Statistical learning homework 2

Dai Yuehao (1800010660@pku.edu.cn)

November 2, 2020

1 ESL 3.16

For best subset selection we select M coefficients into our model, thus we have

$$\hat{y'} = \sum_{i=1}^{p} \hat{\beta}_i I_i x_i,$$

here $I_i = 1$ if $\hat{\beta}_i$ is chosen into the model, $I_i = 0$ if not. Now the residual is

$$r = ||y - \hat{y'}||_2^2 = ||y||_2^2 - \sum_{i=1}^{p} \hat{\beta}_i^2 I_i^2 ||x_i||_2^2 = ||y||_2^2 - \sum_{i=1}^{p} \hat{\beta}_i^2 I_i,$$

hence the best selection of coefficients should be the M largest ones, hence the fomula is

$$\hat{\beta}'_j = \hat{\beta}_j \cdot I[\text{rank}(|\hat{\beta}_j| \leq M)].$$

For ridge regression we have the close form soltion

$$\hat{\boldsymbol{\beta}}' = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}y = (1+\lambda)\boldsymbol{X}^T y,$$

since

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y = \boldsymbol{X}^T y$$

we have the fomula

$$\hat{\beta}'_j = \frac{\hat{\beta}_j}{1+\lambda}.$$

For LASSO we have the corresbonding optimizatioin problem

$$\min f(\boldsymbol{\beta}) = \frac{1}{2}||y - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{i=1}^{p} \beta_i,$$

this is a convex problem hence the optimality condition is

$$\boldsymbol{X}^T y - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T y - \boldsymbol{\beta} \in \partial\lambda||\boldsymbol{\beta}||_1, \tag{1.0.1}$$

since $\boldsymbol{X}^T y = \hat{\boldsymbol{\beta}}$ and if $\hat{\beta}_j > \lambda$ then $\partial|\hat{\beta}_j| = 1$ and we should shrink $\hat{\beta}_j$ to $\hat{\beta}_j - \lambda$ in order to satisfy (1.0.1). Similar operation when $\hat{\beta}_j < -\lambda$. If $0 \leq \hat{\beta}_j \leq \lambda$ we can only shrink it to 0 and the same operation when $-\lambda \leq \hat{\beta}_j \leq 0$. Thereby we obtain the fomula

$$\hat{\beta}'_j = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+.$$

2 ESL 3.30

Let

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda\alpha}\boldsymbol{I} \end{pmatrix}, \quad \tilde{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}$$

then we have

$$||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}||_2^2 = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda\alpha||\boldsymbol{\beta}||_2^2,$$

hence the elastic-net optimizatioin problem can be turned into a lasso problem writen as

$$\min_{\boldsymbol{\beta}} \quad ||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}||_2^2 + \lambda(1-\alpha)||\boldsymbol{\beta}||_1.$$

3 ADMM for group LASSO

Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \cdots), \boldsymbol{\gamma}_G^T)^T$, then we have the equivelant optimization problem

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \quad \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda\sum_{g=1}^{G}||\boldsymbol{\gamma}_g||_2 \quad \text{s.t.} \quad \boldsymbol{\beta}_g = \boldsymbol{\gamma}_g, \quad g = 1, \cdots, G$$

then the augmented Lagrange function is

$$L_\rho(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda\sum_{g=1}^{G}||\boldsymbol{\gamma}_g||_2 + \boldsymbol{\alpha}^T(\boldsymbol{\beta} - \boldsymbol{\gamma}) + \frac{\rho}{2}||\boldsymbol{\beta} - \boldsymbol{\gamma}||_2^2,$$

then we can renew the coefficients

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \left(\boldsymbol{X}^T\boldsymbol{X} + \rho\boldsymbol{I}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{y} + \rho\boldsymbol{\gamma}^{(k)} - \boldsymbol{\alpha}^{(k)}\right),$$

then for given $\boldsymbol{\gamma}_g$ it is a convex optimization problem

$$\min_{\boldsymbol{\gamma}_g} \quad \lambda||\boldsymbol{\gamma}_g||_2 - \boldsymbol{\alpha}^T\boldsymbol{\gamma}_g + \frac{\rho}{2}||\boldsymbol{\beta}_g - \boldsymbol{\gamma}_g||_2^2,$$

the optimality condition is

$$\rho(\boldsymbol{\gamma}_g - \boldsymbol{\beta}_g) - \boldsymbol{\alpha} \in -\lambda\partial||\boldsymbol{\gamma}_g||_2,$$

then in terms of this we renew $\boldsymbol{\gamma}_g^{(k)}$ from $g = 1$ to $g = G$, at last we renew

$$\boldsymbol{\alpha}^{(k+1)} \leftarrow \boldsymbol{\alpha}^{(k)} + \rho\left(\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\gamma}^{(k+1)}\right).$$

4 Normal bound

By the assumption we have

$$P\{Z \geq t\}e^{t^2/(2\sigma^2)} = \int_t^\infty \frac{1}{\sqrt{2\pi}\sigma}e^{-(x^2-t^2)/(2\sigma^2)}dx = \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}e^{-xt/\sigma^2}dx,$$

hence

$$P\{Z \geq t\}e^{t^2/(2\sigma^2)} \leq \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}dx = \frac{1}{2}.$$

On the other hand for all $\varepsilon > 0$ there exists $N > 0$ such that

$$\int_0^N \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} dx > \frac{\sqrt{1-\varepsilon}}{2}$$

holds and there exists $t_0 > 0$ such that for all $t < t_0$ and $x \le N$

$$e^{-xt/\sigma^2} > \sqrt{1-\varepsilon}$$

holds, hence

$$\int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} e^{-xt/\sigma^2} dx > \int_0^N \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} e^{-xt/\sigma^2} dx > \frac{1-\varepsilon}{2},$$

thereby we have

$$\sup_{t>0} \left\{ P\{Z \ge t\} e^{t^2/(2\sigma^2)} \right\} = \frac{1}{2}.$$

5 ESL 4.2

5.1 a

For class $k = 1, 2$ and by the Gaussian assumption we can calculate the Bayes' discriminant function

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_k|) + \ln(\pi_k),$$

by the assumption we let

$$\Sigma_1 = \Sigma_2 = \hat{\mathbf{\Sigma}}, \quad \mu_k = \hat{\mu_k}, \quad \pi_k = \frac{N_k}{N}, \quad k = 1, 2,$$

thus the LDA rule classifies x to class 2 if $\delta_2(x) - \delta_1(x) > 0$. Note that on can write

$$\delta_k(x) = -\frac{1}{2}\left(x^T \hat{\mathbf{\Sigma}}^{-1} x - x^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu_k} - \hat{\mu_k}^T \hat{\mathbf{\Sigma}}^{-1} x + \hat{\mu_k}^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu_k} \right) + \log\left(\frac{N_k}{N}\right),$$

hence

$$\delta_2(x) - \delta_1(x) = x^T \hat{\mathbf{\Sigma}}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) - \frac{1}{2}\left(\hat{\mu}_2^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_2 - \hat{\mu}_1^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_1 \right) - \log\left(\frac{N_1}{N}\right) + \log\left(\frac{N_2}{N}\right)$$

which is exactly our result.

5.2 b

Assume that $x_1, \cdots, x_{N_1}$ are in class 1 the rest are in class 2, then we directly calculate that

$$X^T X = \begin{pmatrix} N & \sum_{i=1}^N x_i^T \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i x_i^T \end{pmatrix}, \quad X^T y = \begin{pmatrix} 0 \\ -N\hat{\mu}_1 + N\hat{\mu}_2 \end{pmatrix},$$

also we have

$$\hat{\mathbf{\Sigma}} = \frac{1}{N-2}\left[\sum_{i=1}^n x_i x_i^T - N_1 \hat{\mu}_1 \hat{\mu}_1^T - N_2 \hat{\mu}_2 \hat{\mu}_2^T \right],$$

thus

$$\sum_{i=1}^{n} x_i x_i^T = (N-2)\hat{\mathbf{\Sigma}} + N_1\hat{\mu}_1\hat{\mu}_1^T + N_2\hat{\mu}_2\hat{\mu}_2^T.$$

Since

$$X^T X \begin{pmatrix}\beta_0\\ \beta\end{pmatrix} = \begin{pmatrix}0\\ -N\hat{\mu}_1 + N\hat{\mu}_2\end{pmatrix}$$

we have

$$N\beta_0 + \sum_{i=1}^{N} x_i^T\beta = 0 \quad \Rightarrow \quad \beta_0 = \left(-\frac{N_1}{N}\hat{\mu}_1^T - \frac{N_2}{N}\hat{\mu}_2^T\right)\beta, \tag{5.2.1}$$

since $\sum_{i=1}^N x_i = N_1\hat{\mu}_1 + N_2\hat{\mu}_2$ we have

$$\sum_{i=1}^{N}\beta_0 x_i + \sum_{i=1}^{N} x_i x_i^T\beta = -N\hat{\mu}_1 + N\hat{\mu}_2$$

$$\Rightarrow \left[(N-2)\hat{\mathbf{\Sigma}} + \frac{N_1N_2}{N}(\hat{\mu}_1-\hat{\mu}_2)^T(\hat{\mu}_1-\hat{\mu}_2)\hat{\mathbf{\Sigma}}_B\right]\beta = N(\hat{\mu}_1-\hat{\mu}_2),$$

thus we obtain

$$\left[(N-2)\hat{\mathbf{\Sigma}} + \frac{N_1N_2}{N}\hat{\mathbf{\Sigma}}_B\right]\beta = N(\hat{\mu}_1-\hat{\mu}_2).$$

5.3 c

Note that

$$\hat{\mathbf{\Sigma}}_B\beta = (\hat{\mu}_1-\hat{\mu}_2)(\hat{\mu}_1-\hat{\mu}_2)^T\beta \propto \hat{\mu}_1-\hat{\mu}_2,$$

thus

$$\beta \propto \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1-\hat{\mu}_2).$$

5.4 d

Assume that we place the label y_i for class i, then we also calculate that

$$X^T y = \begin{pmatrix}N_1y_1 + N_2y_2\\ y_1N_1\hat{\mu}_1 + y_2N_2\hat{\mu}_2\end{pmatrix},$$

then we have

$$N\beta_0 + \sum_{i=1}^{N} x_i^T\beta = N_1y_1 + N_2y_2,$$

hence

$$\beta_0 = \frac{N_1}{N}y_1 + \frac{N_2}{N}y_2 - \left(\frac{N_1}{N}\hat{\mu}_1^T + \frac{N_2}{N}\hat{\mu}_2^T\right)\beta,$$

since

$$\sum_{i=1}^{N}\beta_0 x_i + \sum_{i=1}^{N} x_i x_i^T\beta = y_1N_1\hat{\mu}_1 + y_2N_2\hat{\mu}_2$$

we obtain

$$\left[(N-2)\hat{\mathbf{\Sigma}} + \frac{N_1N_2}{N}\hat{\mathbf{\Sigma}}_B\right]\beta = y_1N_1\hat{\mu}_1 + y_2N_2\hat{\mu}_2 - (N_1\hat{\mu}_1 + N_2\hat{\mu}_2)\left(\frac{N_1}{N}y_1 + \frac{N_2}{N}y_2\right),$$

then we see that

$$y_1N_1\hat{\mu}_1 + y_2N_2\hat{\mu}_2 - (N_1\hat{\mu}_1 + N_2\hat{\mu}_2)\left(\frac{N_1}{N}y_1 + \frac{N_2}{N}y_2\right) = \frac{N_1N_2}{N}(y_1 - y_2)(\hat{\mu}_1 - \hat{\mu}_2)$$

then we conclude our proof.

5.5 e

From (5.2.1) we have

$$\hat{\beta}_0 = -\frac{1}{N}\sum_{i=1}^{N} x_i^T\beta.$$

Next we have

$$\hat{f} = \left(x - \frac{1}{N}\sum_{i=1}^{N} x_i^T\right)^T \beta \propto x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) - \frac{1}{N}\sum_{i=1}^{N} x_i^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2),$$

when $N_1 = N_2$ the right hand side of above can be written as

$$x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$$

which from ESL 4.5(a) we know it is the dicision boudary of LDA. When $N_1 \neq N_2$ they are obviously different.

6 ESL 4.3

We can write the simplified discriminant function of x as

$$\delta_k^{(x)}(x) = x^T\hat{\mathbf{\Sigma}}_x^{-1}\hat{\mu}_k^{(x)} - \frac{1}{2}\hat{\mu}_k^{(x)T}\hat{\mathbf{\Sigma}}_x^{-1}\hat{\mu}_k^{(x)} + \log\pi_k,$$

and the discriminant function of y as

$$\delta_k^{(y)}(y) = y^T\hat{\mathbf{\Sigma}}_y^{-1}\hat{\mu}_k^{(y)} - \frac{1}{2}\hat{\mu}_k^{(y)T}\hat{\mathbf{\Sigma}}_y^{-1}\hat{\mu}_k^{(y)} + \log\pi_k,$$

since $y = \hat{\boldsymbol{B}}^T x$ and $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{B}}$ we have

$$\hat{\mu_k}^{(y)} = \hat{\boldsymbol{B}}^T\hat{\mu_k}^{(x)}, \quad \hat{\mathbf{\Sigma}}_y = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{g_i=k}(y_i - \hat{\mu}_i^{(y)})(y_i - \hat{\mu}_i^{(y)})^T = \hat{\boldsymbol{B}}^T\hat{\mathbf{\Sigma}}_x\hat{\boldsymbol{B}}, \tag{6.0.1}$$

our goal is to verify that $\delta_k^{(x)}(x) - \delta_l^{(x)}(x) = 0$ if and only if $\delta_k^{(y)}(x\hat{\boldsymbol{B}}^T) - \delta_l^{(y)}(x\hat{\boldsymbol{B}}^T) = 0$, if $K \leq p$ we know that $\hat{\mathbf{\Sigma}}_y$ is nonsingular, and from (6.0.1) we need to verify

$$\hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^T\Sigma_x\hat{\boldsymbol{B}})^{-1}\hat{\boldsymbol{B}}^T(\hat{\mu}_k^{(x)} - \hat{\mu}_l^{(x)}) = \hat{\mathbf{\Sigma}}_x^{-1}(\hat{\mu}_k^{(x)} - \hat{\mu}_l^{(x)}).$$

Since $\boldsymbol{Y}$ is an indicator response matrix we have

$$\hat{\mu}_k^{(x)} = \frac{1}{N_k}\sum_{g_i=k} x_i = \frac{1}{N_k}\boldsymbol{X}^T y_k, \tag{6.0.2}$$

thus

$$\hat{\boldsymbol{\Sigma}}_x = \frac{1}{N-K}\left[\sum_{i=1}^{N} x_i x^T - \sum_{k=1}^{K} N_k \hat{\mu}_k^{(x)} \hat{\mu}_k^{(x)T}\right] = \frac{1}{N-K}(\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{P})$$

where

$$\boldsymbol{P} = \sum_{k=1}^{K} \frac{1}{N_k} \boldsymbol{X}^T y_i y_i^T \boldsymbol{X},$$

hence

$$\hat{\boldsymbol{\Sigma}}_y = \frac{1}{N-K}\boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{P})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y},$$

let $\boldsymbol{H} = \boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ then

$$(\hat{\boldsymbol{B}}^T\hat{\boldsymbol{\Sigma}}_x\hat{\boldsymbol{B}})^{-1} = (N-K)(\boldsymbol{I} - \boldsymbol{H}_1)^{-1}\boldsymbol{H}^{-1}$$

where $\boldsymbol{H}_1 = \boldsymbol{H}^{-1}\boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{P}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, hence from (6.0.2) if we have

$$\begin{aligned}
&\hat{\boldsymbol{\Sigma}}_x\hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^T\hat{\boldsymbol{\Sigma}}_x\hat{\boldsymbol{B}})^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{X}^T\boldsymbol{Y}\\
=&(\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{P}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H}_1)^{-1}\boldsymbol{H}^{-1}\boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\\
=&\boldsymbol{X}^T\boldsymbol{Y},
\end{aligned} \tag{6.0.3}$$

then

$$\hat{\boldsymbol{\Sigma}}_x\hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^T\hat{\boldsymbol{\Sigma}}_x\hat{\boldsymbol{B}})^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{X}^T y_k = \hat{\boldsymbol{\Sigma}}_x\hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^T\hat{\boldsymbol{\Sigma}}_x\hat{\boldsymbol{B}})^{-1}\hat{\boldsymbol{B}}^T N_k\hat{\mu}_k^x = \boldsymbol{X}^T y_k = N_k\hat{\mu}_k^x$$

and we finish the proof. Equation (6.0.3) holds if and only if

$$\boldsymbol{P}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{H}_1, \tag{6.0.4}$$

a sufficient condition is $N_{k_1} = N_{k_2}$ for all k_1, k_2, under this condition we have

$$\boldsymbol{P} = \frac{1}{N_1}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}$$

and (6.0.4) holds. General cases have not been solved yet.

If $K > p$ then $\hat{\boldsymbol{\Sigma}}_y$ is singular and it seems that LDA with y can not be done. However if we consider that $\hat{\boldsymbol{Y}}$ is restricted in a p dimensional subspace and we do the LDA in that subspace, then we augment $\boldsymbol{X}$ to

$$\tilde{\boldsymbol{X}} = \begin{pmatrix}\boldsymbol{X} & \boldsymbol{0}\end{pmatrix}$$

where $\tilde{\boldsymbol{X}}$ is a $N \times K$ matrix then

$$\tilde{\boldsymbol{\Sigma}}_x = \begin{pmatrix}\hat{\boldsymbol{\Sigma}}_x & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}$$

and we can also augment $\hat{\boldsymbol{B}}$ to a $K \times K$ nonsingular matrix, say

$$\tilde{\boldsymbol{H}} = \begin{pmatrix}\hat{\boldsymbol{B}}\\ \boldsymbol{0}\end{pmatrix}$$

that still satisfies $y = \tilde{x}\tilde{\boldsymbol{B}}$ and $\hat{\boldsymbol{\Sigma}}_y = \tilde{\boldsymbol{B}}^T\tilde{\boldsymbol{\Sigma}}_x\tilde{\boldsymbol{B}}$, then

$$\hat{\boldsymbol{\Sigma}}_y^{\dagger} = \tilde{\boldsymbol{B}}^{-1}\begin{pmatrix}\hat{\boldsymbol{\Sigma}}_x^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}(\tilde{\boldsymbol{B}}^T)^{-1},$$

then we have

$$\tilde{\boldsymbol{B}}\hat{\boldsymbol{\Sigma}}_y^{\dagger}\tilde{\boldsymbol{B}}^T = \tilde{\boldsymbol{\Sigma}}_x^{\dagger},$$

then we conclude the proof.

7 ESL 4.5

Assume that we classify x_i to class 0 if $x_i < x_0$, the log-likelihood function is

$$l(\beta_0,\beta_1)=\sum_{i=1}^{N}\left[(\beta_0+\beta_1x_i)1_{\{x_i>x_0\}}-\log(1+e^{\beta_0+\beta_1x_i})\right],$$

then the MLE of $\hat{\beta}_0$ and $\hat{\beta}_1$ satisfy

$$\frac{\partial l}{\partial\hat{\beta}_0}=\sum_{i=1}^{N}1_{\{x_i>x_0\}}-\sum_{i=1}^{N}\frac{e^{\beta_0+\beta_1x_i}}{1+e^{\beta_0+\beta_1x_i}}=0$$

and

$$\frac{\partial l}{\partial\hat{\beta}_1}=\sum_{i=1}^{N}x_i1_{\{x_i>x_0\}}-\sum_{i=1}^{N}x_i\frac{e^{\beta_0+\beta_1x_i}}{1+e^{\beta_0+\beta_1x_i}}=0,$$

moreover if consider that x_0 separating two classes means that $p(x_0;\beta_0,\beta_1)=1/2$ then we have $\beta_0+x_0\beta_1=0$, then we substitude this into the one order equation and solve β.

Actually we can see the problem in another way, we want to maxmize the probability of generating class 1 when $x>x_0$, note that this pobability is monotone decreasing for β_1 with respect to all $x_i>x_0$, thus the value of MLE of β_0,β_1 can be reached when $\beta\to-\infty$. In other words, the bigger $|\beta_1|$ is the better when $\beta_0+x_0\beta_1=0$.

7.1 a

Assume that two classes can be separated by a hyperplane H_0, then for all $x\in H_0$ there should be $\beta_0+x^T\boldsymbol{\beta}_1=0$ hence H_0 is exactly $\beta_0+x^T\boldsymbol{\beta}_1=0$, then we want to maxmize the probability of generating class 1 when $\beta_0+x^T\boldsymbol{\beta}_1<0$ which is

$$p(x;\beta_0,\boldsymbol{\beta}_1)=\frac{1}{e^{\beta_0+x^T\boldsymbol{\beta}_1}},$$

this is monotone increasing for $||\boldsymbol{\beta}||$ if it maintains that $\beta_0+x^T\boldsymbol{\beta}_1<0$, or in other words β_0 and $\boldsymbol{\beta}_1$ are mutiplied by the same constant. Hence the situation is similar with that in $\mathbb{R}$.

7.2 b

Assume that M classes are separated by the points $-\infty=x_0<x_1<\cdots<x_{M-1}<x_M=+\infty$ and if $x_{m-1}<x<x_m$ we classify x to class $m-1$. Similarly we should have

$$\beta_{m,0}+x_m\beta_{m,1}=\beta_{m+1,0}+x_m\beta_{m+1,1},\quad m=1,\cdots,M-2, \tag{7.2.1}$$

and

$$\beta_{M-1,0}+x_{M-1}\beta_{M-1,1}=0, \tag{7.2.2}$$

also we want to maxmize the probability of generating class M when $x>x_{M-1}$ note that this pobability is monotone decreasing for $\beta_{M,1}$ with respect to all $x_i>x_{M-1}$, then from equations (7.2.1) and euqation (7.2.2) we can obtain all β. Note that if we multiply all β by a constant bigger than 1 then the probability to generate all the sample points will increase, hence the same as the situation with 2 classes, the bigger all β the better when equations (7.2.1) and (7.2.2) hold.

8 ESL 4.6

8.1 a

By assumption if there is separability there exists a β such that $\beta^T x_i^* > 0$ if $y_i = 1$ and $\beta^T x_i^* < 0$ if $y_i = -1$ which can be written as $y_i \beta^T x_i^* > 0$, thus $y_i \beta^T z_i > 0$, hence there is a certain t such that $y_i \beta^T z_i \geq t$ since the number of samples is finite, then let

$$\beta_{\text{sep}} = \frac{1}{m}\beta$$

that we conclude the proof.

8.2 b

We can calculate that

$$||\beta_{\text{new}} - \beta_{\text{sep}}||^2 = ||\beta_{\text{old}} - \beta_{\text{sep}} + y_i z_i||^2 = ||\beta_{\text{old}} - \beta_{\text{sep}}||^2 + y_i^2 ||z_i||^2 + 2y_i(\beta_{\text{old}} - \beta_{\text{sep}})^T z_i,$$

hence we only need to verify that

$$y_i(\beta_{\text{old}} - \beta_{\text{sep}})^T z_i \leq -1$$

since $y_i^2 ||z_i||^2 = 1$. This is quite trivial because $y_i \beta_{\text{sep}}^T z_i = 1$ and y_i is misclassified thus $y_i \beta_{\text{old}}^T z_i \leq 0$. Thereby we conclude our proof.