Statistical Learning HW 4

Johnson Lin

November 18, 2019

1 Problem 1

ESL Ex. 7.2

Assume that the true probability of observing $Y = 1$ is larger than $\frac{1}{2}$, so we have $f(x_0) > \frac{1}{2}$. In this case, the optimal decision is $G(x_0) = 1$. The Bayes error is then the probability Y is not one. So we have

$$Err_B(x_0) = Pr(Y \neq 1) = Pr(Y \neq G(x_0) = Pr(Y = 0) = 1 - f(x_0)$$

The full error is the probability that Y is not the same label as the one we assign it.

$$\begin{aligned}
Err(x_0) &= Pr(Y \neq \hat{G}(x_o)) \\
&= Pr(Y = 1)Pr(\hat{G}(x_0) = 0) + Pr(Y = 0)Pr(\hat{G}(x_0) = 1) \\
&= f(x_0)Pr(\hat{G}(x_0) = 0) + (1 - f(x_0)(1 - Pr(\hat{G}(x_0) = 0)) \\
&= 1 - f(x_0) + (2f(x_0) - 1)Pr(\hat{G}(x_0) = 0) \\
&= Err_B(x_0) + (2f(x_0) - 1)Pr(\hat{G}(x_0) = 0)
\end{aligned}$$

So we have

$$Pr(\hat{G}(x_0) = 0) = Pr(\hat{G}(x_0) =\neq 1) = Pr(\hat{G}(x_0) \neq G(x_0))$$

Now let's consider the other case where $f(x_0) < \frac{1}{2}$ and $G(x_0) = 0$. Our Bayes error in this case is

$$Err_B(x_0) = Pr(Y \neq 0) = Pr(Y \neq G(x_0)) = Pr(Y = 1) = f(x_0)$$

Our total error is then

$$Err(x_0) = 1 - f(x_0) + (2f(x_0) - 1)Pr(\hat{G}(x_0) = 0)$$

$$= Err_B(x_0) + (1 - 2f(x_0) + (2f(x_0) - 1)Pr(\hat{G}(x_0) = 0)$$

Here we have

$$Pr(\hat{G}(x_0) = 0) = Pr(\hat{G}(x_0) = G(x_0)) = 1 - Pr(\hat{G}(x_0) \neq G(x_0)$$

So we can write the expression of $Err(x_0)$ as

$$Err(x_0) = Err_B(x_0) + (1 - 2f(x_0) + (2f(x_0) - 1)(1 - Pr(\hat{G}(x_0) \neq G(x_0))$$
$$= Err_B(x_0) - (2f(x_0) - 1)Pr(\hat{G}(x_0) \neq G(x_0))$$
$$= Err_B(x_0) + |2f(x_0) - 1|Pr(\hat{G}(x_0) \neq G(x_0))$$

which is what we were trying to show. The next thing we have to show is an approximation of $Pr(\hat{G}(x_0) \neq G(x_0))$. Let's start with the first case where $f(x_0) < \frac{1}{2}$. So we have

$$Pr(\hat{G}(x_0) \neq G(x_0)) = Pr(\hat{G}(x_0) = 1) = Pr(\hat{f}(x_0) > \frac{1}{2})$$
$$= Pr(\frac{\hat{f}(x_0) - E\hat{f}(x_0)}{\sqrt{Var(\hat{f}(x_0))}} > \frac{\frac{1}{2} - E\hat{f}(x_0)}{\sqrt{Var(\hat{f}(x_0))}})$$
$$= 1 - \Phi(\frac{\frac{1}{2} - E\hat{f}(x_0)}{\sqrt{Var(\hat{f}(x_0))}})$$
$$= \Phi(\frac{E\hat{f}(x_0) - \frac{1}{2}}{\sqrt{Var(\hat{f}(x_0))}})$$

Now we look at the second case where $f(x_0) > \frac{1}{2}$. Here we have

$$Pr(\hat{G}(x_0) \neq G(x_0)) = Pr(\hat{G}(x_0) = 0) = Pr(\hat{f}(x_0) < \frac{1}{2})$$
$$= Pr(\frac{\hat{f}(x_0) - E\hat{f}(x_0)}{\sqrt{Var(\hat{f}(x_0))}} < \frac{\frac{1}{2} - E\hat{f}(x_0)}{\sqrt{Var(\hat{f}(x_0))}})$$
$$= \Phi(\frac{\frac{1}{2} - E\hat{f}(x_0)}{\sqrt{Var(\hat{f}(x_0))}})$$

Similarly, we can combine these two cases to get the expression:

$$Pr(\hat{G}(x_0) \neq G(x_0)) = \Phi(\frac{sign(\frac{1}{2} - f(x_0))(E\hat{f}(x_0) - \frac{1}{2})}{\sqrt{Var(\hat{f}(x_0))}})$$

which is what we were trying to show.

2 Problem 2

ESL Ex. 7.6

k-nearest-neighbor regression fit can be expressed as a linear smoother where $\hat{y} = Sy$. Then we have

$$S_{ij} = \begin{cases} \frac{1}{k} & \text{if} X_j \in N_k(X_i) \\ 0 & \text{otherwise} \end{cases}$$

where S_{ij} is the element of S in row i and column j and (X, y) is our training set. $N_k(X_i)$ is the set of k nearest neighbors of X_i. Note that $S_{ii} = \frac{1}{k}$ for all $i = 1, \ldots, N$ since a data point will always lie in its own set of k nearest neighbors. The effective degrees of freedom is equal to $trace(S)$ so we have

$$\begin{aligned} df(S) &= trace(S) \\ &= \sum_{i=1}^{N} S_{ii} \\ &= \sum_{i=1}^{N} \frac{1}{k} \\ &= \frac{N}{k} \end{aligned}$$

3 Problem 3

ESL Ex. 7.7

$$GCV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} (\frac{y_i - \hat{f}(x_i)}{1 - trace(S)/N})^2$$

If we use the approximation $\frac{1}{(1-x)^2} \approx 1 + 2x$ then we have

$$\begin{aligned} GCV(\hat{f}) &\approx \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2 (1 + \frac{2trace(S)}{N}) \\ &= \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2 + \frac{2}{N^2} trace(S) \sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2 \end{aligned}$$

The first term is the in-sample training error. The second term we can approximate

$$\hat{\sigma}_\varepsilon^2 \approx \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2$$

which turns the equation into

$$GCV(\hat{f}) = e\bar{r}r + \frac{2trace(S)}{N}\hat{\sigma}_{\varepsilon}^2$$

We know that $trace(S)$ is the effective number of parameters in the model. We can call it d and this gives us the same expression as C_p in Equation 7.26:

$$GCV(\hat{f}) = e\bar{r}r + \frac{2d}{N}\hat{\sigma}_{\varepsilon}^2$$

$$C_p = e\bar{r}r + 2 \cdot \frac{d}{N}\hat{\sigma}_{\varepsilon}^2$$

4 Problem 4

ESL Ex. 10.2

We need to prove Equation 10.16 which is:

$$f^*(x) = argmin_{f(x)} E_{Y|x}(e^{-Yf(x)}) = \frac{1}{2} log \frac{Pr(Y=1|x)}{Pr(Y=-1|x)}$$

In order to find a $f(x)$ that fulfills the first equivalence, we would take the derivative and set it equal to 0, solving for $f(x)$. This gives us

$$E_{Y|x}(-Ye^{-Yf(x)}) = 0$$

If we evaluate this when our targets are $Y = \pm 1$,

$$-(-1)e^{-(-1)f(x)}Pr(Y=-1|x) - 1(1)e^{-1f(x)}Pr(Y=1|x) = 0$$

$$e^{2f(x)}Pr(Y=-1|x) - Pr(Y=1|x) = 0$$

$$e^{2f(x)} = \frac{Pr(Y=1|x)}{Pr(Y=-1|x)}$$

$$f(x) = \frac{1}{2} log \frac{Pr(Y=1|x)}{Pr(Y=-1|x)}$$

which is what we wanted to show.

5 Problem 5

ESL Ex. 10.5

(a.) We are looking for $f(x)$ which fulfills:

$$f^*(x) = argmin_{f(x)} E_{Y|x}[exp(-\frac{1}{K}Y^T f)]$$

where $\sum_{k=1}^{K} f(x) = 0$. Let $\mathcal{L}(f;\lambda)$ be the Lagrangian defined as

$$\mathcal{L}(f;\lambda) \equiv E_{Y|x}[exp(-\frac{1}{K}Y^T f)] - \lambda(\sum_{k=1}^{K} f_k - 0)$$

We will start by evaluating the expectation

$$\varepsilon \equiv E_{Y|x}[exp(-\frac{1}{K}Y^T f)]$$

$$\varepsilon \equiv E_{Y|x}[exp(-\frac{1}{K}(Y_1 f_1 + \cdots + Y_K f_K))]$$

We will evaluate this using LOTUS $E[f(x)] = \sum f(x_i)p(x_i)$ and given the encoding for vector Y:

$$Y_k = \begin{cases} 1 & k = c \\ -\frac{1}{K-1} & k \neq c \end{cases}$$

Now we have

$$\varepsilon \equiv exp(-\frac{1}{K}(-\frac{1}{K-1}f_1(x) + f_2(x)\cdots + f_K(x)))\text{Prob}(c=1|x)$$

$$+exp(-\frac{1}{K}(f_1(x) - \frac{1}{K-1}f_2(x)\cdots + f_K(x)))\text{Prob}(c=2|x)$$

$$\vdots$$

$$+exp(-\frac{1}{K}(f_1(x) + f_2(x)\cdots - \frac{1}{K-1}f_K(x)))\text{Prob}(c=K|x)$$

The exponential argument in each term can be written as

$$-\frac{1}{K-1} = \frac{K-1-K}{K-1} = 1 - \frac{1}{K-1}$$

So now we have:

$$\varepsilon \equiv exp(-\frac{1}{K}(f_1(x) + f_2(x)\cdots + f_K(x) - \frac{K}{K-1}f_1(x)))\text{Prob}(c=1|x)$$

$$+exp(-\frac{1}{K}(f_1(x) + f_2(x)\cdots + f_K(x) - \frac{K}{K-1}f_2(x)))\text{Prob}(c=2|x)$$

$$\vdots$$

$$+exp(-\frac{1}{K}(f_1(x) + f_2(x)\cdots + f_K(x) - \frac{K}{K-1}f_K(x)))\text{Prob}(c=K|x)$$

Under the constraint $\sum_{k'=1}^{K} f_{k'}(x) = 0$, we now have

$$\varepsilon \equiv exp(-\frac{1}{K}(f_1(x)))\text{Prob}(c=1|x) + exp(-\frac{1}{K}(f_2(x)))\text{Prob}(c=2|x) + \cdots + exp(-\frac{1}{K}(f_K(x)))\text{Prob}(c=K|x)$$

In order to find $f(x)$ we need to take the derivative of the Lagrangian objective function $\mathcal{L}$ and set that equal to 0. We need to add $-\lambda \sum_{k=1}^{K} f_k$ to ε.

$$\frac{\partial \mathcal{L}}{\partial f_k} = \frac{\partial \varepsilon}{\partial f_k} - \lambda = \frac{1}{K-1} exp(\frac{1}{K-1} f_k(x)) \text{Prob}(c = k|x) - \lambda$$

for $1 \leq k \leq K$. The derivative with respect to λ would give back the constraint $\sum_{k'=1}^{K} f_{k'}(x) = 0$. To solve this, we solve the above equation in terms of λ and plug it back into the constraint. We then have

$$f_k(x) = -(K-1) log(\frac{-(K-1)\lambda}{\text{Prob}(c = k|x)})$$

$$f_k(x) = -(K-1) log(\text{Prob}(c = k|x)) - (K-1) log(-(K-1)\lambda).$$

Requiring this expression to sum to 0 means that the following must be true

$$(K-1) \sum_{k'=1}^{K} log(\text{Prob}(c = k'|x)) - K(K-1) log(-(K-1)\lambda) = 0$$

$$sum_{k'=1}^{K} log(\text{Prob}(c = k'|x)) - K log(-(K-1)\lambda) = 0$$

$$sum_{k'=1}^{K} log(\text{Prob}(c = k'|x)) = K log(-(K-1)\lambda)$$

$$log(-(K-1)\lambda) = \frac{1}{K} sum_{k'=1}^{K} log(\text{Prob}(c = k'|x))$$

$$\lambda = -\frac{1}{K-1} exp(\frac{1}{K} sum_{k'=1}^{K} log(\text{Prob}(c = k'|x)))$$

When we plug this back in we get

$$f_k(x) = (K-1) log(\text{Prob}(c = k|x)) - \frac{K-1}{K} \sum_{k'=1}^{K} log(\text{Prob}(c = k'|x))$$

$$= (K-1)(log(\text{Prob}(c = k|x)) - \frac{1}{K} \sum_{k'=1}^{K} log(\text{Prob}(c = k'|x)))$$

for $1 \leq k \leq K$. We can think of this as K equations for the K unknowns $\text{Prob}(c = k|x)$. To find these probabilities, we write the above as

$$\frac{1}{K-1} f_k(x) = log(\text{Prob}(c = k|x)) + log([\prod_{k'=1}^{K} \text{Prob}(c = k'|x)]^{-1/K})$$

$$\text{Prob}(c = k|x) = [\prod_{k'=1}^{K} \text{Prob}(c = k'|x)]^{1/K} e^{\frac{f_k(x)}{K-1}}$$

If we sum both sides from $k' = 1$ to $k' = K$, we have

$$1 = [\prod_{k'=1}^{K} \text{Prob}(c = k'|x)]^{1/K} \sum_{k'=1}^{K} e^{\frac{f_{k'}(x)}{K-1}}$$

$$[\prod_{k'=1}^{K} \text{Prob}(c = k'|x)]^{1/K} \sum_{k'=1}^{K} = \frac{1}{\sum_{k'=1}^{K} e^{\frac{f_{k'}(x)}{K-1}}}$$

$$\text{Prob}(c = k'|x) = \frac{e^{\frac{f_k(x)}{K-1}}}{\sum_{k'=1}^{K} e^{\frac{f_{k'}(x)}{K-1}}}$$

which is what we wanted to show.

(b.) The AdaBoost algorithm is given by Algorithm 10.1. A multiclass boosting algorithm (SAMME) using this loss function would follow the same steps for the most part. The difference being in Step 2c. and Step 3. Step 2c. would be

$$\text{Compute } a_m = log((1 - err_m)/err_m) + log(K - 1)$$

Step 3 is

$$G(x) = argmax x_k \sum_{m=1}^{M} a_m I(G_m(x) = k)$$

Compared to the AdaBoost algorithm, they are actually equivalent when $K = 2$. And when $K > 2$, the difference is $log(K - 1)$. Thus multiclass boosting leads to a reweighting algorithm very similar to AdaBoost.

6 Problem 6

ESL Ex. 10.8

(a.) The log-likelihood function for this problem is given by

$$L(y, p(x)) = \sum_{k=1}^{K} I(y = \mathcal{G}_k) log(p_k(x)) = \sum_{k=1}^{K} I(y = \mathcal{G}_k) f_k(x) - log(\sum_{l=1}^{K} e^{f_l(x)}).$$

$$L'(y, p(x)) = \sum_{k=1}^{K} f_k(x) - log \sum_{l=1}^{K} e^{f_l(x)}$$

$$L''(y, p(x)) = -log \sum_{l=1}^{K} e^{f_l(x)}$$

(b.) Let's consider the total log-likelihood over all samples

$$LL = \sum_{x_i \in R} \sum_{k=1}^{K} y_{ik} f_k(x_i) - \sum_{x_i \in R} log(\sum_{l=1}^{K} e^{f_l(x_i)})$$

If we increment $f_k(x)$ by γ_k, we now have

$$LL(\gamma) = \sum_{x_i \in R} \sum_{k=1}^{K} y_{ik}(f_k(x_i) + \gamma_k) - \sum_{x_i \in R} log(\sum_{l=1}^{K} e^{f_l(x_i) + \gamma_l})$$

We are going to use Newton's algorithm to find the maximum log-likelihood with respect to $K-1$ values γ_k. In order to do so, we need to use the first and second derivatives with respect to these variables.

$$\frac{\partial}{\partial \gamma_k} LL(\gamma) = \sum_{x_i \in R} y_{ik} - \sum_{x_i \in R} \left(\frac{e^{f_k(x_i)+\gamma_k}}{\sum_{l=1}^{K} e^{f_l(x_i)+\gamma_l}} \right)$$

There are two cases to consider for the second derivative, when $k' \neq k$ and $k' = k$. They are respectively

$$\frac{\partial^2}{\partial \gamma_k \partial \gamma_{k'}} LL(\gamma) = -\sum_{x_i \in R} \frac{e^{f_k(x)+\gamma_k} e^{f_{k'}(x)+\gamma_{k'}}}{(\sum_{l=1}^{K} e^{f_l(x_i)+\gamma_l})^2}$$

$$\frac{\partial^2}{\partial \gamma_k \partial \gamma_k} LL(\gamma) = \sum_{x_i \in R} \left(-\frac{e^{2f_k(x)+2\gamma_k}}{(\sum_{l=1}^{K} e^{f_l(x_i)+\gamma_l})^2} + \frac{e^{f_k(x)+\gamma_k}}{(\sum_{l=1}^{K} e^{f_l(x_i)+\gamma_l})} \right)$$

One step of Newton's method will use a value for γ_0 and update to get γ_1:

$$\gamma_1 = \gamma_0 - \left(\frac{\partial^2 LL(\gamma)}{\partial \gamma_k \partial \gamma_{k'}} \right)^{-1} \frac{\partial LL(\gamma)}{\partial \gamma_k}$$

Let's start with $\gamma_0 = 0$, we get

$$\frac{\partial}{\partial \gamma_k} LL(\gamma = 0) = \sum_{x_i \in R} y_{ik} - \sum_{x_i \in R} p_{ik} = \sum_{x_i \in R} (y_{ik} - p_{ik})$$

$$\frac{\partial^2}{\partial \gamma_k \partial \gamma_{k'}} LL(\gamma = 0) = -\sum_{x_i \in R} p_{ik} p_{ik'} \text{ for } k' \neq k$$

$$\frac{\partial^2}{\partial \gamma_k \partial \gamma_k} LL(\gamma = 0) = \sum_{x_i \in R} (-p_{ik}^2 + p_{ik}) \text{ for } k' = k$$

where $p_{ik} = \frac{e^{f_k(x_i)}}{\sum_{l=1}^{K} f_l(x_i)}$. If the Hessian is diagonal then the matrix inverse becomes a sequence of scalar inverses, and the first Newton iteration becomes

$$\gamma_k^1 = \frac{\sum_{x_i \in R} (y_{ik} - p_{ik})}{\sum_{x_i \in R} p_{ik}(1 - p_{ik})}$$

for $1 \leq k \leq K-1$ which is what we wanted to show.

(c.) Given the form for normalized gammas, $\hat{\gamma}_k$, we have:

$$\hat{\gamma}_k = a\gamma_k^1 + b$$

We want $\hat{\gamma}_k$ to sum to zero. This requires that

$$\sum_{k=1}^{K} \hat{\gamma}_k = a \sum_{k=1}^{K} \gamma_k^1 + bK = 0$$

So we have

$$b = -\frac{a}{K}\sum_{k=1}^{K}\gamma_k^1$$

Plugging that back in, we get

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

$$\hat{\gamma}_k = a(\gamma_k^1 - \frac{1}{K}\sum_{k=1}^{K}\gamma_k^1)$$

where $a = \frac{K-1}{K}$ which is what we wanted to show.