Simple Book Example

TeXstudio Team

January 2013

Contents

Chapter 1

The First Chapter

Chapter 2

Regression

2.1 Evaluating Regression Models Performance

R square

$$SS_{hot} = SUM(y_i - y_{avg})^2 \tag{2.1}$$

$$SS_{hot} = SUM(y_i - y_{avg})^2 \tag{2.2}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{2.3}$$

Here, we want $SS_{res} \rightarrow Min$

The problem here is: if we have n features (regressor) have been already existed in our regression model. We will add a new feature, we want to know if it helps to improve the performance of hour model. The solution is: we compare the difference of R^2 with and without the new feature. There are two situations here:

1. R^2 increase. Because the new feature decrease the SS_{res}
2. R^2 keeps unchanged. Because the new feature does not help the model. It has no effect on the dependant variable. The coefficient of the new feature is zero.

When add a new feature, it is bias as the R^2 is always increase. So the adjust R square is proposed.

$$AdjR^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \tag{2.4}$$

where p is the number of regressors and n is the sample size.

One example is shown as follows:

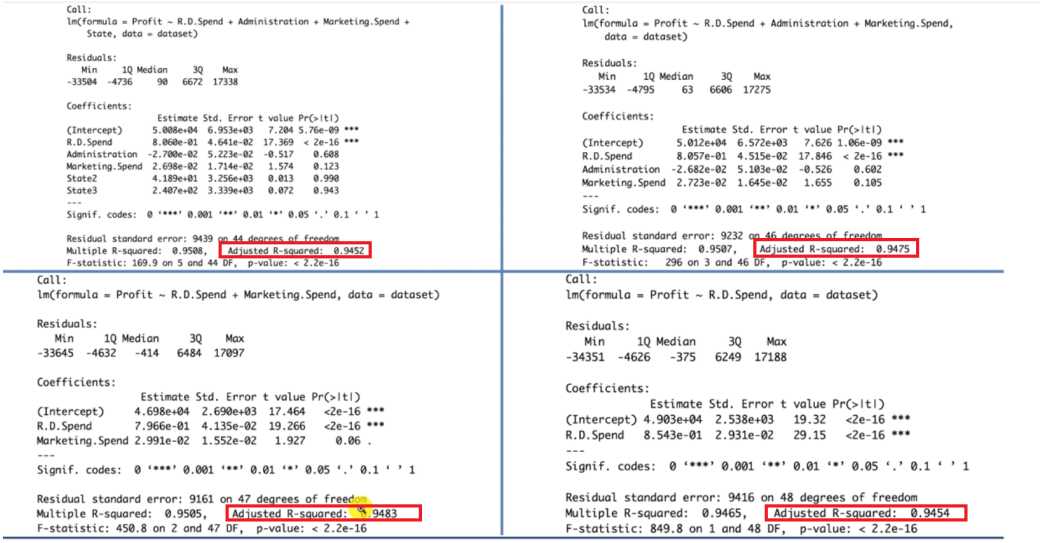

Figure 2.1: example of adjust square error

Regression Model	Pros	Cons
Linear Regression	Works on any size of dataset, gives information about relevance of features	The Linear Regression Assumptions
Polynomial Regression	works on any size of dataset, works very well on non linear problems	Need to choose the right polynomial degree for a good bias/variance tradeoff
SVR	Easily adaptable, works very well on non linear problems, not biased by outliers	Compulsory to apply feature scaling, not well known, more difficult to understand
Decision Tree Regression	Interpretability, no need for feature scaling, works on both linear/nonlinear problems	Poor results on too small datasets, overfitting can easily occur
Random Forest Regression	powerful and accurate, good performance on many problems, including non linear	Nointerpretability, overfitting can easily occur, need to choose the number of trees

Table 2.1: Comparison of different regression models

Table 2.1 shows the comparison of different regression models from comparison.

How to address overfitting problems:Regularization.

Chapter 3

Classification

Chapter 4

Clustering

4.1 K-Means Clustering

4.2 Hierarchical Clustering