Saarland University

The Elements of Stastical Learning

Assignement 5

Due Date: 03.01.2018

Thibault Schowing *Mat. 2571837*
Sarah Mcleod *Mat. 2566398*
December 28, 2017

Problem 1

Principal Components Analysis The first principal component is the direction of maximum variance in the data. Show that this first principal component also minimizes the residual sum of squares, which is here the squared distance between the projected data point and the original data point.

The RSS is given by:

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

The least squares regression model is given by the equation 6.16 and 6.17 of ISRL:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \qquad \text{(6.16, ISLR)}$$

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i, \quad i = 1, ..., n \qquad \text{(6.17, ISLR)}$$

In the RSS equation, $f(x)$ correspond to the value projected on the first component line and is given by: $f(x) = \phi(x_i - \bar{x})$ (6.19, ISLR). So we have RSS:

$$RSS = \sum_{i=1}^{n}(\theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i - \phi(x_i - \bar{x}))^2$$

The first component has the highest variance so $Var(f(x))$ is a maximum and so the RSS is minimized.

NOT SURE OF THE $\phi(x_i - \bar{x})$ thing !!!!

Problem 2

a

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

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.

b

The odds formula is as below:

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

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:

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

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:

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

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:

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

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

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

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:

$$log(\frac{Pr(Y=k|X)}{Pr(Y=k-1|X)}) = \beta_0 + \beta_1 X_1 + ... + \beta_k - 1 X_k - 1$$

Which is equal to:

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

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:

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

Problem 3

Problem 4