Statistical learning homework 4

Dai Yuehao (1800010660@pku.edu.cn)

December 16, 2020

1 ESL 7.2

Let $p = P\{\hat{G}(x_0) = G(x_0)|x_0\}$ we have

$$\begin{aligned}
\mathrm{Err}(x_0) &= P\{Y \neq \hat{G}(x_0)|X = x_0\} \\
&= P\{Y \neq G(x_0)|X = x_0\} \cdot p + P\{Y = G(x_0)|X = x_0\} \cdot (1-p) \\
&= P\{Y \neq G(x_0)|X = x_0\} - P\{Y \neq G(x_0)|X = x_0\} \cdot (1-p) + P\{Y = G(x_0)|X = x_0\} \cdot (1-p) \\
&= \mathrm{Err}_{\mathrm{B}}(x_0) + [P\{Y = G(x_0)|X = x_0\} - P\{Y \neq G(x_0)|X = x_0\}] \cdot (1-p) \\
&= \mathrm{Err}_{\mathrm{B}}(x_0) + |2f(x_0) - 1| P\{\hat{G}(x_0) \neq G(x_0)|X = x_0\}.
\end{aligned}$$

Now let $\sigma^2 = \mathrm{Var}[f(x_0)]$ and $\mu = \mathbb{E}[\hat{f}(x_0)]$ we have

$$\begin{aligned}
P\{\hat{G}(x_0) \neq G(x_0), X = x_0\} &= P\{\hat{f}(x_0) < 0.5, G(x_0) = 1\} + P\{\hat{f}(x_0) > 0.5|G(x_0) = 0\} \\
&\approx \Phi\left(\frac{0.5-\mu}{\sigma}\right) P\{G(x_0) = 1\} + \left[1 - \Phi\left(\frac{0.5-\mu}{\sigma}\right)\right] P\{G(x_0) = 0\} \\
&= \Phi\left(\frac{0.5-\mu}{\sigma}\right) P\{G(x_0) = 1\} + \Phi\left(\frac{\mu-0.5}{\sigma}\right) P\{G(x_0) = 0\} \\
&= \Phi\left(\frac{\mathrm{sign}(0.5 - f(x_0))(\mu - 0.5)}{\sigma}\right),
\end{aligned}$$

hence we conclude the proof.

2 ESL 7.6

We have

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i = \boldsymbol{S} y,$$

here $\boldsymbol{S}$ is a binary matrix with values $1/k$ or 0 such that the sum of each row sums is 1. Moreover, the diagonal entries are all $1/k$, since x_i should be the neighbour of x_i's, hence

$$\mathrm{df}(\boldsymbol{S}) = \mathrm{tr}(\boldsymbol{S}) = \frac{N}{k}.$$

3 ESL 7.7

Let $x = \mathrm{tr}(S)/N$ we can calculate that

$$\begin{aligned}
\mathrm{GCV} &= \frac{1}{N}\sum_i^N \left[\frac{y_i - \hat{y}_i}{1 - \mathrm{tr}(S)/N}\right]^2 \approx \frac{1}{N}\sum_i^N (y_i - \hat{y}_i)^2 \left[1 + 2\frac{\mathrm{tr}(S)}{N}\right] \\
&= \frac{1}{N}\sum_i^N (y_i - \hat{y}_i)^2 + 2\frac{d}{N^2}\sum_i^N (y_i - \hat{y}_i)^2 \\
&\approx \overline{\mathrm{err}} + 2\frac{d}{N}\sigma_\epsilon^2 \\
&= C_p.
\end{aligned}$$

4 ESL 10.2

The goal is to find

$$f^*(x) = \arg\min_{f(x)} \mathbb{E}_{Y|x}\left[e^{-Yf(x)}\right],$$

taking the derivative of $f(x)$ and setting it to zero we have

$$\frac{\partial}{\partial f} E_{Y|x}\left[e^{-Yf(x)}\right] = E_{Y|x}\left[-Ye^{-Yf(x)}\right] = 0,$$

thus

$$e^{f(x)}P\{Y = -1|x\} - e^{-f(x)}P\{Y = 1|x\} = 0,$$

hence

$$f(x) = \frac{1}{2}\log\left(\frac{P\{Y = 1|x\}}{P\{Y = -1|x\}}\right).$$

5 ESL 10.5

5.1 a

The optimization problem can be written as

$$\begin{aligned}
\min_f \quad & \mathbb{E}\left[\exp\left(-\frac{1}{K}Y^T f\right)\right] \\
\text{s.t.} \quad & \mathbf{1}^T f = 0,
\end{aligned}$$

the Lagrange function is

$$\mathcal{L}(f, \nu) = \mathbb{E}\left[\exp\left(-\frac{1}{K}Y^T f\right)\right] + \nu\left(\mathbf{1}^T f\right),$$

then

$$\frac{\partial \mathcal{L}}{\partial f} = -\frac{1}{K}\mathbb{E}\left[Y\exp\left(-\frac{1}{K}Y^T f\right)\right] + \nu\mathbf{1}^T = 0,$$

on the other hand we have

$$\mathbb{E}\left[Y\exp\left(-\frac{1}{K}Y^T f\right)\right] = \sum_{k=1}^K P\{G = \mathcal{G}_k\}Y_k \exp\left(\frac{1}{K(K-1)}\sum_{i=1}^K f_i - \frac{1}{K-1}f_k\right),$$

hence the entries of the above are all equal, which means

$$P\{G = \mathcal{G}_k\}\exp\left(-\frac{1}{K-1}f_k\right) = P\{G = \mathcal{G}_t\}\exp\left(-\frac{1}{K-1}f_t\right), \quad \forall k, t.$$

now we fix f_1 and then

$$\log P\{G=\mathcal{G}_k\} - \frac{1}{K-1}f_k = \log P\{G=\mathcal{G}_1\} - \frac{1}{K-1}f_1,$$

hence

$$f_k = f_1 + (K-1)\log\frac{P\{G=\mathcal{G}_k\}}{P\{G=\mathcal{G}_1\}},$$

since $1^T f = 0$ we have

$$Kf_1 + (K-1)\sum_{k=2}^{K}\log\frac{P\{G=\mathcal{G}_k\}}{P\{G=\mathcal{G}_1\}} = 0$$

which yields that

$$f_1 = \frac{K-1}{K}\left[(K-1)\log P\{G=\mathcal{G}_1\} - \sum_{k=2}^{K}\log P\{G=\mathcal{G}_k\}\right] = (K-1)\log P\{G=\mathcal{G}_1\} - \frac{K-1}{K}\sum_{k=1}^{K}\log P\{G=\mathcal{G}_k\},$$

hence

$$f_k = (K-1)\log P\{G=\mathcal{G}_k\} - \frac{K-1}{K}\sum_{k=1}^{K}\log P\{G=\mathcal{G}_k\}.$$

Finally we get

$$f_k^* = (K-1)\log P\{G=\mathcal{G}_k\} - \frac{K-1}{K}\sum_{k=1}^{K}\log P\{G=\mathcal{G}_k\}.$$

5.2 b

Now the optimization problem is

$$(\beta_m, G_m) = \arg\min_{\beta,G}\sum_i L[y_i, f_{m-1}(x_i) + \beta G(x_i)]$$

which can be written as

$$(\beta_m, G_m) = \arg\min_{\beta,G}\sum_i w_i^{(m)}\exp\left[-\frac{\beta}{K}y_i^T G(x_i)\right]$$

here $w_i^{(m)} = \exp\left[-y_i^T f_{m-1}(x_i)\right]$. Now we take the derivative of β and set it to zero we have

$$\sum_i w_i^{(m)} y_i^T G(x_i)\exp\left[-\frac{\beta}{K}y_i^T G(x_i)\right] = 0,$$

since when $G(x_i) = y_i$ we have

$$y_i^T G(x_i) = \frac{K-1}{(K-1)^2} + 1 = \frac{K}{K-1},$$

when $G(x_i) \neq y_i$ we have

$$y_i^T G(x_i) = \frac{K-2}{(K-1)^2} - \frac{2}{K-1} = \frac{-K}{(K-1)^2},$$

hence

$$\sum_{y_i=G(x_i)} w_i^{(m)} y_i^T G(x_i)\exp\left[-\frac{1}{K-1}\beta\right] + \sum_{y_i\neq G(x_i)} w_i^{(m)} y_i^T G(x_i)\exp\left[\frac{1}{(K-1)^2}\beta\right] = 0$$

which yields

$$\exp\left[\frac{K}{(K-1)^2}\beta\right] = (K-1)\frac{\sum_{y_i=G(x_i)} w_i^{(m)}}{\sum_{y_i\neq G(x_i)} w_i^{(m)}},$$

hence

$$\beta_m = \frac{(k-1)^2}{K}\left[\log\left(\frac{1-\mathrm{err}_m}{\mathrm{err}_m}\right) + \log(K-1)\right].$$

6 ESL 10.8

6.1 a

First we have

$$p_k(x) = \frac{e^{f_k(x)}}{\sum_j e^{f_j(x)}},$$

the log-likelihood can be written as

$$\begin{aligned} l &= \sum_{x_i \in R} \sum_{k=1}^{K} y_{ik} \log p_k(x_i) = \sum_{x_i \in R} \sum_{k=1}^{K} y_{ik} \left[\log e^{f_k(x_i)+\gamma_k} - \log \left(\sum_j e^{f_j(x_i)+\gamma_j} \right) \right] \\ &= \sum_{x_i \in R} \sum_{k=1}^{K} y_{ik} \left[f_k(x_i) + \gamma_k \right] - \sum_{x_i \in R} \sum_{k=1}^{K} y_{i,k} \log \left(\sum_j e^{f_j(x_i)+\gamma_j} \right) \\ &= \sum_{x_i \in R} \sum_{k=1}^{K} y_{ik} \left[f_k(x_i) + \gamma_k \right] - \sum_{x_i \in R} \log \left(\sum_j e^{f_j(x_i)+\gamma_j} \right). \end{aligned}$$

The first derivative is

$$\frac{\partial l}{\partial \gamma_k} = \sum_{x_i \in R} y_{i,k} - \sum_{x_i \in R} \frac{e^{f_k(x_i)+\gamma_k}}{\sum_j e^{f_j(x_i)+\gamma_j}}.$$

The second derivatives are

$$\frac{\partial^2 l}{\partial \gamma_k^2} = - \sum_{x_i \in R} \frac{e^{f_k(x_i)+\gamma_k} \sum_j e^{f_j(x_i)+\gamma_j} - e^{2f_k(x_i)+2\gamma_k}}{\left(\sum_j e^{f_j(x_i)+\gamma_j} \right)^2}$$

and

$$\frac{\partial^2 l}{\partial \gamma_k \gamma_m} = \sum_{x_i \in R} \frac{e^{f_k(x_i)+\gamma_k} e^{f_m(x_i)+\gamma_m}}{\left(\sum_j e^{f_j(x_i)+\gamma_j} \right)^2}.$$

6.2 b

The Newton update can be written as

$$\gamma_{new} = \left(\frac{\partial^2 l}{\partial \gamma^2} \right)^{-1} \frac{\partial l}{\partial \gamma},$$

since we only use the diagonal of the Hessian matrix, we have

$$\gamma_k^1 = \frac{\partial l}{\partial \gamma_k} \cdot \left(\frac{\partial^2 l}{\partial \gamma_k^2} \right)^{-1} \Bigg|_{\gamma_k = 0} = \frac{\sum_{x_i \in R} (y_{i,k} - p_k(x_i))}{\sum_{x_i \in R} p_k(x_i)(1 - p_k(x_i))}$$

6.3 c

We can verify that

$$\sum_{k=1}^{K} \hat{\gamma}_k = \frac{K-1}{K} \sum_{k=1}^{K} \gamma_k^1 - K \cdot \frac{K-1}{K} \sum_{k=1}^{K} \frac{1}{K} \gamma_k^1 = 0.$$

7 ESL 12.1

Consider (12.8), since $C > 0$, the optimal ξ_i should satisfy

$$\xi_i[y_i(x_i^T\beta + \beta_0) - 1 + \xi_i] = 0,$$

hence

$$\xi_i = 1 - y_i(x_i^T\beta + \beta_0) \quad \text{or} \quad \xi_i = 0$$

must hold, hence $\xi_i = [1 - y_i(x_i^T\beta + \beta_0)]_+$, hence (12.8) can be written as

$$\min_{\beta,\beta_0} \quad \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}[1 - y_i(x_i^T\beta + \beta_0)]_+,$$

with $\lambda = 1/C$ the optimization problem is also equivalent to

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

thereby we conclude the proof.

8 ESL 12.2

If we choose K such that

$$\langle h(x_j), h(x_i)\rangle = K(x_j, x_i),$$

then

$$f(x_j) = h(x_j)^T\beta + \beta_0 = \sum_{i=1}^{N}\alpha_i y_i \langle h(x_j), h(x_i)\rangle + \beta_0 = \sum_{i=1}^{N}\alpha_i y_j K(x_j, x_i) + \beta_0$$

for some α_i, where

$$\beta = \sum_{i=1}^{N}\alpha_i y_i \langle \cdot, h(x_i)\rangle,$$

let $\alpha_i' = \alpha_i y_i$ hence

$$\beta^T\beta = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \langle h(x_j), h(x_i)\rangle = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i'\alpha_j' K(x_j, x_i) = (\alpha')^T K(\alpha'),$$

hence

$$h(x_j)^T\beta = \sum_{i=1}^{N}\alpha_i' K(x_j, x_i), \quad \beta^T\beta = (\alpha')^T K(\alpha'),$$

hence the solution to (12.29) is the same as the solution to (12.25) for a particular kernel $K(x_j, x_i) = \langle h(x_j), h(x_i)\rangle$, we have concluded the proof.