Saarland University

The Elements of Stastical Learning

Assignement 2

Due Date: 15.11.2017

Thibault SCHOWING
Sarah MCLEOD
November 12, 2017

Problem 1

Derive the variance formula:

$$Var(\frac{1}{k}\sum_{i=1}^{k} X_i) = \rho\sigma^2 + \frac{1-\rho}{k}\sigma^2$$

Problem 2

Problem 3

TheR^2 statistic is a common measure of model fit corresponding to the fraction of variance in the data that is explained by the model. In general, R^2 is given by the formula

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

The objective here is to show that for univariate regression, $R^2 = Cor(X,Y)^2$

Let's take first the RSS and TSS formula

$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ =amount of variability left unaccounted after the regression

$TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ = total variance in Y

$TSS - RSS$ = amount of variance removed/explained by the regression

So we have

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

And the correlation formula

$$\widehat{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

For an univariate linear regression, the approximation has the form $(af + b)$ and we'll suppose an exact approximation with $(y - f)^2 = 0$. We will also assume that f is the model that minimize the square-error and that there is no shift of scaling that improve f.

We can write it like this

$$\sum_{i=1}^{n}(f_i - y_i)^2 \leq \sum_{i=1}^{n}(af_i + b - y_i)^2$$

Consider the second part as $g(a,b) = \sum_{i=1}^{n}(af_i + b - y_i)^2$. As f minimizes the loss, g is optimized at $g(1,0)$ and so it is the optimum. We can derive it

$$\frac{d}{da}g(a,b)_{(a=1,b=0)} = \sum_{i=1}^{n} 2(af_i + b - y_i)f_i = \sum_{i=1}^{n} 2(f_i - y_i)f_i = 0$$

So $yf - ff = 0 \rightarrow yf = ff$

And

$$\frac{d}{db}g(a,b)_{(a=1,b=0)} = \sum_{i=1}^{n} 2(af_i + b - y_i) = \sum_{i=1}^{n} 2(f_i - y_i) = 0$$

So $\bar{y} = \bar{f} \rightarrow$ Mean is the normalized sum.

We can simplify R^2. From the derivative we have $yf = ff$ and $\bar{f} = \bar{y}$. To simplify we assume $\bar{y}$ and $\bar{f} = 0$

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i)^2} = 1 - \frac{yy - 2yf + ff}{yy}$$

Since we have $yf = ff$

$$R^2 = 1 - \frac{yy - ff}{yy} = \frac{ff}{yy}$$

For the correlation we have

$$\rho = \frac{\sum\limits_{i=1}^{n}(f_i y_i)}{\sqrt{\sum\limits_{i=1}^{n} f_i^2}\sqrt{\sum\limits_{i=1}^{n} y_i^2}} = \frac{fy}{\sqrt{(ff)(yy)}} = \frac{ff}{\sqrt{(ff)(yy)}} = \sqrt{\frac{ff}{yy}}$$

So $R^2 = \rho^2$

Problem 4