Introduction to Statistical Learning
Chapter 4 Exercise

Penghao Chen

March 10, 2019

1. Plugging (4.2) into (4.3), we get:

$$\begin{aligned}\frac{p(X)}{1-p(X)} &= \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}} \\ &= e^{\beta_0+\beta_1 X}\end{aligned}$$

   Thus the equivalence between (4.2) and (4.3) is proved.

   Since the logarithm transformation is monotonously increasing, maximizing logistic function (4.2) and maximizing logit (4.4) agree with each other.

2. *Proof.* Taking log of (4.12), we get:

$$\begin{aligned}log\left[\frac{\frac{\pi_k}{\sqrt{2\pi}\sigma}exp(-\frac{(x-\mu_k)^2}{2\sigma^2})}{\sum_{l=1}^{K}\frac{\pi_l}{\sqrt{2\pi}\sigma}exp(-\frac{(x-\mu_l)^2}{2\sigma^2})}\right] &= log\left[\frac{\pi_k\sigma exp(-\frac{(x-\mu_k)^2}{2\sigma^2})}{\sum_{l=1}^{K}\pi_l exp(-\frac{(x-\mu_l)^2}{2\sigma^2})}\right] \\ &= log\left[\pi_k exp(-\frac{(x-\mu_k)^2}{2\sigma^2})\right] - log\left[\sum_{l=1}^{K}\pi_l exp(-\frac{(x-\mu_l)^2}{2\sigma^2})\right]\end{aligned}$$

   Since the second term is independent of k, we can treat it as a constant term and omit it in the maximization problem. Continue:

$$\begin{aligned}log\left[\pi_k exp(-\frac{(x-\mu_k)^2}{2\sigma^2})\right] &= log(\pi_k) + (-\frac{(x-\mu_k)^2}{2\sigma^2}) \\ &= \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} - \frac{x^2}{2\sigma^2} + log(\pi_k)\end{aligned}$$

Again, the term $log(\pi_k)$ is constant and we can omit it in the maximization problem. This gives us:

$$\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$$

which is (4.13)

□

3. *Proof.* Plugging (4.11) into (4.10), we get:

$$\begin{aligned} P(Y = k|X = x) &= \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \\ &= \frac{\frac{\pi_k}{\sqrt{2\pi}\sigma_k} exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2})}{\sum_{l=1}^{K} \frac{\pi_l}{\sqrt{2\pi}\sigma_l} exp(-\frac{(x-\mu_l)^2}{2\sigma_l^2})} \\ &= \frac{\frac{\pi_k}{\sigma_k} exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2})}{\sum_{l=1}^{K} \frac{\pi_l}{\sigma_l} exp(-\frac{(x-\mu_l)^2}{2\sigma_l^2})} \end{aligned}$$

Maximizing this function is equivalent to maximizing its log transformation, since natural logarithm is monotonously increasing. Taking log, we get:

$$log\left[\frac{\pi_k}{\sigma_k} exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2})\right] - log\left[\sum_{l=1}^{K} \frac{\pi_l}{\sigma_l} exp(-\frac{(x-\mu_l)^2}{2\sigma_l^2})\right]$$

Notice the second term is a constant given x, and therefore can be omitted in the optimization problem. This gives us:

$$\begin{aligned} log(\pi_k) - log(\sigma_k) - \frac{(x-\mu_k)^2}{2\sigma_k^2} &= -\frac{x^2}{2\sigma_k^2} + \frac{\mu_k}{\sigma_k^2}x - \frac{\mu_k^2}{2\sigma_k^2} + log(\frac{\pi - k}{\sigma_k}) \\ &= \delta_k(x) \end{aligned}$$

It is obvious that the discriminant function, $\delta_k(x)$, is quadratic with respect to x. The classifier based upon it is hence quadratic.

□

4. (The following calculations are only approximate.)

(a) 10%

(b) 1%

(c) $(10\%)^{100}$

(d) Since $\lim_{p\to\infty}(10\%)^p = 0$, when p is very large, no training observation will be close to the test observation.

(e) For $p = 1$: $x = 10\%$

For $p = 2$: $x^2 = 10\% \Rightarrow x \approx 31.6\%$

For $p = 100$: $x^{100} = 10\% \Rightarrow x \approx 97.7\%$

As p is increasing, the length of each side is also increasing dramatically. When p is very large, the KNN method has very little filtering power.

5.

(a) On the training set, we expect QDA to perform better because it is more flexible. On the test set, we expect LDA to perform better because QDA is likely to overfit the training set when the Bayes decision boundary is linear.

(b) On the training set, we expect QDA to perfrom better becasue it is more flexible. On the test set, we still expect QDA to perform better becasue LDA has higher bias when the Bayes decision boundary is non-linear.

(c) As the sample size increases, we expect the precision accuracy of QDA relative to LDA to improve, since the weekness of QDA is overfitting to the data hence higher variance, however, large sample size would decrease the classifier's variance.

(d) False. More flexibility leads to higher variance of the classifier. Given limited data points and a linear decision boundary, such higher variance would lead to worse test error than LDA.

6.

(a)

$$\begin{aligned} P(gettingA) &= \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} \\ &= \frac{1}{1+e^{-(-6+0.05\times 40+1\times 3.5)}} \\ &\approx 37.75\% \end{aligned}$$

(b)

$$\begin{aligned} \frac{1}{1+e^{-(-6+0.05\times x_1+1\times 3.5)}} &= 50\% \\ \Rightarrow x_1 &= 50 \end{aligned}$$

7. Several quantities to compute before applying Bayes' theorem:

$$\begin{aligned} f_1(x) &= \frac{\pi_1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu_1)^2}{2\sigma^2}) \\ &= \frac{\pi_1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2}) \end{aligned}$$

$$\begin{aligned} f_2(x) &= \frac{\pi_2}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu_2)^2}{2\sigma^2}) \\ &= \frac{\pi_2}{\sqrt{2\pi}\sigma} exp(-\frac{2}{9}) \end{aligned}$$

Now, apply Bayes' theorem, and we get:

$$\begin{aligned} P(Yes|X) &= \frac{\pi_1 f_1(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \\ &= \frac{\frac{\pi_1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2})}{\frac{\pi_1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2}) + \frac{\pi_2}{\sqrt{2\pi}\sigma} exp(-\frac{2}{9})} \\ &= \frac{0.8e^{\frac{1}{2}}}{0.8e^{\frac{1}{2}} + 0.2e^{-\frac{2}{9}}} \\ &\approx 75.2\% \end{aligned}$$

8. For KNN method when K=1, the training error rate is 0%. Given this, it can be estimated that the test error rate for KNN method when K=1 is greater than 30%. Hence the logistic regression should be preferred.

9.

(a) We have the following relations:

$$\begin{aligned} P(default) + P(nondefault) &= 1 \\ \frac{P(default)}{P(nondefault)} &= 0.37 \end{aligned}$$

Solving this, we get: $P(default) \approx 0.27$. Roughly 27% of people will default.

(b)

$$\frac{16\%}{1-16\%} \approx 0.19$$

The odds that she will default is 0.19.