Homework 2 Machine Learning

Marco Treglia

December 5, 2016

1 Linear Regression

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

One variable, denoted x, is regarded as the predictor, explanatory, or independent variable. The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

2 Mathematical step

Since we are interested in summarizing the trend between two quantitative variables, the natural question arises "what is the best fitting line?" In order to examine which of the two lines is a better fit, we first need to introduce some common notation:

y_i denotes the observed response for experimental unit i

x_i denotes the predictor value for experimental unit i

$\hat{y_i}$ is the predicted response (or fitted value) for experimental unit i

Then, the equation for the best fitting line is:

$$\hat{y_i} = b_0 + b_1 x_i$$

Where for computing b_0 and b_1 as we need first calculate the correlation r_{xy}, the standar deviation σ and the mean μ of x_i and y_i :

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2}$$

Then b_1 and b_0 are :

$$b_1 = r_{xy} \frac{\sigma_x}{\sigma_y}$$

$$b_0 = \mu_y - b_1 \mu_x$$

In general, when we predict the actual response to y_i, we make a prediction error (or residual error) of size:

$$e_i = y_i - \hat{y}$$

And for have a better

2.1