Saarland University

The Elements of Stastical Learning

Assignment 3

Due Date: 29.11.2017

Thibault SCHOWING Mat. 2571837
Sarah MCLEOD Mat. 2566398
November 30, 2017

Problem 1

To show that the median distance from the origin to the closest data point is given by the expression:

$$d(p,N) = \left(1 - \frac{1}{2}^{\frac{1}{N}}\right)^{\frac{1}{p}}$$

First we note that since the data points are uniformly distributed, $P(data_point > d) = 1 - P(data_point < d)$.
Next, we since we have a spherical ball with uniformly distributed points, the probability that a point falls into the sphere is: ↷ explain more -0.5

$$P(data_point > d) = \frac{G(p) - G(p)d^p}{G(p)}$$

which simplifies to :

$$P(data_point > d) = 1 - d^p \checkmark$$

Since d is the median that means half the points are on one side and half the points are on the other. Therefore $P(data_point > d) = \frac{1}{2}$ ✓
then for a data point we have:

$$\frac{1}{2} = 1 - d^p$$

no -0.5

generalizing this to all data points (and assuming all the points were generated independently):

$$\frac{1}{2} = (1 - d^p)^N \checkmark$$

then we get:

$$\frac{1}{2} = (1 - d^p)^N$$

$$1 - \frac{1}{2}^{\frac{1}{N}} = d^p$$

$$(1 - \frac{1}{2}^{\frac{1}{N}})^{\frac{1}{p}} = d \checkmark$$

For $N = 100$ and $p = 10$, $d(p,N) \approx 0.608$. This is more than halfway to the boundary, which means most of the data points are closer to the boundary, where prediction is much harder for the KNN algorithm.✓

(7)

Problem 2

a

In which setting is logistic regression applicable? Why is linear regression not applicable in such a setting?

1.5

The logistic regression is applicable when the output space is restricted to being categorical (e.g. Y = 0 or 1). Linear regression is not applicable because the values are in the real domain and though can give values out of the desired range (e.g. negative values, when the output is between 0 and 1). It will also give an order to the output value (like Blue ¿ Red) which often makes no sense. ✗ masking -0.5

b

The odds formula is as below:

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

2

In a general meaning, odds and probabilities have the same signification but mathematically they are not the same. While a probability is expressed with a value between 0 and 1, the odds can have value from 0 to infinity.

In we take a classical example, we have a bag with 20 red balls, 40 blue and 40 yellow, so a total of 100 balls. The probability to get a red ball is $\frac{20}{100}$ but the odd is not the same. The denominator in the probability contains all the possibilities. In the odds, we remove the count of red balls so we have an odd to get a red ball of 20:80 which equals to $\frac{20}{80} = \frac{1}{4}$. So we have an odd of $\frac{1}{4}$ to pick up a red ball.

c

Prove that the equation $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$ is equivalent to $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$.

So we have $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$.
We simplify as follows:

1

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

So:

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

We then replace in $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$ and the two terms cancel:

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

d

In the previous exercise we saw that $Odds(X) = \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$.

Here we have:

0.5 you only change X_i here -0.5

$$\frac{Odds(X_i + \Delta)}{Odds(X_i)} = \frac{e^{\beta_0+\beta_1(X+\Delta)}}{e^{\beta_0+\beta_1(X)}}$$

This is equal to:

$$e^{\beta_0+\beta_1 X+\beta_1\Delta} e^{-\beta_0-\beta_1 X} = e^{\beta_1\Delta}$$

If the random variable X is an important feature and so have a high β_1, any small change in X will have a big effect on the odds. The bigger the β_1 the bigger the effects.

e

We have

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} = 0.5$$

The Odds are:

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

1

So we have:

$$Odds(X) = \frac{0.5}{0.5} = e^{\beta_0+\beta_1 X} = 1$$

So

$$\beta_0 + \beta_1 X = 0$$

And

$$\frac{-\beta_0}{\beta_1} = X \checkmark$$

If the probability is 0.5, it means that X lies on the hyperplane (a.k.a Decision boundary) which equation is $\beta_0 + \beta_1 X = 0$. It means that it cannot be classified. ✓

f

The book introduces the conditional probabilities and the log-odds for 2-way logistic regression. Extend this model to logistic regression for k response classes.

For classification we have that: $p(X) = Pr(Y = 1|X)$ so we can rewrite:

$$log - odds(X) = log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X$$

as:

$$log(\frac{Pr(Y=1|X)}{Pr(Y=0|X)}) = \beta_0 + \beta_1 X$$

We can generalize it as:

$$log(\frac{Pr(Y=k|X)}{Pr(Y=k-1|X)}) = \beta_0 + \beta_1 X_1 + ... + \beta_k - 1 X_k - 1$$ no -1

Which is equal to:

What about the intercept

$$log(\frac{Pr(Y=k|X)}{Pr(Y=k-1|X)}) = \sum_{i=1}^{k-1} \beta_i X_i$$

4

Remove the $log()$:

$$\frac{Pr(Y=k|X)}{Pr(Y=k-1|X)} = e^{\sum_{i=1}^{k-1} \beta_i X_i}$$

Multiply by the denominator to have $Pr(Y=k|X)$:

$$Pr(Y=k|X) = e^{\sum_{i=1}^{k-1} \beta_i X_i} \cdot Pr(Y=k-1|X)$$

And for $Pr(Y=k-1|X)$:

$$Pr(Y=k-1|X) = Pr(Y=k|X) \cdot \frac{1}{e^{\sum_{i=1}^{k-1} \beta_i X_i}}$$

We can substitute $Pr(Y=k-1|X)$ in the previous equation and we have:

$$Pr(Y=k|X) = \frac{e^{\sum_{i=1}^{k-1} \beta_i X_i}}{1 + e^{\sum_{i=1}^{k-1} \beta_i X_i}}$$ ✓

So we finally have:

$$Pr(Y=k-1|X) = \frac{1}{1 + e^{\sum_{i=1}^{k-1} \beta_i X_i}}$$ ✓

10

Problem 3

0 **a**

b

The difference between LDA and QDA is that in QDA, the response variables are still drawn from a Normal distribution but each class has its own covariance matrix. ✓ Thus, the posterior probability $Pr(Y = k|X = x)$ abbreviated $p_k(x)$ below, cannot be reduced. (4.11 ISLR)

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

5 The denominator here is still constant. It is a sum over all the classes. So, we have to maximize the numerator. We will here consider the log of the numerator and thus we have:

$$\delta_k(x) = log\left(\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} exp\left(-\frac{1}{2\sigma_k^2(x-\mu_k)^2}\right)\right)$$

We expand the $exp()$ content and separate the log and get:

$$\delta_k(x) = log(\pi_k) - log(\sqrt{2\pi}\sigma_k) - \frac{x^2}{2\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + \frac{x\mu_k}{2\sigma_k^2} \checkmark$$

We can see that the $\frac{x^2}{2\sigma_k^2}$ term is quadratic. Because of the σ_k values, that are different in opposition with the LDA, it is not possible here to cancel it. So the Bayes classifier here is not linear ✓

(5)

Problem 4

2 1. See R code for training and test set splits. ✓

3 2. For the LDA model fit according to the data split in (a) the error is calculated as the percentage of correctly classified items. The training error is 0.056 and the test error is 0.080. ✓

0 3. ✗

3 4. For the phonemes aa and ao the training error is 0.1064 and the test error is 0.2141.

6 5. For a QDA model fit using all phoneme data the train error is 0, and the test error is 0.158. For a QDA model fit on only phonemes aa and ao the train error is 0 and the test error is 0.3394. The training error is lower in the QDA model, however this model is also overfitting the data, as evidenced by the test error being much higher than the LDA model. In this example we would prefer the LDA model.

6. For the LDA model the confusion matrix for aa and ao is:

	aa	ao	total
aa	439	80	519
ao	56	703	759
Total	495	783	

✗ confusion matrices on test data -1.5

1.5 The confusion matrix for the QDA model for aa and ao is:

	aa	ao	total
aa	519	0	519
ao	0	759	759
Total	519	759	

These tables show that, at least on the training data, the QDA model has better sensitivity and better specificity. The QDA correctly identifies all training instances of each aa and ao phoneme. The LDA model on the other hand misclassified aa as ao on 80 instances and misclassified ao as aa 56 times. ✓

15.5

97.5