac{\mathbf{x}_1^T\mathbf{y}}{\mathbf{x}_1^T\mathbf{x}_1} & & & \\ & \frac{\mathbf{x}_2^T\mathbf{y}}{\mathbf{x}_2^T\mathbf{x}_2} & & \\ & & \ddots & \\ & & & \frac{\mathbf{x}_p^T\mathbf{y}}{\mathbf{x}_p^T\mathbf{x}_p} \end{bmatrix} \end{aligned}$$

$$\therefore \ \ \hat{\beta}_j = \frac{< \mathbf{x}_j, \mathbf{y} >}{< \mathbf{x}_j, \mathbf{x}_j >}$$

**Eq 3.27:** (ESL p.53)

$$\hat{\beta}_1 = \frac{< \mathbf{x} - \overline{x}\mathbf{1}, \mathbf{y} >}{< \mathbf{x} - \overline{x}\mathbf{1}, \mathbf{x} - \overline{x}\mathbf{1} >}$$

**Proof** :

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \tag{3.3}$$

When there is an intercept,

$$Y = X\beta + \beta_0 + \epsilon$$

For univariate case

$$y = \mathbf{x}_1\beta_1 + \mathbf{x}_0\beta_0 + \epsilon$$

where

$$\mathbf{x}_0 = 1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \leftarrow \ N \times 1$$

$$\text{RSS}(\beta) = (y - \mathbf{x}_0\beta_0 - \mathbf{x}_1\beta_1)^T(\mathbf{y} - \mathbf{x}_0\beta - \mathbf{x}_1\beta_1)$$

To find minimum RSS w.r.t. $\beta_0$ and $\beta_1$,

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta_0} = -2\mathbf{x}_0^T(\mathbf{y} - \mathbf{x}_0\beta_0 - \mathbf{x}_1\beta_1) = 0$$

$$\Rightarrow \quad -\sum_{i=1}^{N} y_i + N\beta_0 + \sum_{i=1}^{N} x_{1i}\beta_1 = 0 \tag{1}$$

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta_1} = -2\mathbf{x}_1^T(\mathbf{y} - \mathbf{x}_0\beta_0 - \mathbf{x}_1\beta_1) = 0$$

$$\Rightarrow \quad -\mathbf{x}_1^T\mathbf{y} + \sum_{i=1}^{N} x_{1i}\beta_0 + \sum_{i=1}^{N} x_{1i}^2\beta_1 = 0 \tag{2}$$

Defining $\overline{x} = \dfrac{1}{N}\displaystyle\sum_1^N x_{1i}$

Eq (1) becomes
$$-\sum_1^N y_i + N\beta_0 + N\overline{x}\beta_1 = 0 \tag{3}$$

Eq (2) becomes
$$-\mathbf{x}_1^T\mathbf{y} + N\overline{x}\beta_0 + \sum_1^N x_{1i}^2\beta_1 = 0 \tag{4}$$

To eliminate $\beta_0$ from Eqs (3) and (4),

Eq (3) $\times\,\overline{x}$ – Eq (4):

$$\Rightarrow \quad -\mathbf{x}_1^T\mathbf{y} + \sum_1^N \overline{x}y_i + \sum_1^N x_{1i}^2\beta_1 - N\overline{x}^2\beta_1 = 0$$

$$\begin{aligned}\Rightarrow \quad \beta_1 &= \frac{x_1^T\mathbf{y} - \sum_1^N \overline{x}y_i}{\sum_1^N x_{1i}^2 - N\overline{x}^2} \\ &= \frac{\sum_1^N x_{1i}y_i - \sum_1^N \overline{x}y_i}{\sum_1^N x_{1i}^2 - \sum_1^N \overline{x}^2} \\ &= \frac{\sum_1^N (x_{1i}-\overline{x})y_i}{\sum_1^N (x_{1i}^2 - \overline{x}^2)}\end{aligned}$$

To show $<\mathbf{x}_1 - \overline{x}\,\mathbf{1}, \mathbf{x}_1 - \overline{x}\,\mathbf{1}> = \displaystyle\sum_1^N (x_{1i}^2 - \overline{x}^2)$,

$$\begin{aligned}<\mathbf{x}_1 - \overline{x}\,\mathbf{1}, \mathbf{x}_1 - \overline{x}\,\mathbf{1}> &= (\mathbf{x}_1 - \overline{x}\,\mathbf{1})^T \cdot (\mathbf{x}_1 - \overline{x}\,\mathbf{1}) \\ &= (x_{11}-\overline{x},\, x_{12}-\overline{x},\, \cdots, x_{1N}-\overline{x}) \begin{pmatrix} x_{11}-\overline{x} \\ x_{12}-\overline{x} \\ \vdots \\ x_{1N}-\overline{x} \end{pmatrix} \\ &= (x_{11}-\overline{x})^2 + (x_{12}-\overline{x})^2 + \cdots + (x_{1N}-\overline{x})^2 \\ &= (x_{11}^2 + x_{12}^2 + \cdots + x_{1N}^2) - 2\overline{x}(x_{11} + x_{12} + \cdots + x_{1N}) + \overline{x}^2 N \\ &= \sum_1^N x_{1i}^2 - N\overline{x}^2 \\ &= \sum_1^N (x_{1i}^2 - \overline{x}^2)\end{aligned}$$

$$\therefore \quad \hat{\beta}_1 = \frac{< \mathbf{x} - \overline{x}\,\mathbf{1}, \mathbf{y} >}{< \mathbf{x} - \overline{x}\,\mathbf{1}, \mathbf{x} - \overline{x}\,\mathbf{1} >}$$

(Here $\mathbf{x}$ is $\mathbf{x}_1$ above.)

## Eq 3.31: (ESL p.55)

$$\mathbf{X} = \mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\boldsymbol{\Gamma} = \mathbf{Q}\mathbf{R}$$

$$\text{where } \mathbf{Q} = \mathbf{Z}\mathbf{D}^{-1},\ \mathbf{R} = \mathbf{D}\boldsymbol{\Gamma}$$

**Proof** :

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} \tag{3.30}$$

$\boldsymbol{\Gamma}$: upper triangular matrix

$$D_{jj} = \|\mathbf{z}_j\|$$

$\mathbf{R} = \mathbf{D}\boldsymbol{\Gamma}$ an upper triangular matrix, since $\boldsymbol{\Gamma}$ is an upper triangular matrix and $\mathbf{D}$ is a diagonal matrix.

We need to show $\mathbf{Q} = \mathbf{Z}\mathbf{D}^{-1}$ is an orthonormal matrix.

$$\mathbf{Q}^T\mathbf{Q} = (\mathbf{Z}\mathbf{D}^{-1})^T(\mathbf{Z}\mathbf{D}^{-1}) = \mathbf{D}^{-1}\mathbf{Z}^T\mathbf{Z}\mathbf{D}^{-1}$$

$$\mathbf{Z} = (\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_p) \qquad (\leftarrow\ N \times (p+1))$$

$$\text{where } \mathbf{z}_i = \begin{bmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{Ni} \end{bmatrix}$$

$$\mathbf{D}^{-1} = \begin{bmatrix} |\mathbf{z}_0|^{-1} & & & \\ & |\mathbf{z}_1|^{-1} & & \\ & & \ddots & \\ & & & |\mathbf{z}_p|^{-1} \end{bmatrix} \qquad \leftarrow (p+1) \times (p+1)$$

$$\mathbf{Z}^T\mathbf{Z} = \begin{bmatrix} \mathbf{z}_0^T \\ \mathbf{z}_1^T \\ \vdots \\ \mathbf{z}_p^T \end{bmatrix} \begin{bmatrix} \mathbf{z}_0^T & \mathbf{z}_1^T & \cdots & \mathbf{z}_p^T \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{z}_0^T\mathbf{z}_0 & \mathbf{z}_0^T\mathbf{z}_1 & \cdots & \mathbf{z}_0^T\mathbf{z}_p \\ \mathbf{z}_1^T\mathbf{z}_0 & \mathbf{z}_1^T\mathbf{z}_1 & \cdots & \mathbf{z}_1^T\mathbf{z}_p \\ \vdots & & & \\ \mathbf{z}_p^T\mathbf{z}_0 & \mathbf{z}_p^T\mathbf{z}_1 & \cdots & \mathbf{z}_p^T\mathbf{z}_p \end{bmatrix}$$

Using the above two equations for $\mathbf{D}^{-1}$ and $\mathbf{Z}^T\mathbf{Z}$,

$$\mathbf{D}^{-1}(\mathbf{Z}^T\mathbf{Z}) = \begin{bmatrix} |\mathbf{z}_0|^{-1} & & & \\ & |\mathbf{z}_1|^{-1} & & \\ & & \ddots & \\ & & & |\mathbf{z}_p|^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{z}_0^T\mathbf{z}_0 & \mathbf{z}_0^T\mathbf{z}_1 & \cdots & \mathbf{z}_0^T\mathbf{z}_p \\ \mathbf{z}_1^T\mathbf{z}_0 & \mathbf{z}_1^T\mathbf{z}_1 & \cdots & \mathbf{z}_1^T\mathbf{z}_p \\ \vdots & & & \\ \mathbf{z}_p^T\mathbf{z}_0 & \mathbf{z}_p^T\mathbf{z}_1 & \cdots & \mathbf{z}_p^T\mathbf{z}_p \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\mathbf{z}_0^T\mathbf{z}_0}{|\mathbf{z}_0|} & & & \\ & \frac{\mathbf{z}_1^T\mathbf{z}_1}{|\mathbf{z}_1|} & & \\ & & \ddots & \\ & & & \frac{\mathbf{z}_p^T\mathbf{z}_p}{|\mathbf{z}_p|} \end{bmatrix}$$

Finally,

$$(\mathbf{D}^{-1}\mathbf{Z}^T\mathbf{Z})\mathbf{D}^{-1} = \begin{bmatrix} |\mathbf{z}_0| & & & \\ & |\mathbf{z}_1| & & \\ & & \ddots & \\ & & & |\mathbf{z}_p| \end{bmatrix} \begin{bmatrix} |\mathbf{z}_0|^{-1} & & & \\ & |\mathbf{z}_1|^{-1} & & \\ & & \ddots & \\ & & & |\mathbf{z}_p|^{-1} \end{bmatrix}$$

$$= \mathbf{I}$$

## Eq 3.32 & 3.33: (ESL p.55)

$$\hat{\beta} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}$$
$$\hat{\mathbf{y}} = \mathbf{Q}\mathbf{Q}^T\mathbf{y}$$

**Proof** :

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.6}$$
$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \tag{3.7}$$
$$\mathbf{X} = \mathbf{Q}\mathbf{R} \tag{3.31}$$

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= [(\mathbf{Q}\mathbf{R})^T\mathbf{Q}\mathbf{R}]^{-1}(\mathbf{Q}\mathbf{R})^T\mathbf{y} \\
&= (\mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\
&\text{(since } \mathbf{Q}^T\mathbf{Q} = \mathbf{I}) \\
&= (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\
&= \mathbf{R}^{-1}(\mathbf{R}^T)^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\
&\text{(since } (\mathbf{R}^T)^{-1}\mathbf{R}^T = \mathbf{I}) \\
&= \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}
\end{aligned}$$

$$\therefore \quad \hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = (\mathbf{Q}\mathbf{R})\mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y} = \mathbf{Q}\mathbf{Q}^T\mathbf{y}$$

## Eq 3.46: (ESL p.66)

$$\begin{aligned}
\mathbf{X}\hat{\beta}^{\text{ls}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{U}\mathbf{U}^T\mathbf{y}
\end{aligned}$$

**Proof** :

$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$

$\mathbf{U} : N \times p$ orthonormal $\quad \rightarrow \mathbf{U}^T\mathbf{U} = \mathbf{I}$

$\mathbf{D} : p \times p$ diagonal

$\mathbf{V} : p \times p$ orthonormal $\quad \rightarrow \mathbf{V}^T\mathbf{V} = \mathbf{I}$

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= (\mathbf{U}\mathbf{D}\mathbf{V}^T)^T\mathbf{U}\mathbf{D}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T \\ &\text{(since } \mathbf{U}^T\mathbf{U} = \mathbf{I}) \\ &= \mathbf{V}\mathbf{D}^T\mathbf{D}\mathbf{V}^T\end{aligned}$$

If $\mathbf{A}$ and $\mathbf{B}$ are square matrices, $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
Since $\mathbf{V}\mathbf{D}^T$ and $\mathbf{D}\mathbf{V}^T$ are square matrices,

$$(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{V}\mathbf{D}^T\mathbf{D}\mathbf{V}^T)^{-1} = (\mathbf{D}\mathbf{V}^T)^{-1}(\mathbf{V}\mathbf{D}^T)^{-1}$$

$$\begin{aligned}\therefore\quad \mathbf{X}\hat{\beta}^{\text{ls}} &= \mathbf{U}\mathbf{D}\mathbf{V}^T(\mathbf{D}\mathbf{V}^T)^{-1}(\mathbf{V}\mathbf{D}^T)^{-1}(\mathbf{U}\mathbf{D}\mathbf{V}^T)^T\mathbf{y} \\ &= \mathbf{U}\mathbf{U}^T\mathbf{y}\end{aligned}$$

## Eq 3.47: (ESL p.66)

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ridge}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y}\end{aligned}$$

**Proof** :
Using $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^T\mathbf{D}\mathbf{V}^T$ from the derivation of Eq (3.46),

$$\begin{aligned}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} &= \mathbf{V}\mathbf{D}^T\mathbf{D}\mathbf{V}^T + \lambda\mathbf{I}\mathbf{V}\mathbf{V}^T \\ &= \mathbf{V}(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})\mathbf{V}^T\end{aligned}$$

Since all three matrices have $p \times p$ dimension; $\mathbf{V},\ (\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I}),\ \mathbf{V}^T$.

$$\Rightarrow\quad (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} = (\mathbf{V}^T)^{-1}(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{V}^{-1}$$

$$\therefore\ \ \mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{U}\mathbf{D}\mathbf{V}^T)\left[(\mathbf{V}^T)^{-1}(\mathbf{D}^T\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{V}^{-1}\right](\mathbf{V}\mathbf{D}^T\mathbf{U}^T)\mathbf{y}$$
$$= \mathbf{U}\mathbf{D}(\mathbf{D}^2+\lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y}$$

## Eq 3.49: (ESL p.66)

$$\mathrm{Var}(\mathbf{z}_1) = \mathrm{Var}(\mathbf{X}v_1) = \frac{d_1^2}{N}$$

**Proof** :

We know that

$$\mathrm{Covar}(X) = \frac{\mathbf{X}^T\mathbf{X}}{N} = \frac{\mathbf{V}\mathbf{D}^2\mathbf{V}^T}{N}$$

Let's calculate $\mathrm{Var}(v_1^T\mathbf{X})$ instead of $\mathrm{Var}(\mathbf{X}v_1)$, since they are the same.

$$\begin{aligned}
\mathrm{Var}(v_1\mathbf{X}) &= \mathrm{E}[(v_1^T\mathbf{X} - v_1^T\mathbf{X})(v_1^T\mathbf{X} - v_1^T\mathbf{X})^T] \\
&= v_1^T\mathrm{E}[(\mathbf{X}-\mathbf{X})(\mathbf{X}-\mathbf{X})^T]v_1 \\
&= v_1^T\mathrm{Covar}(\mathbf{X})v_1 \\
&= v_1^T\frac{\mathbf{V}\mathbf{D}^2\mathbf{V}^T}{N}v_1 \\
&= (v_1^T\mathbf{V})\frac{\mathbf{D}^2}{N}(\mathbf{V}^Tv_1) \\
&= \frac{1}{N}\begin{bmatrix}1 & 0 & 0 & \cdots & 0\end{bmatrix}\begin{bmatrix}d_1^2 & & & \\ & d_2^2 & & \\ & & \ddots & \\ & & & d_p^2\end{bmatrix}\begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix} \\
&= \frac{1}{N}\begin{bmatrix}d_1^2 & 0 & 0 & \cdots & 0\end{bmatrix}\begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix} \\
&= \frac{1}{N}d_1^2
\end{aligned}$$

# Chapter 5. Basis Expansions and Regularization

## Eq 5.18: (ESL p.154)

$$\text{RSS} = (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \lambda \mathbf{f}^T \mathbf{K} \mathbf{f}$$

**Proof** :

From Exercise 5.9,

$$\begin{aligned} \mathbf{K} &= \mathbf{N}^{-1}\mathbf{\Omega}_N \mathbf{N}^{-1} \\ \Rightarrow \quad \mathbf{\Omega}_N &= \mathbf{N}^T \mathbf{K} \mathbf{N} \end{aligned}$$

$$\text{RSS} = (\mathbf{y} - \mathbf{N}\theta)^T(\mathbf{y} - \mathbf{N}\theta) + \lambda \theta^T \mathbf{\Omega}_N \theta \tag{5.11}$$

Since $\mathbf{f} = \mathbf{N}\theta$ (Eq 5.13),

$$\begin{aligned} \text{RSS} &= (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \lambda \theta^T (\mathbf{N}^T \mathbf{K} \mathbf{N})\theta \\ &= (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \lambda (\mathbf{N}\theta)^T \mathbf{K} (\mathbf{N}\theta) \end{aligned}$$

$$\therefore \quad \text{RSS} = (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \lambda \mathbf{f}^T \mathbf{K} \mathbf{f}$$

## Eq 5.31 & 5.32: (ESL p.162)

$$\begin{aligned} \frac{\partial l(\theta)}{\partial \theta} &= \mathbf{N}^T(\mathbf{y} - \mathbf{p}) - \lambda \mathbf{\Omega}\theta \\ \frac{\partial^2 l(\theta)}{\partial \theta\, \partial \theta^T} &= -\mathbf{N}^T \mathbf{W} \mathbf{N} - \lambda \mathbf{\Omega} \end{aligned}$$

**Proof** :

$$l(f;\lambda) = \sum_{i=1}^{N} [y_i f(x_i) - \log(1 + e^{f(x_i)})] - \frac{1}{2}\lambda \int \{f''(t)\}^2 dt \tag{5.30}$$

$$\begin{aligned} \Rightarrow \quad l(\theta) &= \mathbf{y}\mathbf{N}\theta - \log(1 + e^{\mathbf{N}\theta}) - \frac{1}{2}\lambda \int \theta^T \mathbf{N}''^T \mathbf{N}\theta dt \\ \frac{\partial l(\theta)}{\partial \theta} &= \mathbf{y}\mathbf{N} - \frac{Ne^{N\theta}}{1 + e^{\mathbf{N}\theta}} - \lambda \int \theta^T \mathbf{N}''^T \mathbf{N}'' dt \end{aligned}$$

Since $\mathbf{p} = \dfrac{e^{\mathbf{N}\theta}}{1+e^{\mathbf{N}\theta}}$,

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \theta} &= \mathbf{y}^T\mathbf{N} - \mathbf{N}\mathbf{p} - \lambda\mathbf{\Omega}\theta \\
&= \mathbf{N}^T(\mathbf{y}-\mathbf{p}) - \lambda\mathbf{\Omega}\theta \\
\text{where } \mathbf{\Omega} &= \int \mathbf{N}''^T\mathbf{N}''dt \qquad \text{(see Eq 5.11)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l(\theta)}{\partial\theta\,\partial\theta^T} &= -\mathbf{N}\frac{\mathbf{N}\theta e^{\mathbf{N}\theta}}{1+e^{\mathbf{N}\theta}} + \frac{\mathbf{N}e^{\mathbf{N}\theta}\mathbf{N}e^{\mathbf{N}\theta}}{(1+e^{\mathbf{N}\theta})^2} - \lambda\mathbf{\Omega} \\
&= -\mathbf{N}^2\mathbf{p} + \mathbf{N}^2\mathbf{p}\mathbf{p}^T - \lambda\mathbf{\Omega} \\
&= -\mathbf{N}^T\mathbf{p}(1-\mathbf{p})^T\mathbf{N} - \lambda\mathbf{\Omega} \\
&= -\mathbf{N}^T\mathbf{W}\mathbf{N} - \lambda\mathbf{\Omega} \\
\text{where } \mathbf{W} &= \mathbf{p}(1-\mathbf{p})^T
\end{aligned}$$

## Eq 5.33: (ESL p.162)

$$\theta^{new} = (\mathbf{N}^T\mathbf{W}\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}^T\mathbf{W}(\mathbf{N}\theta^{old} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p}))$$

**Proof** :

We want to find $\theta$ for $\dfrac{\partial l(\theta)}{\partial\theta} = 0 \quad \text{(Eq 5.31)}$.

The Newton-Raphson method is to find f(x) = 0, so in our case $f(x) = \dfrac{\partial l(\theta)}{\partial\theta}$.

The Newton-Raphson method for f(x) says,

$$\mathbf{w}_1 = \mathbf{w}_0 - \frac{f(\mathbf{w}_0)}{f'(\mathbf{w}_0)}$$

So in our case it will be

$$\begin{aligned}
\theta^{new} &= \theta^{old} - \frac{\left(\dfrac{\partial l(\theta)}{\partial\theta}\right)}{\left(\dfrac{\partial^2 l(\theta)}{\partial\theta\,\partial\theta^T}\right)} \\
&= \theta^{old} + \frac{\mathbf{N}^T(\mathbf{y}-\mathbf{p}) - \lambda\mathbf{\Omega}\theta^{old}}{\mathbf{N}^T\mathbf{W}\mathbf{N} - \lambda\mathbf{\Omega}} \\
&= \frac{\theta^{old}(\mathbf{N}^T\mathbf{W}\mathbf{N} - \lambda\mathbf{\Omega}) + \mathbf{N}^T(\mathbf{y}-\mathbf{p}) - \lambda\mathbf{\Omega}\theta^{old}}{\mathbf{N}^T\mathbf{W}\mathbf{N} - \lambda\mathbf{\Omega}}
\end{aligned}$$

$$\begin{aligned}
\text{Numerator } &= \mathbf{N}^T\mathbf{W}[\theta^{old}(\mathbf{N} - (\mathbf{N}^T\mathbf{W})^{-1}\lambda\mathbf{\Omega}) + (\mathbf{N}^T\mathbf{W})^{-1}\mathbf{N}^T(\mathbf{y}-\mathbf{p}) - (\mathbf{N}^T\mathbf{W})^{-1}\lambda\mathbf{\Omega}\theta^{old}] \\
&= \mathbf{N}^T\mathbf{W}[\theta^{old}(\mathbf{N} - \mathbf{W}^{-1}(\mathbf{N}^T)^{-1}\lambda\mathbf{\Omega}) + \mathbf{W}^{-1}(\mathbf{N}^T)^{-1}\mathbf{N}^T(\mathbf{y}-\mathbf{p}) \\
&\quad - \mathbf{W}^{-1}(\mathbf{N}^T)^{-1}\lambda\mathbf{\Omega}\theta^{old}] \\
&= \mathbf{N}^T\mathbf{W}(\mathbf{N}\theta^{old} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p}))
\end{aligned}$$

$$\therefore \ \theta^{new} = (\mathbf{N}^T\mathbf{W}\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}^T\mathbf{W}(\mathbf{N}\theta^{old} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p}))$$

## ESL p.187 :

"If we adopt the convention that $B_{i,1} = 0$ if $\tau_i = \tau_{i+1}$, then by induction $B_{i,m} = 0$ if $\tau_i = \tau_{i+1} = \cdots = \tau_{i+m}$."

**Proof** : (Ref: Wikipedia/B-Spline)

$$B_{i,m}(x) = \frac{x-\tau_i}{\tau_{i+m-1}-\tau_i}B_{i,m-1}(x) + \frac{\tau_{i+m}-x}{\tau_{i+m}-\tau_{i+1}}B_{i+1,m-1}(x) \tag{5.78}$$

To prove this, the notation at the Wiki is more convenient.

$$B_{i,k+1}(x) = w_{i,k}B_{i,k}(x) + [1 - w_{i+1,k}(x)]B_{i+1,k}(x) \tag{1}$$

$$\text{where } \ w_{i,k}(x) = \begin{cases} \dfrac{x-\tau_i}{\tau_{i+k}-\tau_i} & \text{if } \tau_{i+k} \neq \tau_i \\ 0 & \text{otherwise} \end{cases}$$

Let's show Eq (5.78) and Eq (1) are equivalent.

If we use m = k + 1 in Eq (5.78), it becomes

$$B_{i,k+1} = \frac{x-\tau_i}{\tau_{i+k}-\tau_i}B_{i,k} + \frac{\tau_{i+k+1}-x}{\tau_{i+k+1}-\tau_{i+1}}B_{i+1,k} \tag{2}$$

$$\text{If we define } \frac{x-\tau_i}{\tau_{i+k}-\tau_i} \equiv w_{i,k}, \text{ then}$$

$$1 - w_{i+1,k} = 1 - \frac{x-\tau_{i+1}}{\tau_{i+k+1}-\tau_{i+1}} = \frac{\tau_{i+k+1}-x}{\tau_{i+k+1}-\tau_{i+1}}$$

This is the same as the second term in Eq (2). Therefore Eq (5.78) and Eq (1) are equivalent. Now let's prove $B_{i,m} = 0$ if $\tau_i = \tau_{i+1} = \cdots = \tau_{i+m}$. We already have $B_{i,1} = 0$ if $\tau_i = \tau_{i+1}$ from convention.

Let's show $B_{i,2} = 0$ if $\tau_i = \tau_{i+1} = \tau_{i+2}$.

From Eq (1), when k = 1,

$$B_{i,2} = w_{i,1}B_{i,1} + [1 - w_{i+1,1}]B_{i+1,1} = 0$$

$B_{i+1,1} = 0$ if $\tau_i = \tau_{i+1}$. This is simply from $B_{i,1} = 0$ if $\tau_i = \tau_{i+1}$. (Since i is any number.)

We keep doing this until k = m - 1.

$$B_{i,m} = w_{i,m-1}B_{i,m-1} + [1 - w_{i+1,m-1}]B_{i+1,m-1}$$

where we know $B_{i,m-1} = 0$.

We have to show $B_{i+1,m-1} = 0$ if $\tau_i = \tau_{i+1} = \cdots = \tau_{i+m}$. So far we have $B_{i,m-1} = 0$ if $\tau_i = \tau_{i+1} = \cdots = \tau_{i+m-1}$. If we simply use $i \to i+1$, then $B_{i+1,m-1} = 0$ if $\tau_i = \tau_{i+1} = \cdots = \tau_{(i+1)+m-1}$.

$$\therefore \;\; B_{i,m} = 0 \;\text{ if }\; \tau_i = \tau_{i+1} = \cdots = \tau_{i+m}$$

# Chapter 7. Model Assessment and Selection

## Eq 7.10: (ESL p.223)

$$\begin{aligned}\mathrm{Err}(x_0) &= \mathrm{E}[(Y - \hat{f}_k(x_0))^2 | X = x_0] \\ &= \sigma_\epsilon^2 + \left[ f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)}) \right]^2 + \frac{\sigma_\epsilon^2}{k}\end{aligned}$$

**Proof** :

This is a K-Nearest Neighbor problem. The important thing to understand about this problem is that the inputs $x_i$ are fixed for simplicity. When $x_i$'s are fixed in KNN, there will be only $y_i$ differences due to the intrinsic noise $\epsilon$.

$$Y = f(X) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

$$\Rightarrow \quad y_i = f(x_i) + \epsilon$$

where Y is a random variable and $x_i$ is an input.

The prediction per a sample will be

$$\begin{aligned}\hat{f}(x_i) &= \frac{1}{k}\sum_{l=1}^{k} y_{(l)} \qquad (l \text{ is NN for } x_i)\\ &= \frac{1}{k}\sum_{l=1}^{k}[f(x_{(l)}) + \epsilon_{(l)}]\end{aligned}$$

Expected prediction over the samples ($\mathcal{T}$) will be,

$$\mathrm{E}_{\mathcal{T}}\hat{f}(x_i) = \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})$$

This solves the second term in Eq (7.10).

As for the third term $\sigma_\epsilon^2/k$,

$$\begin{aligned}\mathrm{Var}_{\mathcal{T}}(\hat{f}(x_i)) &= \mathrm{Var}_{\mathcal{T}}\left[\frac{1}{k}\sum_{l=1}^{k} f(x_{(l)}) + \frac{1}{k}\sum_{l=1}^{k}\epsilon_{(l)}\right]\\ &= \mathrm{Var}_{\mathcal{T}}\left(\frac{1}{k}\sum_{l=1}^{k}\epsilon_{(l)}\right)\\ &= \frac{1}{k^2}\mathrm{Var}_{\mathcal{T}}(\epsilon_{(1)} + \epsilon_{(2)} + \cdots + \epsilon_{(k)})\\ &= \frac{1}{k^2}(k\sigma_\epsilon^2) \qquad (\leftarrow \mathrm{Var}_{\mathcal{T}}(\epsilon_{(l)}) = \sigma_\epsilon^2)\\ &= \frac{1}{k}\sigma_\epsilon^2\end{aligned}$$

## **Eq 7.11:** (ESL p.224)

$$\begin{aligned}\mathrm{Err}(x_0) &= \mathrm{E}[(Y - \hat{f}_p(x_0))^2 | X = x_0]\\ &= \sigma_\epsilon^2 + [f(x_0) - \mathrm{E}\hat{f}_p(x_0)]^2 + \|\mathbf{h}(x_0)\|^2\sigma_\epsilon^2\\ \text{where } \mathbf{h}(x_0) &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0\end{aligned}$$

**Proof** :

For a linear model, we have $y = X\beta + \epsilon$.

$$\begin{aligned}
\hat{f}_p(x) &= x^T\hat{\beta} \\
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \qquad \text{(from least squares)} && (3.6)\\
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon) \\
&= \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon
\end{aligned}$$

We have $\mathrm{E}_{\mathcal{T}}(\epsilon) = 0$ and $\mathrm{E}(\epsilon\epsilon^T) = \sigma^2\mathbf{I}$.

$$\Rightarrow \quad \mathrm{E}_{\mathcal{T}}(\hat{\beta}) = \beta$$

$$\begin{aligned}
\mathrm{Var}_{\mathcal{T}}(\hat{\beta}) &= \mathrm{E}_{\mathcal{T}}(\hat{\beta}\hat{\beta}^T) - \mathrm{E}_{\mathcal{T}}(\hat{\beta})\mathrm{E}_{\mathcal{T}}(\hat{\beta}^T) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \qquad \text{(See Weatherwax p.9)} \\
\mathrm{Var}_{\mathcal{T}}(\hat{y}_0) &= \mathrm{Var}_{\mathcal{T}}(x_0^T\hat{\beta}) \\
&= x_0^T\mathrm{Var}_{\mathcal{T}}(\hat{\beta})x_0 \\
&= x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 x_0
\end{aligned}$$

Let's calculate $\mathbf{h}(x_0)^T \cdot \mathbf{h}(x_0)$,

$$\begin{aligned}
[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0]^T\,[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0] &= x_0^T[(\mathbf{X}^T\mathbf{X})^{-1}]^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0 \\
\text{(Since } [(\mathbf{X}^T\mathbf{X})^{-1}]^T = [(\mathbf{X}^T\mathbf{X})^T]^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}) \\
&= x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0
\end{aligned}$$

$$\therefore \quad \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) = \mathrm{Var}_{\mathcal{T}}(\hat{f}_p(x_0)) = \|\mathbf{h}(x_0)\|^2\sigma_\epsilon^2$$

## Eq 7.41: (ESL p.235)

$$p(\mathcal{M}_m|\mathbf{Z}) \approx \frac{e^{-\frac{1}{2}\mathrm{BIC}_m}}{\sum_{l=1}^{M} e^{-\frac{1}{2}\mathrm{BIC}_l}}$$

**Proof** :

$$\log p(\mathbf{Z}|\mathcal{M}_m) = \log p(\mathbf{Z}|\hat{\theta}_m, \mathcal{M}_m) - \frac{d_m}{2}\log N + O(1) \qquad (7.40)$$

$$\mathrm{BIC} = -2\,\mathrm{loglik} + (\log N)d \qquad (7.35)$$

If we define our loss function to be $-2\log p(\mathbf{Z}|\hat{\theta}_m, \mathcal{M}_m)$, Eq (7.40) becomes after dropping O(1), (multiply by -2)

$$-2\log p(\mathbf{Z}|\mathcal{M}_m) = -2\log p(\mathbf{Z}|\hat{\theta}_m, \mathcal{M}_m) + d_m \log N$$

where $\log p(\mathbf{Z}|\mathcal{M}_m) = \text{BIC}$, and $\log p(\mathbf{Z}|\hat{\theta}_m, \mathcal{M}_m) = \text{loglik}$.

$$\Rightarrow \quad \text{BIC}_m = -2\,\text{loglik} + d_m \log N$$

This is same as Eq (7.35).

Since $\text{BIC}_m = -2\log p(\mathbf{Z}|\mathcal{M}_m)$,

$$p(\mathbf{Z}|\mathcal{M}_m) = \exp\left(-\frac{1}{2}\text{BIC}_m\right)$$

$$\begin{aligned}\text{Posterior} &= \frac{p(\mathcal{M}_m)\cdot\exp\left(-\frac{1}{2}\text{BIC}_m\right)}{p(\mathbf{Z})} \\ &= \frac{p(\mathcal{M}_m)\cdot\exp\left(-\frac{1}{2}\text{BIC}_m\right)}{\sum_{l=1}^{M} p(\mathbf{Z}|\mathcal{M}_l)} \\ &= \frac{p(\mathcal{M}_m)\cdot\exp\left(-\frac{1}{2}\text{BIC}_m\right)}{\sum_{l=1}^{M} \exp\left(-\frac{1}{2}\text{BIC}_l\right)}\end{aligned}$$

## Eq 7.59: (ESL p.252)

$$\hat{\gamma} = \hat{p}_1(1-\hat{q}_1) + (1-\hat{p}_1)\hat{q}_1$$

**Proof** :

$$\begin{aligned}\hat{\gamma} &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N} L(y_i, \hat{f}(x_{i'})) \\ &= \frac{1}{N^2}\left[\sum_{i\in p_1}\sum_{i'\in q_0} L(y_i, \hat{f}(x_{i'})) + \sum_{i\in p_0}\sum_{i'\in q_1} L(y_i, \hat{f}(x_{i'}))\right]\end{aligned}$$

In the above equation, L = 0 for the cases of i and i′ other than the values defined in the equation. Since L = 1 in the above equation,

$$\begin{aligned}\hat{\gamma} &= \frac{1}{N^2}\left[\sum_{i\in p_1}\sum_{i'\in q_0} 1 + \sum_{i\in p_0}\sum_{i'\in q_1} 1\right] \\ &= \frac{1}{N}\sum_{i\in p_1}\left(\sum_{i'\in q_0}\frac{1}{N}\right) + \frac{1}{N}\sum_{i\in p_0}\left(\sum_{i'\in q_1}\frac{1}{N}\right)\end{aligned}$$

Since $\sum_{i' \in q_0} \frac{1}{N} = 1 - \hat{q}_1$, and $\sum_{i' \in q_1} \frac{1}{N} = \hat{q}_1$,

$$\hat{\gamma} = \frac{1}{N} \sum_{i \in p_1} (1 - \hat{q}_1) + \frac{1}{N} \sum_{i \in p_0} (\hat{q}_1)$$

Since $\frac{1}{N} \sum_{i \in p_1} = \hat{p}_1$, and $\frac{1}{N} \sum_{i \in p_0} = 1 - \hat{p}_1$,

$$\therefore \quad \hat{\gamma} = \hat{p}_1 (1 - \hat{q}_1) + (1 - \hat{p}_1)(\hat{q}_1)$$

# Chapter 8. Model Inference and Averaging

## Eq 8.30: (ESL p.271)

$$p(\theta|z) \sim \mathcal{N}\left(\frac{z}{1 + 1/\tau}, \frac{1}{1 + 1/\tau}\right)$$

**Proof** :

$$p(z) \sim \mathcal{N}(\theta, 1) \tag{8.29}$$

Since Eq (8.29) is equivalent to $p(z|\theta)$,

$$\Rightarrow \quad p(z|\theta) = \frac{1}{(2\pi)^{1/2}} \cdot \exp\left\{-\frac{1}{2}(z - \theta)^2\right\} \qquad \text{likelihood}$$

$$\text{Posterior } p(\theta|z) = \frac{p(z|\theta) \cdot p(\theta)}{p(z)}$$

$$\begin{aligned}
p(\theta|z) &\sim p(z|\theta) \cdot p(\theta) \\
&= \frac{1}{(2\pi)^{1/2}} \cdot \exp\left\{-\frac{1}{2}(z - \theta)^2\right\} \frac{1}{(2\pi)^{1/2}\tau^{1/2}} \exp\left(-\frac{1}{2\tau}\theta^2\right) \\
&= \frac{1}{2\pi\, \tau^{1/2}} \cdot \exp\left\{-\frac{1}{2}(z - \theta)^2 - \frac{\theta^2}{2\tau}\right\} \\
&= \frac{1}{2\pi\, \tau^{1/2}} \cdot \exp\left\{-\frac{1}{2}\left[(z - \theta)^2 + \frac{\theta^2}{2\tau}\right]\right\}
\end{aligned}$$

$$
\begin{aligned}
(z-\theta)^2 + \frac{\theta^2}{2\tau} &= z^2 - 2z\theta + \theta^2 + \frac{\theta^2}{\tau} \\
&= \left(1 + \frac{1}{\tau}\right)\theta^2 - 2z\theta + z^2 \\
&= \left(1 + \frac{1}{\tau}\right)\left[\left(\theta - \frac{z}{1+1/\tau}\right)^2 + \frac{z^2}{(1+1/\tau)} - \frac{z^2}{(1+1/\tau)^2}\right] \\
\therefore\ \ p(\theta|z) &\sim \mathcal{N}\left(\frac{z}{1+1/\tau}, \frac{1}{1+1/\tau}\right)
\end{aligned}
$$

**Eq 8.54:** (ESL p.289)

$$
\mathrm{E}(\zeta|\mathbf{Z}) = \sum_{m=1}^{M} \mathrm{E}(\zeta|\mathcal{M}_m, \mathbf{Z}) p(\mathcal{M}_m|\mathbf{Z})
$$

**Proof** :

$$
\begin{aligned}
\mathrm{E}(\zeta|\mathbf{Z}) &= \int \zeta p(\zeta|\mathbf{Z}) d\zeta \\
&= \int \zeta \sum_m p(\zeta|\mathcal{M}_m, \mathbf{Z}) \cdot p(\mathcal{M}_m|\mathbf{Z}) d\zeta \\
&= \sum_m \int \zeta p(\zeta|\mathcal{M}_m, \mathbf{Z}) d\zeta \cdot p(\mathcal{M}_m|\mathbf{Z}) \\
(\text{Since } &\int \zeta p(\zeta|\mathcal{M}_m, \mathbf{Z}) d\zeta = \mathrm{E}(\zeta|\mathcal{M}_m, \mathbf{Z})) \\
&= \sum_m \mathrm{E}(\zeta|\mathcal{M}_m, \mathbf{Z}) \cdot p(\mathcal{M}_m|\mathbf{Z})
\end{aligned}
$$

# Chapter 9. Additive Models, Trees

**Eq 9.11:** (ESL p.307)

$$
\hat{c}_m = \mathrm{ave}(y_i | x_i \in R_m)
$$

**Proof** :

$$L = \sum_i (y_i - f(x_i))^2$$
$$\frac{\partial L}{\partial c_m} = \sum_i 2(y_i - f(x_i)) \cdot \left(-\frac{\partial f(x_i)}{\partial c_m}\right)$$
$$f(x_i) = \sum_{m=1}^{M} c_m I(x_i \in R_m) \tag{9.10}$$
$$\frac{\partial f(x_i)}{\partial c_m} = I(x_i \in R_m)$$
$$\frac{\partial L}{\partial c_m} = \sum_i 2(y_i - f(x_i)) \cdot (-I(x_i \in R_m)) = 0$$

$\Rightarrow$ Two conditions:

1. $I(x_i \in R_m) = 1$

2. $\sum_i (y_i - f(x_i)) = \sum_i \left[ y_i - \sum_{m=1}^{M} c_m I(x_i \in R_m) \right] = 0$

$$\Rightarrow \quad \sum_i [y_i | x_i \in R_m - C_m I(x_i \in R_m)] = 0$$
$$\Rightarrow \quad \sum_i [y_i | x_i \in R_m] = C_m \sum_i I(x_i \in R_m)$$

where $\sum_i I(x_i \in R_m) = \#$ of $x_i$'s belonging to $R_m$.

$$\therefore \quad c_m = \frac{\sum_i [y_i | x_i \in R_m]}{\# \text{ of } x_i\text{'s belonging to } R_m}$$
$$= \text{ave}(y_i | x_i \in R_m)$$

## ESL p. 309:

"For example, in a two-class problem with 400 observations in each class (denote this by (400, 400)), suppose one split created nodes (300, 100) and (100, 300), while the other created nodes (200, 400) and (200, 0). Both splits produce a mis-classification rate of 0.25."

**Proof** :

1. First case

$$\begin{tikzcd}[column sep=small] & \overset{c_1 \quad c_2}{(400, 400)} \arrow[dl] \arrow[dr] & \\ \underset{\text{node 1 } (R_1)}{\underset{c_1 \quad c_2}{(300, 100)}} & & \underset{\text{node 2 } (R_2)}{\underset{c_1 \quad c_2}{(100, 300)}} \end{tikzcd}$$

$N_{n_1} = N_{R_1} = 400$

$N_{n_2} = N_{R_2} = 400$

Misclassification error:

$$\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$$

Let's calculate $\hat{p}_{mk}$'s,

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{i \in R_m} I(y_i = k)$$

$n_1$: node 1, and $n_2$: node 2

$$\hat{p}_{n_1 c_1} = \frac{1}{400} \sum_{x_i \in R_1} I(y_i = c_1) = \frac{1}{400} \cdot 300 = \frac{3}{4}$$

$$\hat{p}_{n_1 c_2} = \frac{1}{400} \sum_{x_i \in R_1} I(y_i = c_2) = \frac{1}{400} \cdot 100 = \frac{1}{4}$$

$$\hat{p}_{n_2 c_1} = \frac{1}{400} \sum_{x_i \in R_2} I(y_i = c_1) = \frac{100}{400} = \frac{1}{4}$$

$$\hat{p}_{n_2 c_2} = 1 - \hat{p}_{n_2 c_1} = \frac{3}{4}$$

ME:

For node 1: $k(n_1) = c_1$

$$\frac{1}{N_{n_1}} \sum_{x_i \in R_1} I(y_i \neq k(n_1)) = \frac{1}{400} \sum_{x_i \in R_1} I(y_i \neq c_1) = \frac{1}{4}$$

For node 2: $k(n_2) = c_2$

$$\frac{1}{N_{n_2}} \sum_{x_i \in R_1} I(y_i \neq c_2) = \frac{1}{400} \cdot 100 = \frac{1}{4}$$

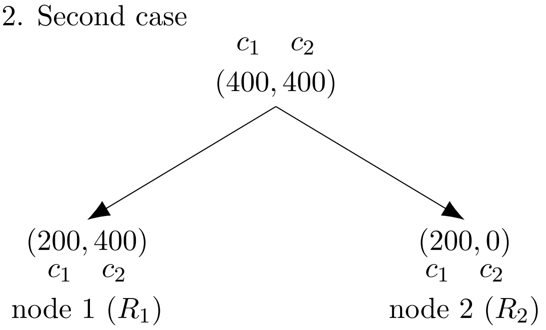
We can calculate this from $1 - \hat{p}_{mk(m)}$.

$$\text{For node 1,}\ \ \text{ME}_1 = 1 - \hat{p}_{n_1 k(n_1)} = 1 - \frac{3}{4} = \frac{1}{4}$$
$$\text{For node 2,}\ \ \text{ME}_2 = 1 - \hat{p}_{n_2 k(n_2)} = 1 - \frac{3}{4} = \frac{1}{4}$$

Now to calculate the combined ME, we use the suggestion given in p.309, saying that we need to weight the node impurity measures by the number $N_{m_L}$ and $N_{m_R}$ of observations in the two child nodes.

$$\begin{aligned}\text{ME}_{combined} &= \frac{N_{n_1}}{N_{n_1} + N_{n_2}} \cdot \text{ME}_1 + \frac{N_{n_2}}{N_{n_1} + N_{n_2}} \cdot \text{ME}_2 \\ &= \frac{400}{800} \cdot \frac{1}{4} + \frac{400}{800} \cdot \frac{1}{4} = \frac{1}{4}\end{aligned}$$

2. Second case




$N_{n_1} = 600$

$N_{n_2} = 200$

$$\begin{aligned}\hat{p}_{n_1 c_1} &= \frac{1}{600} \sum_{x_i \in R_1} I(y_i = c_1) = \frac{1}{600} \cdot 200 = \frac{1}{3} \\ \hat{p}_{n_1 c_2} &= \frac{2}{3} \\ \hat{p}_{n_2 c_1} &= \frac{1}{200} \sum_{x_i \in R_2} I(y_i = c_1) = \frac{1}{200} \cdot 200 = 1 \\ \hat{p}_{n_2 c_2} &= 0\end{aligned}$$

For node 1: $k(n_1) = c_2$

$$\text{ME}_1 = 1 - \frac{2}{3} = \frac{1}{3}$$

For node 2: $k(n_2) = c_1$

$$\text{ME}_2 = 1 - 1 = 0$$

$$\text{ME}_{combined} = \frac{600}{800} \cdot \frac{1}{3} + \frac{200}{800} \times 0 = \frac{1}{4}$$

where $\frac{600}{800}$ is from weight $= \frac{N_{n_1}}{N_{n_1} + N_{n_2}}$.

# Chapter 10. Boosting and Additive Trees

## Eq 10.11: (ESL p.344)

$$\sum_{i=1}^{N} w_i^{(m)} \exp\left(-\beta y_i G(x_i)\right) = (e^{\beta} - e^{-\beta}) \cdot \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)) + e^{-\beta} \sum_{i=1}^{N} w_i^{(m)}$$

**Proof** :

$$\begin{aligned}
\sum_{i=1}^{N} w_i^{(m)} \exp\left(-\beta y_i G(x_i)\right) &= e^{-\beta} \sum_{y_i = G(x_i)} w_i^{(m)} + e^{\beta} \sum_{y_i \neq G(x_i)} w_i^{(m)} \\
&= e^{-\beta} \sum_{i=1}^{N} w_i^{(m)} I(y_i = G(x_i)) + e^{\beta} \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)) \\
&= e^{-\beta} \sum_{i=1}^{N} w_i^{(m)} I(y_i = G(x_i)) + \left\{ e^{-\beta} \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)) \right. \\
&\quad \left. - e^{-\beta} \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)) \right\} + e^{\beta} \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)) \\
&= e^{-\beta} \sum_{i=1}^{N} w_i^{(m)} [I(y_i = G(x_i)) + I(y_i \neq G(x_i))] \\
&\quad + (e^{\beta} - e^{-\beta}) \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i))
\end{aligned}$$

Since $I(y_i = G(x_i)) + I(y_i \neq G(x_i)) = 1$,

$$\therefore \quad \sum_{i=1}^{N} w_i^{(m)} \exp\left(-\beta y_i G(x_i)\right) = e^{-\beta} \sum_{i=1}^{N} w_i^{(m)} + (e^{\beta} - e^{-\beta}) \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i))$$

**Eq 10.12:** (ESL p.344)

$$\beta_m = \frac{1}{2}\log\frac{1-\text{err}_m}{\text{err}_m}$$
$$\text{where } \text{err}_m = \frac{\sum_{i=1}^N w_i^{(m)} I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i^{(m)}}$$

**Proof** :

$$f(\beta) = (e^\beta - e^{-\beta}) \sum_{i=1}^N w_i^{(m)} I(y_i \neq G_m(x_i)) + e^{-\beta} \sum_{i=1}^N w_i^{(m)} \tag{10.11}$$
$$\frac{\partial f(\beta)}{\partial \beta} = \beta(e^\beta + e^{-\beta}) \sum_{i=1}^N w_i^{(m)} I(y_i \neq G_m(x_i)) - \beta e^{-\beta} \sum_{i=1}^N w_i^{(m)}$$
$$= 0$$

$$(e^{2\beta}+1) \sum_{i=1}^N w_i^{(m)} I(y_i \neq G_m(x_i)) - \sum_{i=1}^N w_i^{(m)} = 0$$
$$(e^{2\beta}+1) = \frac{1}{\text{err}_m}$$
$$\therefore \quad \beta_m = \frac{1}{2}\log\frac{1-\text{err}_m}{\text{err}_m}$$

**Eq 10.18:** (ESL p.346)

$$-l(Y, f(x)) = \log(1 + e^{-2Yf(x)})$$

**Proof** :

$$l(Y, f(x)) = Y'\log p(x) + (1-Y')\log(1-p(x))$$
$$\text{where } Y' \in \{0,1\} \text{ and } Y \in \{1,-1\}$$

$$p(x) = p(Y=1|x) = \frac{1}{1+e^{-2f(x)}} = \sigma(2f(x)) \tag{10.17}$$

And we know

$$1 - p(x) = 1 - \frac{1}{1+e^{-2f(x)}} = \frac{1}{1+e^{2f(x)}} = \sigma(-2f(x))$$

$$l(Y, f(x)) = \begin{cases} \log(p(x)) & \text{if } Y' = 1 \\ \log(1 - p(x)) & \text{if } Y' = 0 \end{cases}$$

$$\log p(x) = -\log(1 + e^{-2f(x)}) \qquad \leftarrow \text{Y = 1 case}$$

$$\log(1 - p(x)) = -\log(1 + e^{2f(x)}) \qquad \leftarrow \text{Y = -1 case}$$

If we write both cases in one equation with $Y \in \{1, -1\}$,

$$\therefore \quad -l(Y, f(x)) = \log(1 + e^{-2Yf(x)})$$

## **Eq 10.19:** (ESL p.348)

$$\mathrm{E}(Y|x) = 2p(Y = 1|x) - 1$$

**Proof** :

From PRML Eq (1.89),

$$\begin{aligned} E(Y|x) &= \sum_{Y \in \{1,-1\}} Y p(Y|x) \\ &= 1 \cdot p(Y = 1|x) + (-1) \cdot p(Y = -1|x) \\ &= p(Y = 1|x) - [1 - p(Y = 1|x)] \\ &= 2p(Y = 1|x) - 1 \end{aligned}$$

## **Eq 10.52:** (ESL p.370)

$$f_k(X) = \ln p_k(X) - \frac{1}{K} \sum_{l=1}^{K} \ln p_l(X)$$

**Proof** :

$$p_k(x) = \frac{e^{f_k(x)}}{1 + \sum_{l=1}^{K} e^{f_l(x)}} \tag{10.21}$$

$$\text{with constraint } \sum_{k=1}^{K} f_k(x) = 0.$$

Using Eq (10.21), let's prove the RHS of Eq (10.52) becomes $f_k(X)$.

$$\begin{aligned}
&\ln p_k(X) - \frac{1}{K}\sum_{l=1}^{K} \ln p_l(X) \\
&\quad = f_k(X) - \ln\left(1+\sum_{l=1}^{K} e^{f_l(x)}\right) - \frac{1}{K}\sum_{l=1}^{K}\left\{f_k(X) - \ln\left(1+\sum_{k=1}^{K} e^{f_k(x)}\right)\right\} \\
&\quad = f_k(X) - \ln\left(1+\sum_{l=1}^{K} e^{f_l(x)}\right) - \frac{1}{K}\sum_{l=1}^{K} f_k(X) + \frac{1}{K}\sum_{l=1}^{K}\ln\left(1+\sum_{k=1}^{K} e^{f_k(x)}\right)
\end{aligned}$$

In the above equation, the third term $= 0$ since from the constraint , $\sum_{l=1}^{K} f_k(X) = 0$.

The fourth term's sum over $l$ does not affect inside the bracket.

$$\begin{aligned}
\Rightarrow \quad \ln p_k(X) - \frac{1}{K}\sum_{l=1}^{K} \ln p_l(X) &= f_k(X) - \ln\left(1+\sum_{l=1}^{K} e^{f_l(x)}\right) + \frac{K}{K}\ln\left(1+\sum_{l=1}^{K} e^{f_l(x)}\right) \\
&= f_k(X)
\end{aligned}$$

## Eq 10.54: (ESL p.376)

$$\mathrm{E}(Y|X) = \mathrm{E}(Y|Y>0, X)\cdot p(Y>0|X)$$

**Proof** :

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

$$p(Y,X) = p(Y,X|Y>0)\cdot p(>0) + p(Y,X|Y=0)\cdot P(Y=0) \qquad \leftarrow \text{sum rule}$$

$$\begin{aligned}
\Rightarrow \quad p(Y|X) &= \frac{p(Y,X|Y>0)\cdot p(Y>0)}{p(X)} + \frac{p(Y,X|Y=0)\cdot p(Y=0)}{p(X)} \\
&= p(Y|Y>0,X)\cdot p(Y>0) + p(Y|Y=0,X)\cdot p(Y=0)
\end{aligned}$$

$$
\begin{aligned}
\mathrm{E}(Y|X) &= \int Y p(Y) X) dY \\
&= \int p(Y|Y>0,X)\cdot p(Y>0)dY + \int p(Y|Y=0,X)\cdot p(Y=0)dY \\
&\text{(second term = 0, since Y = 0)} \\
&= \int p(Y|Y>0,X)\cdot p(Y>0)dY \\
&\text{(since p(Y > 0) is a constant,)} \\
&= \int p(Y|Y>0,X)dY \cdot p(Y>0) \\
&= \mathrm{E}(Y|Y>0,X)\cdot p(Y>0)
\end{aligned}
$$

# Chapter 12. Support Vector Machines

## Eq 12.33: (ESL p.433)

$$\beta_\lambda = \frac{1}{\lambda}\sum_{i=1}^{N}\alpha_i y_i x_i$$

**Proof** :

$$\min_{\beta,\beta_0}\frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \tag{12.8}$$

$$\text{subject to } \xi_i \geq 0, \quad y_i(x_i^T\beta+\beta_0)\geq 1-\xi_i \quad \forall_i$$

$$\min_{\beta,\beta_0}\sum_{i=1}^{N}[1-y_i f_i(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2 \tag{12.25}$$

Since the solutions of $\beta$ are identical for both Eqs (12.8) and (12.25), we can write Eq (12.25) in the form of (12.8) with $\lambda$ instead of C. ($\frac{1}{\lambda} \leftrightarrow c$)

$$\min_{\beta,\beta_0}\frac{\lambda}{2}\|\beta\|^2 + \sum_{i=1}^{N}\xi_i \tag{12.8a}$$

Then the new Lagrange function is

$$L'_p = \frac{\lambda}{2}\|\beta\|^2 + \sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1-\xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i \tag{12.9a}$$

$$\frac{\partial L'_p}{\partial \beta} = 0 \quad \rightarrow \quad \beta = \frac{1}{\lambda}\sum_{i=1}^{N}\alpha_i y_i x_i \qquad \text{(This proves Eq (12.33))}$$

$$\frac{\partial L'_p}{\partial \beta_0} = 0 \quad \rightarrow \quad 0 = \sum_{i=1}^{N}\alpha_i y_i$$

$$\frac{\partial L'_p}{\partial \xi_i} = 0 \quad \rightarrow \quad \alpha_i = 1 - \mu_i \quad \forall_i$$

$$\text{and } \alpha_i, \mu_i, \xi_i \geq 0 \quad \forall_i$$

From $\alpha_i = 1 - \mu_i, \quad \alpha_i \leq 1$,

$$\Rightarrow \quad 0 \leq \alpha_i \leq 1.$$

# Chapter 14. Unsupervised Learning

## Eq 14.58: (ESL p.540)

$$\min_{\beta,\mathbf{R}}\|\mathbf{X}_2 - \beta\mathbf{X}_1\mathbf{R}\|_F$$

$$\text{Solutions: } \hat{\mathbf{R}} = \mathbf{U}\mathbf{V}^T, \quad \hat{\beta} = \frac{\mathrm{Tr}(\mathbf{D})}{\|\mathbf{X}_1\|_F^2}$$

**Proof:**

$\beta$ is a positive scalar.

Based on Eq (14.58), Lagrangian will be

$$L(\beta, \mathbf{R}, \mathbf{A}) = \mathrm{Tr}[(\mathbf{X}_2 - \beta\mathbf{X}_1\mathbf{R})^T(\mathbf{X}_2 - \beta\mathbf{X}_1\mathbf{R})] + \mathrm{Tr}[\mathbf{A}(\mathbf{R}^T\mathbf{R} - \mathbf{I})]$$

First term:

$$\begin{aligned}(\mathbf{X}_2 - \beta\mathbf{X}_1\mathbf{R})^T(\mathbf{X}_2 - \beta\mathbf{X}_1\mathbf{R}) &= (\mathbf{X}_2^T - \beta\mathbf{R}^T\mathbf{X}_1^T)(\mathbf{X}_2 - \beta\mathbf{X}_1\mathbf{R}) \\ &= \mathbf{X}_2^T\mathbf{X}_2 - \mathbf{X}_2^T\beta\mathbf{X}_1\mathbf{R} - \beta\mathbf{R}^T\mathbf{X}_1^T\mathbf{X}_2 + \beta^2\mathbf{R}^T\mathbf{X}_1^T\mathbf{X}_1\mathbf{R}\end{aligned}$$

$$\frac{\partial}{\partial\beta}\text{Tr}(\beta\mathbf{X}_2^T\mathbf{X}_1\mathbf{R}) = \text{Tr}(\mathbf{X}_2^T X_1\mathbf{R})$$
$$\frac{\partial}{\partial\beta}\text{Tr}(\beta\mathbf{R}^T\mathbf{X}_1^T\mathbf{X}_2) = \text{Tr}(\mathbf{R}^T\mathbf{X}_1^T\mathbf{X}_2)$$
$$\frac{\partial}{\partial\beta}\text{Tr}(\beta^2\mathbf{R}^T\mathbf{X}_1^T\mathbf{X}_1\mathbf{R}) = 2\beta\text{Tr}(\mathbf{R}^T\mathbf{X}_1^T\mathbf{X}_1\mathbf{R})$$

Using the trace relations; $\text{Tr}(\mathbf{A}^T\mathbf{B}) = \text{Tr}(\mathbf{B}^T\mathbf{A}) = \text{Tr}(\mathbf{A}\mathbf{B}^T) = \text{Tr}(\mathbf{B}\mathbf{A}^T)$,

$$\text{Tr}(\mathbf{X}_2^T\mathbf{X}_1\mathbf{R}) = \text{Tr}(\mathbf{X}_2\mathbf{R}^T\mathbf{X}_1^T) \quad (1)$$
$$\text{Tr}(\mathbf{R}^T\mathbf{X}_1^T\mathbf{X}_2) = \text{Tr}(\mathbf{X}_2^T\mathbf{X}_1\mathbf{R}) = \text{Tr}(\mathbf{X}_2\mathbf{R}^T\mathbf{X}_1^T) \quad (2)$$

To maximize L w.r.t. $\beta$,

$$\Rightarrow \quad \frac{\partial L(\beta,\mathbf{R},A)}{\partial\beta} = -\text{Tr}(\mathbf{X}_2^T\mathbf{X}_1\mathbf{R}) - \text{Tr}(R^T\mathbf{X}_1^T\mathbf{X}_2) + 2\beta\text{Tr}(\mathbf{R}^T\mathbf{X}_1^T\mathbf{X}_1\mathbf{R}) = 0$$

Utilizing Eqs (1) and (2),

$$\begin{aligned}\frac{\partial L(\beta,\mathbf{R},A)}{\partial\beta} &= -2\text{Tr}(\mathbf{X}_2\mathbf{R}^T\mathbf{X}_1^T) + 2\beta\text{Tr}(\mathbf{R}^T\mathbf{X}_1^T\mathbf{X}_2\mathbf{R}) \\ &= -2\text{Tr}(\mathbf{X}_2\mathbf{R}^T\mathbf{X}_1^T) + 2\beta\text{Tr}(\mathbf{X}_1^T\mathbf{X}_2) \\ &= 0\end{aligned}$$

Therefore we have

$$\hat{\beta} = \frac{\text{Tr}(\mathbf{R}^T\mathbf{X}_1^T\mathbf{X}_2)}{\|\mathbf{X}_1\|_F} \quad (3)$$

$\frac{\partial L}{\partial\mathbf{R}} = 0$ will produce the same solution for $\mathbf{R}$ as in Eq (14.57), since the only change is $\hat{\beta}\tilde{\mathbf{X}}_1$ instead of $\tilde{\mathbf{X}}_1$. So $\hat{\mathbf{R}} = \mathbf{U}\mathbf{V}^T$.

Plugging this into Eq (3), (and $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{U}\mathbf{D}\mathbf{V}^T$)

$$\begin{aligned}\hat{\beta} &= \frac{\text{Tr}[\mathbf{V}\mathbf{U}^T \cdot \mathbf{U}\mathbf{D}\mathbf{V}^T]}{\|\mathbf{X}_1\|_F} = \frac{\text{Tr}[\mathbf{V}\mathbf{D}\mathbf{V}^T]}{\|\mathbf{X}_1\|_F} \\ &= \frac{\text{Tr}[\mathbf{V}^T\mathbf{V}\mathbf{D}]}{\|\mathbf{X}_1\|_F} = \frac{\text{Tr}[\mathbf{D}]}{\|\mathbf{X}_1\|_F}\end{aligned}$$

where $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ is used.

# Chapter 17. Undirected Graphical Models

**Eq 17.6:** (ESL p.630)

$$p(Y|Z=z) \sim \mathcal{N}(\mu_Y + (z-\mu_Z)^T \Sigma_{ZZ}^{-1}\sigma_{ZY}, \sigma_{YY} - \sigma_{ZY}^T \mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY})$$

$$\text{where } \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix}$$

**Proof** :

Let's use the method described in §2.3.1 Conditional Gaussian Distributions in PRML. In the book, the derivations are based on the condition on $x_b$.

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}, \qquad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{pmatrix}$$

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab} \\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb} \end{pmatrix}$$

where $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$.

But here, the order is reversed; conditioned on $x_a$. So if we rearrange $\mathbf{\Sigma}$ to match PRML's, we can use the same equations.

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{zz} & \sigma_{zy} \\ \sigma_{zy}^T & \sigma_{yy} \end{pmatrix} \longrightarrow \mathbf{\Sigma}' = \begin{pmatrix} \sigma_{yy} & \sigma_{zy}^T \\ \sigma_{zy} & \mathbf{\Sigma}_{zz} \end{pmatrix}$$

$$\mathbf{\Lambda}' = \begin{pmatrix} \mathbf{\Lambda}_{yy} & \mathbf{\Lambda}_{yz} \\ \mathbf{\Lambda}_{zy} & \mathbf{\Lambda}_{zz} \end{pmatrix}$$

We know that

$$(\mathbf{\Sigma}')^{-1} = \begin{pmatrix} \sigma_{yy} & \sigma_{zy}^T \\ \sigma_{zy} & \mathbf{\Sigma}_{zz} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{\Lambda}_{yy} & \mathbf{\Lambda}_{yz} \\ \mathbf{\Lambda}_{zy} & \mathbf{\Lambda}_{zz} \end{pmatrix}$$

From PRML Eq (2.76),

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}$$

where $\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$.

Using PRML Eq (2.96),

$$p(x_a|x_b) = \mathcal{N}(x_a|\mu_{a|b}, \mathbf{\Lambda}_{aa}^{-1}) \tag{PRML Eq 2.96}$$
$$p(Y|Z) = \mathcal{N}(\mu_{y|z}, \mathbf{\Lambda}_{yy}^{-1})$$

Using PRML Eq (2.97),

$$\begin{aligned}
\mu_{a|b} &= \mu_a - \mathbf{\Lambda}_{aa}^{-1}\mathbf{\Lambda}_{ab}(x_b - \mu_b) && \text{(PRML Eq 2.97)}\\
\mu_{y|z} &= \mu_y - \mathbf{\Lambda}_{yy}^{-1}\mathbf{\Lambda}_{yz}(z - \mu_z) \\
\mathbf{\Lambda}_{yy}^{-1} &= \mathbf{M}^{-1} = (\sigma_{yy} - \sigma_{zy}^T\mathbf{\Sigma}_{zz}^{-1}\sigma_{zy}) && (1)\\
\mathbf{\Lambda}_{yz} &= -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} = -\mathbf{M}\sigma_{zy}^T\mathbf{\Sigma}_{zz}^{-1} \\
\Rightarrow \quad \mu_{y|z} &= \mu_y + \mathbf{M}^{-1}\cdot\mathbf{M}\sigma_{zy}^T\mathbf{\Sigma}_{zz}^{-1}(z - \mu_z) \\
&= \mu_y + \sigma_{zy}^T\mathbf{\Sigma}_{zz}^{-1}(z - \mu_z)
\end{aligned}$$

By taking transpose on the second term,

$$\mu_{y|z} = \mu_y + (z - \mu_z)^T\mathbf{\Sigma}_{zz}^{-1}\sigma_{zy} \tag{2}$$

Eqs (1) and (2) prove Eq (17.6).

## Eq 17.8: (ESL p.631)

$$\theta_{zy} = -\theta_{yy}\cdot\mathbf{\Sigma}_{zz}^{-1}\sigma_{zy}$$
$$\text{where } \frac{1}{\theta_{yy}} = \sigma_{yy} - \sigma_{zy}^T\mathbf{\Sigma}_{zz}^{-1}\sigma_{zy}$$

**Proof** :

Referring to §2.3.1 Conditional Gaussian Distributions in PRML, $\mathbf{\Theta}$ here corresponds to $\mathbf{\Lambda}$ in PRML.

Using

$$\mathbf{\Sigma}' = \begin{pmatrix} \sigma_{yy} & \sigma_{zy}^T \\ \sigma_{zy} & \mathbf{\Sigma}_{zz} \end{pmatrix}^{-1} \quad \text{and} \quad \mathbf{\Theta}' = \begin{pmatrix} \theta_{yy} & \theta_{yz} \\ \theta_{zy} & \theta_{zz} \end{pmatrix}$$

$$\begin{aligned}\theta_{zy} &= -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} \\ &= -\boldsymbol{\Sigma}_{zz}^{-1}\sigma_{zy}\theta_{yy} \\ \text{where } {\theta_{yy}}^{-1} = \mathbf{M}^{-1} &= \sigma_{yy} - \sigma_{zy}^T\boldsymbol{\Sigma}_{zz}^{-1}\sigma_{zy}\end{aligned}$$

## Eq 17.33: (ESL p.639)

$$\frac{\partial \boldsymbol{\Phi}(\boldsymbol{\Theta})}{\partial \theta_{jk}} = \sum_{x \in \chi} x_j x_k \cdot p(x, \boldsymbol{\Theta})$$

**Proof** :

$$\boldsymbol{\Phi}(\boldsymbol{\Theta}) = \log \sum_{x \in \chi} \left[ \exp\left( \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right) \right] \tag{17.29}$$

$$\begin{aligned}\frac{\partial \boldsymbol{\Phi}(\boldsymbol{\Theta})}{\partial \theta_{jk}} &= \frac{\sum_{x \in \chi} \exp\left( \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right) \cdot x_j x_k}{\sum_{x \in \chi} \left[ \exp\left( \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right) \right]} \\ &= \sum_{x \in \chi} x_j x_k \cdot \frac{\exp\left( \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right)}{\sum_{x \in \chi} \left[ \exp\left( \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right) \right]} \qquad (1)\end{aligned}$$

$$\begin{aligned}p(X, \boldsymbol{\Theta}) &= \exp\left[ \sum_{(j,k) \in E} \theta_{jk} x_j x_k - \boldsymbol{\Phi}(\boldsymbol{\Theta}) \right] \qquad (17.28) \\ &= \frac{\exp\left[ \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right]}{\exp[\boldsymbol{\Phi}(\boldsymbol{\Theta})]} \\ &= \frac{\exp\left[ \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right]}{\exp\left\{ \log \sum_{x \in \chi} \left[ \exp\left( \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right) \right] \right\}} \\ &= \frac{\exp\left[ \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right]}{\sum_{x \in \chi} \left[ \exp\left( \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right) \right]} \qquad (2)\end{aligned}$$

Eq (2) is equal to the fraction part of Eq (1).

$$\therefore \quad \frac{\partial \boldsymbol{\Phi}(\boldsymbol{\Theta})}{\partial \theta_{jk}} = \sum_{x \in \chi} x_j x_k \cdot p(x, \boldsymbol{\Theta})$$

# Chapter 18. High-Dimensional Problems

## Eq 18.52: (ESL p.692)

$$p(t_j) \sim \pi_0 \cdot F_0 + (1 - \pi_0)F_1$$

**Proof** :

$$\begin{aligned} p(t_j, z_j) &= p(z_j) \cdot [(t_j|z_j) \\ p(t_j) &= \int p(t_j, z_j) dz_j \\ &= \int p(z_j) \cdot p(t_j|z_j) dz_j \\ &= \pi_0 p(t_j|z_j = 0) + (1 - \pi_0) p(t_j|z_j = 1) \end{aligned}$$

Since $F_0 = p(t_j|z_j = 0)$ and $F_1 = p(t_j|z_j = 1)$,

$$\therefore \ p(t_j) \sim \pi_0 \cdot F_0 + (1 - \pi_0)F_1$$