An Introduction to Statistical Learning: Chapter 4

Due on January 5,2016 at 3:10pm

G. James et.al.

Tony Jiang

1. Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.

Answer:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad (4.2)$$

$$\begin{aligned}\frac{p(X)}{1-p(X)} &= \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}} \\ &= \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{\frac{1+e^{\beta_0+\beta_1 X}-e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}} \\ &= e^{\beta_0+\beta_1 X} \quad (4.3)\end{aligned}$$

2. It was stated in the text that classifying an observation to the class for which (4.12) is largest is equivalent to classifying an observation to the class for which (4.13) is largest. Prove that this is the case. In other words, under the assumption that the observations in the kth class are drawn from a $N(\mu_k, \sigma^2)$ distribution, the Bayes' classifier assigns an observation to the class for which the discriminant function is maximized.

Answer:

Recall 4.12 has the form:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} exp\big(-\frac{1}{2\sigma^2}(x-\mu_k)^2\big)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} exp\big(-\frac{1}{2\sigma^2}(x-\mu_l)^2\big)}$$

and 4.13

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$$

Taking the log of 4.12 on both sides:

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

To maximize this , we only need to maximize all items related to class k. that is to say, we

only care about what is a function of k (the rest are the same to a given x).

$$\begin{aligned}
\arg\max_x p_k(x) &= \arg\max_x log(p_k(x)) \\
&= \arg\max_x \quad log\Big(\pi_k \frac{1}{\sqrt{2\pi}\sigma} exp\big(-\frac{1}{2\sigma^2}(x-\mu_k)^2\big)\Big) \\
&= \arg\max_x \quad \Big(log(\pi_k) + log(\frac{1}{\sqrt{2\pi}\sigma}) - [\frac{1}{2\sigma^2}(x^2 - 2\mu_k x + \mu_k^2)]\Big) \\
&= \arg\max_x \quad \Big(log(\pi_k) - \frac{x^2}{2\sigma^2} + \frac{2\mu_k x}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2}\Big) \\
&= \arg\max_x \quad \Big(log(\pi_k) - \frac{x^2}{2\sigma^2} + \frac{\mu_k x}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}\Big) \\
&= \arg\max_x \quad \Big(x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)\Big)
\end{aligned}$$

The last line bears the same form as 4.13.

3. This problem relates to the QDA model,in which the observations within each class are drawn from a normal distribution with a class-specific mean vector and a class specific covariance matrix. We consider the simple case where $p = 1$; i.e. there is only one feature.

Suppose that we have K classes, and that if an observation belongs to the kth class then X comes form a one-dimensional normal distribution, X $N(\mu_k, \sigma_k^2)$. Recall that the density function for the one-dimensional normal distribution is given in (4.11). Prove that in this case, the Bayes' classifier is *not* linear. Argue that it is in fact quadratic.

Hint: For this problem, you should follow the arguments laid out in Section 4.4.2, but without making the assumption that $\sigma_1^2 = ... = \sigma_K^2$.

Answer:

$$
\begin{aligned}
\arg\max_x p_k(x) &= \arg\max_x log(p_k(x)) \\
&= \arg\max_x \quad log\Big(\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} exp\big(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\big)\Big) \\
&= \arg\max_x \quad \Big(log(\pi_k) + log(\frac{1}{\sqrt{2\pi}\sigma_k}) - [\frac{1}{2\sigma_k^2}(x^2 - 2\mu_k x + \mu_k^2)]\Big) \\
&= \arg\max_x \quad \Big(log(\pi_k) - \frac{x^2}{2\sigma_k^2} + \frac{2\mu_k x}{2\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2}\Big) \\
&= \arg\max_x \quad \Big(log(\pi_k) - \frac{x^2}{2\sigma_k^2} + \frac{\mu_k x}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2}\Big) \\
&= \arg\max_x \quad \Big(-x^2 \cdot \frac{1}{2\sigma_k^2} + x \cdot \frac{\mu_k}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + log(\pi_k)\Big)
\end{aligned}
$$

It does contain the x^2 term.

4. When the number of features p is large, there tends to be a deterioration in the performance of KNN and other *local* approaches that perform prediction using only observations that are *near* the test observation for which a prediction must be made. This phenomenon is known as the $curseofdimensionality$, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this curse.
(a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly (evenly) distributed on [0,1]. Associated with each observation is a response value. Suppose that we wish to predict a test observation?s response using only observations that are within 10 % of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$, curse of dimensionality we will use observations in the range [0.55,0.65]. On average, what fraction of the available observations will we use to make the prediction?
(b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1, X_2) are uniformly distributed on [0,1]×[0,1]. We wish to predict a test observation?s response using only observations that are within 10 % of the range of X_1 and within 10 % of the range of X_2 closest to that test observation. For instance, in order to predict the response for a test observation with X1 = 0.6 and X2 = 0.35, we will use observations in the range [0.55, 0.65] for X_1 and in the range [0.3, 0.4] for X_2. On average, what fraction of the available observations will we use to make the prediction?
(c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation?s response using observations within the 10 % of each feature?s range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

(d) Using your answers to parts (a)?(c), argue that a drawback of KNN when p is large is that there are very few training observations ?near? any given test observation.
(e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10 % of the train- ing observations. For p = 1,2, and 100, what is the length of each side of the hypercube? Comment on your answer.
Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When p = 1, a hypercube is simply a line segment, when p = 2 it is a square, and when p = 100 it is a 100-dimensional cube.

Answers:
(a). 10% of the observations will be used to make the prediction (because we are using the 10% of the range of [0,1].
(b): $10\% \times 10\% = 1\%$.
(c): $10\%^{100} = 10^{-98}\%$ (d): When p increases, the probability of having observations near the test observation decreases exponentially.
(e): when $p = 1, L = 0.1$;
when $p = 2, L = \sqrt{0.1} = 0.3162$; when $p = 100, L = \sqrt[100]{0.1}$.

5. We now examine the differences between LDA and QDA.
(a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?
(b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?
(c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?
(d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

Answers:
(a): If the Bayes decision boundary is linear, QDA is expected to perform better on the training set than LDA because it has higher flexiblity. On the test set, LDA is expected to perform better than QDA because LDA is closer to the true model. QDA will overfit the model.
(b). If the Bayes decision boundary is non-linear, QDA is expected to perform better on the training set than LDA because it has higher flexiblity. On the test set, QDA is expected to perform better than LDA because QDA is closer to the true model.
(c). We expect the test prediction accuracy of QDA relative to LDA to improve, in general, as the the sample size nn increases because a more flexibile method will yield a better fit as more samples can be fit and variance is offset by the larger sample sizes.

(d) False. QDA will suffer from "overfitting" thus yield a higher test error rate than LDA.

6. Suppose we collect data for a group of students in a statistics class with variables $X_1 =$ hours studied, $X_2 =$ undergrad GPA, and $Y =$ receive an A. We fit a logistic regression and produce estimated coefficient, $\beta_0 = -6$, $\beta_1 = 0.05$, $\beta_2 = 1$.
(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.
(b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

Answers:

(a).

$$\begin{aligned} p(Y = A|X_1 = 40 \quad and \quad X_2 = 3.5) &= \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2}} \\ &= \frac{e^{-6+0.05\times 40+1\times 3.5}}{1 + e^{-6+0.05\times 40+1\times 3.5}} \\ &= \frac{e^{-0.5}}{1 + e^{-0.5}} \\ &= 0.3775 \\ &= 37.75\% \end{aligned}$$

(b).

$$\begin{aligned} p(Y = A|X_1 \quad and \quad X_2 = 3.5) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2}} &= 50\% = 0.5 \\ e^{\beta_0+\beta_1 X_1+\beta_2 X_2} &= 1 \\ e^{-6+0.05\times X_1+1\times 3.5} &= 1 \\ -6 + 0.05 \times X_1 + 1 \times 3.5 &= 0 \\ X_1 &= 50(hours) \end{aligned}$$

7. Suppose that we wish to predict whether a given stock will issue a dividend this year (?Yes? or ?No?) based on X, last year?s percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\sigma^2 = 36$. Finally, 80 % of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Hint: Recall that the density function for a normal random variable is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$. *You will need to use Bayes' theorem.*

Answers:

$$
\begin{aligned}
Pr(X=4|Y=Yes) &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2} \\
&= \frac{1}{\sqrt{2\pi\times 36}}e^{-(4-10)^2/2\times 36} \\
Pr(X=4|Y=No) &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2} \\
&= \frac{1}{\sqrt{2\pi\times 36}}e^{-(4-0)^2/2\times 36} \\
Pr(Y=Yes|X=4) &= \frac{Pr(Y=Yes)*Pr(X=4|Y=Yes)}{Pr(Y=Yes)*Pr(X=4|Y=Yes)+Pr(Y=No)*Pr(X=4|Y=No)} \\
&= \frac{0.8\times\frac{1}{\sqrt{2\pi\times 36}}e^{-(4-10)^2/2\times 36}}{0.8\times\frac{1}{\sqrt{2\pi\times 36}}e^{-(4-10)^2/2\times 36}+0.2\times\frac{1}{\sqrt{2\pi\times 36}}e^{-(4-0)^2/2\times 36}} \\
&= 0.7518 \\
&= 75.18\%
\end{aligned}
$$

8. Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20 % on the training data and 30 % on the test data. Next we use 1-nearest neighbors (i.e. $K = 1$) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

Answers:
We should choose logistic regression. When K=1, KNN yields a 0% test error so we know the test error rate is 36% which is greater than logistic regression method (30%). We always prefer methods with lower test error rate.

9. This problem has to do with odds.
(a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?
(b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will de- fault?

Answers:

(a)

$$
\begin{aligned}
odds &= \frac{p(X)}{1-p(X)} \\
p(X) &= \frac{odds}{1+odds} \\
&= \frac{0.37}{1+0.37} \\
&= 0.27 = 27\%
\end{aligned}
$$

(b)

$$
\begin{aligned}
odds &= \frac{p(X)}{1-p(X)} \\
&= \frac{0.16}{1-0.16} \\
&= 0.1904
\end{aligned}
$$