Gauss-Markov theorem

i) Since both $\hat{\theta}$ and $\tilde{\theta}$ are linear combinations of the elements of y, we can write:

$c^T = a^T (X^T X)^{-1} X^T + d$, where $d \in \mathbb{R}^{1 \times N}$

Then:

$\tilde{\theta} = \hat{\theta} + dy \Rightarrow$

$E[\tilde{\theta}] = E[\hat{\theta}] + E[dy] \Rightarrow$ $\tilde{\theta}, \hat{\theta}$ are unbiased

$a^T w = a^T w + dE[Xw + \varepsilon] \Rightarrow$ $E[\varepsilon] = 0$

$dXw = 0 \Rightarrow$ We want this to hold for all w, since w is unobservable.

$dX = 0$

For the variance of $\tilde{\theta}$ we have:

$$\begin{aligned}
Var(\tilde{\theta}) &= Var(c^T y) \\
&= c^T \cdot Cov(y) \cdot c \\
&= c^T (\sigma^2 I_{N \times N}) c \\
&= \sigma^2 c^T c \\
&= \sigma^2 [a^T (X^T X)^{-1} X^T + d][X (X^T X)^{-1} a + d^T] \\
&= \sigma^2 [a^T (X^T X)^{-1} X^T X (X^T X)^{-1} a + dX (X^T X)^{-1} a \\
&\quad + a^T (X^T X)^{-1} X^T d^T + dd^T] \\
&= \sigma^2 [a^T (X (X^T X)^{-1})^T (X (X^T X)^{-1}) a + dd^T]
\end{aligned}$$

$I_{N \times N}$ is the $N \times N$ identity matrix.

$dX = 0$

$$= a^T (X(X^TX)^{-1})^T \sigma^2 I_{N\times N} (X(X^TX)^{-1}) a + \sigma^2 \|d\|^2$$
$$= a^T Cov[(X(X^TX)^{-1})^T y] a + \sigma^2 \|d\|^2$$
$$= a^T Cov[\hat{w}] a + \sigma^2 \|d\|^2$$
$$= Var[\hat{\theta}] + \sigma^2 \|d\|^2$$
$$\geq Var[\hat{\theta}].$$

ii) Since both $\hat{w}$ and $\tilde{w}$ are linear transformations of y, we can write:

$C = (X^TX)^{-1}X^T + D$, where $D \in \mathbb{R}^{d\times N}$

Then:

$\tilde{w} = \hat{w} + Dy$

$E[\tilde{w}] = E[\hat{w}] + E[Dy] \Rightarrow$ $\hat{w}, \tilde{w}$ are unbiased

$w = w + D\,E[Xw + \varepsilon] \Rightarrow$ $E[\varepsilon] = 0$

$DXw = 0 \Rightarrow$ We want this to hold for all w, since w is unobservable.

$DX = 0$

For the covariance of $\tilde{w}$, we have:

$$Cov(\tilde{w}) = Cov(Cy)$$
$$= C \cdot Cov(y) \cdot C^T$$
$$= C \cdot (\sigma^2 I_{N\times N}) \cdot C^T$$

$$= \sigma^2 C C^T$$

$$= \sigma^2 \left[(X^T X)^{-1} X^T + D \right] \left[X (X^T X)^{-1} + D^T \right]$$

$$= \sigma^2 \left[(X^T X)^{-1} X^T X (X^T X)^{-1} + (X^T X)^{-1} X^T D^T + D X (X^T X)^{-1} + D D^T \right]$$

$DX = 0$

$$= \sigma^2 \left[(X^T X)^{-1} X^T \left((X^T X)^{-1} X^T \right)^T + D D^T \right]$$

$$= (X^T X)^{-1} X^T \cdot (\sigma^2 I_{N \times N}) \cdot \left((X^T X)^{-1} X^T \right)^T + \sigma^2 D D^T$$

$$= \mathrm{Cov}\left[(X^T X)^{-1} X^T y \right] + \sigma^2 D D^T$$

$$= \mathrm{Cov}(\hat{w}) + \sigma^2 D D^T$$

Therefore, we get:

$\tilde{\Sigma} = \hat{\Sigma} + \sigma^2 D D^T \Rightarrow$

$\tilde{\Sigma} - \hat{\Sigma} = D D^T$

Since DD^T is always a positive semidefinite matrix, we have $\tilde{\Sigma} \geq \hat{\Sigma}$.

Ridge regression

Since we have a Gaussian sampling model (likelihood), the Gaussian prior serves as a conjugate prior. In other words, when the prior and the likelihood are Gaussian, the posterior is also Gaussian. In a Gaussian distribution, the mean and the mode match. Therefore, it suffices to show that the ridge regression estimate w_{RR} is the point where the density of the posterior $P(w|y)$ is maximized. By Bayes theorem, we have:

$$p(w|y) = \frac{p(y|w) \cdot p(w)}{p(y)} \Rightarrow$$

$$\log p(w|y) = \log p(y|w) + \log p(w) - \log p(y) \Rightarrow$$

$$\log p(w|y) = \log \frac{1}{\sqrt{(2\pi)^N |\sigma^2 I|}} - \frac{1}{2}(y - Xw)^T (\sigma^2 I)^{-1} (y - Xw)$$

$$+ \log \frac{1}{\sqrt{(2\pi)^d |\tau I|}} - \frac{1}{2} w^T (\tau I)^{-1} w - \log p(y) \Rightarrow$$

c is a constant (not a function of w)

$$\log p(w|y) = -\frac{1}{2}\left[\frac{(y - Xw)^T (y - Xw)}{\sigma^2} + \frac{w^T w}{\tau}\right] + c \Rightarrow$$

$$\sigma^2 \cdot \log p(w|y) = -\frac{1}{2}\left[(y - Xw)^T (y - Xw) + \frac{\sigma^2}{\tau} w^T w\right] + c \cdot \sigma^2$$

Therefore, finding the maximum of the density with respect to w, is equivalent to minimizing the expression:

$$(y-Xw)^T(y-Xw) + \frac{\sigma^2}{\tau} w^T w$$

On the other hand, for the ridge regression problem, if each x_{ij} gets replaced by $x_{ij} - \bar{x}_j$ and w_0 is set equal to $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, the minimization problem for the remaining d parameters can be written in matrix form as:

$$\min_w (y-Xw)^T(y-Xw) + \lambda w^T w.$$

It is easy to see that the two problems are equivalent for $\lambda = \frac{\sigma^2}{\tau}$.

Sigmoid: the beginning

i) $$\sigma(-a) = \frac{1}{1+e^{a}} = \frac{e^{-a}}{e^{-a}+1} = \frac{1+e^{-a}-1}{1+e^{-a}} = 1 - \frac{1}{1+e^{-a}} = 1-\sigma(a)$$

ii) We set $y = \sigma(a) > 0$ and solve for a.

$$y = \frac{1}{1+e^{-a}} \iff$$

$$1+e^{-a} = \frac{1}{y} \iff$$

$$e^{-a} = \frac{1}{y} - 1 \iff$$

$$-a = \ln \frac{1-y}{y} \iff$$

$$a = \ln \frac{y}{1-y}.$$

Therefore, $\sigma^{-1}(y) = \ln \frac{y}{1-y}$.

Sigmoid: the posterior

We have $a = \ln \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$

$$= \ln \frac{p(x|C_1)}{p(x|C_2)} + \ln \frac{p(C_1)}{p(C_2)}$$

$$= \ln p(x|C_1) - \ln p(x|C_2) + \ln \frac{p(C_1)}{p(C_2)}.$$

In the above, we use $p(x|C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}$

and we get:

$$a = -\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) + \ln \frac{p(C_1)}{p(C_2)}$$

$$= -\frac{1}{2}\left[x^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu_1 - \mu_1^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1} \mu_1\right]$$

$$+ \frac{1}{2}\left[x^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu_2 - \mu_2^T \Sigma^{-1} x + \mu_2^T \Sigma^{-1} \mu_2\right] + \ln \frac{p(C_1)}{p(C_2)}$$

Here we use the symmetry of the dot product.

$$= \frac{1}{2}\left[\mu_1^T (\Sigma^{-1})^T x + \mu_1^T \Sigma^{-1} x - \mu_2^T (\Sigma^{-1})^T x - \mu_2^T \Sigma^{-1} x\right]$$

$$- \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}$$

Note that the covariance matrix Σ (and therefore its iverse Σ^{-1}) is symmetric.

As a result, we get:

$$a = \mu_1^T (\Sigma^{-1})^T x - \mu_2^T (\Sigma^{-1})^T x - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}$$

$$= w^T x + b,$$

where $w = \Sigma^{-1}(\mu_1 - \mu_2)$

and $b = -\frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}$

One-of-K

The likelihood function is:

$$p(\{\varphi_n, y_n\} \mid \pi_1, \pi_2, \dots, \pi_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[p(\varphi_n \mid C_k) \cdot \pi_k \right]^{y_{nk}}$$

Equivalently, we can maximize the log-likelihood:

$$\log p(\{\varphi_n, y_n\} \mid \pi_1, \pi_2, \dots, \pi_K) = \sum_{n=1}^{N} \sum_{k=1}^{K} y_{nk} \left[\log p(\varphi_n \mid C_k) + \log \pi_k \right]$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{N} y_{nk} \cdot \log p(\varphi_n \mid C_k) + \sum_{n=1}^{N} \sum_{k=1}^{K} y_{nk} \cdot \log \pi_k$$

Since the first term does not depend on $\pi_1, \pi_2, \dots, \pi_K$, we only have to maximize the second term but taking into account the constraint $\sum_{k=1}^{K} \pi_k = 1$.

We do that by adding a Lagrange multiplier and finding the saddle point of the expression:

$$L(\pi_1, \pi_2, \dots, \pi_K, \lambda) = \sum_{n=1}^{N} \sum_{k=1}^{K} y_{nk} \cdot \log \pi_k + \lambda \cdot \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

Setting the partial derivatives to 0, we get:

$$\frac{\partial L}{\partial \pi_k} = 0 \text{ for } k = 1, \dots, K \Leftrightarrow$$

$$\sum_{n=1}^{N} y_{nk} \cdot \frac{1}{\pi_k} + \lambda = 0 \text{ for } k = 1, \dots, K \Leftrightarrow$$

$$\pi_k = -\frac{N_k}{\lambda} \text{ for } k = 1, \dots, K$$

For the partial derivative with respect to λ, we have:

$$\frac{\partial L}{\partial \lambda} = 0 \iff \sum_{k=1}^{K} \pi_k - 1 = 0$$

By the system of $K+1$ equations, we get:

$$\sum_{k=1}^{K} \left(-\frac{N_k}{\lambda}\right) - 1 = 0 \iff -\frac{N}{\lambda} = 1 \iff \lambda = -N$$

Therefore, for $k = 1, 2, \dots, K$, we have:

$$\pi_k = \frac{N_k}{N}.$$

Sigmoid: derivative and loss

i) We have:

$$\frac{d\sigma}{da} = -\frac{1}{(1+e^{-a})^2} \cdot (-e^{-a})$$

$$= \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}+1-1}{1+e^{-a}}$$

$$= \sigma(a) \cdot \left(1 - \frac{1}{1+e^{-a}}\right)$$

$$= \sigma(a) \cdot (1 - \sigma(a))$$

ii) We have:

$$\nabla_w L(\phi, y, w) = -\sum_{n=1}^{N} y_n \nabla_w \ln \sigma(w^T \varphi_n) + (1-y_n) \nabla_w \ln(1-\sigma(w^T\varphi_n))$$

$$= -\sum_{n=1}^{N} y_n \frac{1}{\sigma(w^T\varphi_n)} \cdot \sigma(w^T\varphi_n) \cdot (1-\sigma(w^T\varphi_n)) \cdot \varphi_n$$

$$- \sum_{n=1}^{N} (1-y_n) \frac{1}{1-\sigma(w^T\varphi_n)} \cdot (-\sigma(w^T\varphi_n)) \cdot (1-\sigma(w^T\varphi_n)) \cdot \varphi_n$$

$$= -\sum_{n=1}^{N} y_n (1-\sigma(w^T\varphi_n)) \cdot \varphi_n$$

$$+ \sum_{n=1}^{N} \sigma(w^T\varphi_n) \cdot \varphi_n - \sum_{n=1}^{N} y_n \cdot \sigma(w^T\varphi_n) \cdot \varphi_n$$

$$= \sum_{n=1}^{N} (\sigma(w^T\varphi_n) - y_n) \cdot \varphi_n$$

Linearly separable

Assume we have N data points $\{y_n, \varphi(x_n)\}_{n=1}^{N}$, where the first N_1 belong to C_1 and the remaining $N-N_1$ belong to class C_2. The likelihood for a logistic regression model is given by:

$$p(\{y_n, \varphi(x_n)\}) = \prod_{n=1}^{N_1} P(C_1 | \varphi(x_n))\, p(\varphi(x_n)) \cdot \prod_{n=N_1+1}^{N} \left[1 - P(C_1 | \varphi(x_n))\right] p(\varphi(x_n))$$

$$= \prod_{n=1}^{N_1} \frac{1}{1+e^{-w^T\varphi(x_n)}}\, p(\varphi(x_n)) \cdot \prod_{n=N_1+1}^{N} \frac{e^{-w^T\varphi(x_n)}}{1+e^{-w^T\varphi(x_n)}}\, p(\varphi(x_n))$$

We can see that $p(\varphi(x_n))$ is independent of w and the fractions take values in $[0,1]$. If we choose a w such that $w^T\varphi(x_n) > 0 \; (*) \; \forall n \in \{1, \ldots, N_1\}$ and $\|w\| \to \infty$, we can easily see that the first N_1 fractions converge to 1. Similarly, if we choose a w such that $w^T\varphi(x_n) < 0 \; (**)$ $\forall n \in \{N_1, \ldots, N\}$ and $\|w\| \to \infty$, the last $N-N_1$ fractions converge to 1. Since the dataset is linearly separable, a w simultaneously satisfying $(*)$, $(**)$ is guaranteed to exist.

Linear Discriminant Analysis

i) Let $X \in \mathbb{R}^{N \times (d+1)}$ be an augmented matrix containing the samples x_i as rows, where the first dimension (column) is fixed to 1: $X = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_N^T \end{bmatrix}$

Since $\begin{bmatrix} w_0 \\ w \end{bmatrix}$ is the least squares estimate, we have:

$$\begin{bmatrix} w_0 \\ w \end{bmatrix} = (X^T X)^{-1} X^T y \Leftrightarrow$$

$$X^T X \begin{bmatrix} w_0 \\ w \end{bmatrix} = X^T y \Leftrightarrow$$

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{bmatrix} \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_N^T \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \Leftrightarrow$$

$$\begin{bmatrix} N & \sum_{i=1}^{N} x_i^T \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i x_i^T \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} y_i \cdot x_i \end{bmatrix} \Leftrightarrow$$

$$N \cdot w_0 + \left(\sum_{i=1}^{N} x_i^T \right) \cdot w = \sum_{i=1}^{N} y_i \quad (1)$$

and

$$\left(\sum_{i=1}^{N} x_i \right) w_0 + \left(\sum_{i=1}^{N} x_i x_i^T \right) w = \sum_{i=1}^{N} y_i x_i \quad (2)$$

For the targets of the two classes coded as $-\frac{N}{N_1}, \frac{N}{N_2}$, we can rewrite (1) as:

$$N \cdot w_0 + \left[N_1 \mu_1^T + N_2 \mu_2^T\right] w = N_1 \cdot \left(-\frac{N}{N_1}\right) + N_2 \cdot \left(\frac{N}{N_2}\right) \iff$$

$$N \cdot w_0 + \left[N_1 \mu_1^T + N_2 \mu_2^T\right] w = 0 \quad (3)$$

Additionally, we have:

$$\Sigma = \frac{1}{N-2}\left[\sum_{i=1}^{N_1} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{i=N_1+1}^{N_2} (x_i - \mu_2)(x_i - \mu_2)^T\right] \iff$$

$$\Sigma = \frac{1}{N-2}\left[\sum_{i=1}^{N_1}\left[x_i x_i^T - x_i \mu_1^T - \mu_1 x_i^T + \mu_1 \mu_1^T\right] + \sum_{i=N_1+1}^{N_2}\left[x_i x_i^T - x_i \mu_2^T - \mu_2 x_i^T + \mu_2 \mu_2^T\right]\right] \iff$$

$$\Sigma = \frac{1}{N-2}\left[\sum_{i=1}^{N} x_i x_i^T - N_1 \mu_1 \mu_1^T - N_1 \mu_1 \mu_1^T + N_1 \mu_1 \mu_1^T - N_2 \mu_2 \mu_2^T - N_2 \mu_2 \mu_2^T + N_2 \mu_2 \mu_2^T\right] \iff$$

$$\Sigma = \frac{1}{N-2} \sum_{i=1}^{N} x_i x_i^T - \frac{N_1}{N-2} \mu_1 \mu_1^T - \frac{N_2}{N-2} \mu_2 \mu_2^T \iff$$

$$\sum_{i=1}^{N} x_i x_i^T = (N-2)\Sigma + N_1 \mu_1 \mu_1^T + N_2 \mu_2 \mu_2^T \quad (4)$$

Also, we have:

$$\sum_{i=1}^{N} y_i x_i = \sum_{i=1}^{N_1} y_i x_i + \sum_{i=N_1+1}^{N_2} y_i x_i = \left(-\frac{N}{N_1}\right) N_1 \mu_1 + \left(\frac{N}{N_2}\right) N_2 \mu_2 = N(\mu_2 - \mu_1) \quad (5)$$

By combining (2), (3), (4) and (5), we get:

$$-\frac{1}{N}(N_1\mu_1+N_2\mu_2)(N_1\mu_1^T+N_2\mu_2^T)w+\left[(N-2)\Sigma+N_1\mu_1\mu_1^T+N_2\mu_2\mu_2^T\right]w$$

$$=N(\mu_2-\mu_1) \iff$$

$$\left[\left(N_1-\frac{N_1^2}{N}\right)\mu_1\mu_1^T-\frac{N_1N_2}{N}\mu_2\mu_1^T-\frac{N_1N_2}{N}\mu_1\mu_2^T+\left(N_2-\frac{N_2^2}{N}\right)\mu_2\mu_2^T+(N-2)\Sigma\right]w$$

$$=N(\mu_2-\mu_1) \iff$$ Here we use $N=N_1+N_2$

$$\frac{N_1N_2}{N}\left[\mu_1\mu_1^T-\mu_2\mu_1^T-\mu_1\mu_2^T+\mu_2\mu_2^T\right]w+(N-2)\Sigma w=N(\mu_2-\mu_1) \iff$$

$$\frac{N_1N_2}{N}(\mu_2-\mu_1)(\mu_2-\mu_1)^Tw+(N-2)\Sigma w=N(\mu_2-\mu_1)$$

This gives us:

$$\left[(N-2)\Sigma+N\Sigma_B\right]w=N(\mu_2-\mu_1)$$

where $\Sigma_B=\frac{N_1N_2}{N^2}(\mu_2-\mu_1)(\mu_2-\mu_1)^T$.

ii) We have that $\Sigma_B w = \frac{N_1 N_2}{N^2}(\mu_2-\mu_1)(\mu_2-\mu_1)^T w$, where $(\mu_2-\mu_1)^T w$ is a scalar. Therefore, $\Sigma_B \cdot w$ is clearly in the direction of $\mu_2-\mu_1$.

By writing $\Sigma_B w = \lambda(\mu_2-\mu_1)$ in the result of (i), we get:

$$(N-2)\Sigma w + N\lambda(\mu_2-\mu_1) = N(\mu_2-\mu_1) \Leftrightarrow$$

$$(N-2)\Sigma w = N(1-\lambda)(\mu_2-\mu_1) \Leftrightarrow$$

$$w = \frac{N(1-\lambda)}{N-2}\Sigma^{-1}(\mu_2-\mu_1) \text{ where } \frac{N(1-\lambda)}{N-2} \text{ is a scalar.}$$

iii) Let the samples of classes C_1, C_2 have arbitrary targets t_1, t_2. Then, equation (3) becomes:

$$N \cdot w_0 + [N_1 \mu_1^T + N_2 \mu_2^T] w = N_1 \cdot t_1 + N_2 \cdot t_2 \quad (3')$$

Similarly, equation (5) becomes:

$$\sum_{i=1}^{N} y_i x_i = t_1 N_1 \mu_1 + t_2 N_2 \mu_2 \quad (5')$$

By combining (2), (3'), (4), (5'), we get:

$$-\frac{1}{N}(N_1\mu_1 + N_2\mu_2)(N_1\mu_1^T + N_2\mu_2^T)w + \frac{1}{N}(N_1\mu_1 + N_2\mu_2)(N_1 t_1 + N_2 t_2)$$

$$+ [(N-2)\Sigma + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T] w = t_1 N_1 \mu_1 + t_2 N_2 \mu_2$$

$$\left[(N-2)\Sigma + N\Sigma_B\right]w = N_1 \cdot t_1 \cdot \mu_1 + N_2 \cdot t_2 \cdot \mu_2 - \frac{1}{N} N_1^2 t_1 \mu_1 - \frac{N_1 N_2}{N} t_2 \mu_1 - \frac{N_1 N_2}{N} t_1 \mu_2 - \frac{N_2^2}{N} t_2 \mu_2 \Leftrightarrow$$

$$\left[(N-2)\Sigma + N\Sigma_B\right]w = \frac{N_1 N_2}{N}\left[t_1 \mu_1 + t_2 \mu_2 - t_2 \mu_1 - t_1 \mu_2\right] \Leftrightarrow$$

$$\left[(N-2)\Sigma + N\Sigma_B\right]w = \frac{N_1 N_2}{N}(t_2 - t_1)(\mu_2 - \mu_1)$$

Therefore, we get a similar result to question (i), as long as $t_1 \neq t_2$.