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:

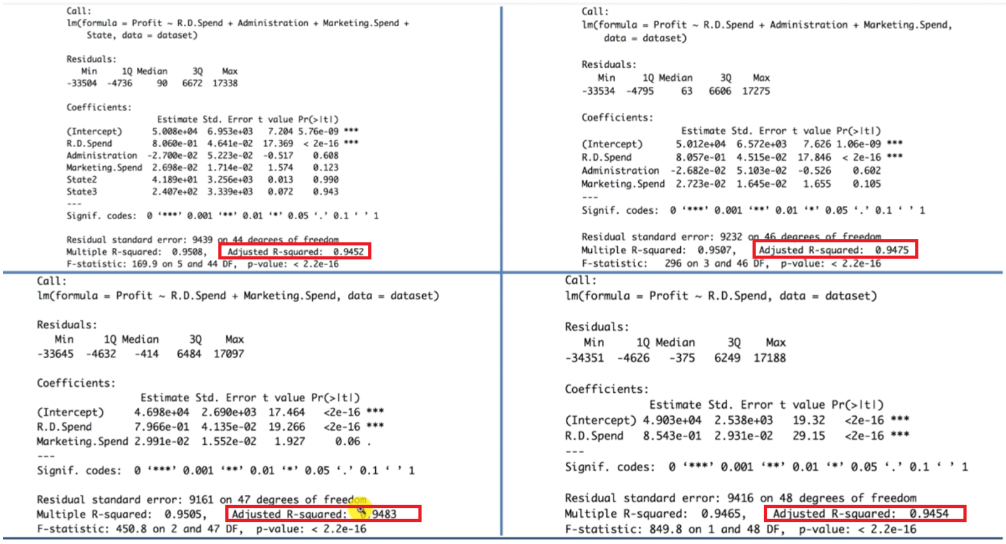
1. R^2 increase. Because the new feature decrease the SS_{res}
2. R^2 keeps unchanged. Because the new feature does not help the model. It has no effect on the dependant variable. The coefficient of the new feature is zero.

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:



```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +
    State, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33504  -4736     90   6672  17338

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.008e+04  6.953e+03   7.204 5.76e-09 ***
R.D.Spend       8.060e-01  4.641e-02  17.369  < 2e-16 ***
Administration -2.700e-02  5.223e-02  -0.517    0.608
Marketing.Spend 2.698e-02  1.714e-02   1.574    0.123
State2          4.189e+01  3.256e+03   0.013    0.990
State3          2.407e+02  3.339e+03   0.072    0.943
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9439 on 44 degrees of freedom
Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9452
F-statistic: 169.9 on 5 and 44 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,
    data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33534  -4795     63   6606  17275

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.012e+04  6.572e+03   7.626 1.06e-09 ***
R.D.Spend       8.057e-01  4.515e-02  17.846  < 2e-16 ***
Administration -2.682e-02  5.103e-02  -0.526    0.602
Marketing.Spend 2.723e-02  1.645e-02   1.655    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9232 on 46 degrees of freedom
Multiple R-squared:  0.9507,    Adjusted R-squared:  0.9475
F-statistic:   296 on 3 and 46 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-34351  -4626   -375   6249  17188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.903e+04  2.538e+03   19.32   <2e-16 ***
R.D.Spend   8.543e-01  2.931e-02   29.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9416 on 48 degrees of freedom
Multiple R-squared:  0.9465,    Adjusted R-squared:  0.9454
F-statistic: 849.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

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

3.1 Logistic Regression

Logistic regression is a linear regression, which returns the possibility.

$$y = b_0 + b_1 * x \tag{3.1}$$

$$p = \frac{1}{1 + e^{-y}} \tag{3.2}$$

How to evaluate the performance of classification: $confusion\_matrix(y_{true}, y_{pred})$

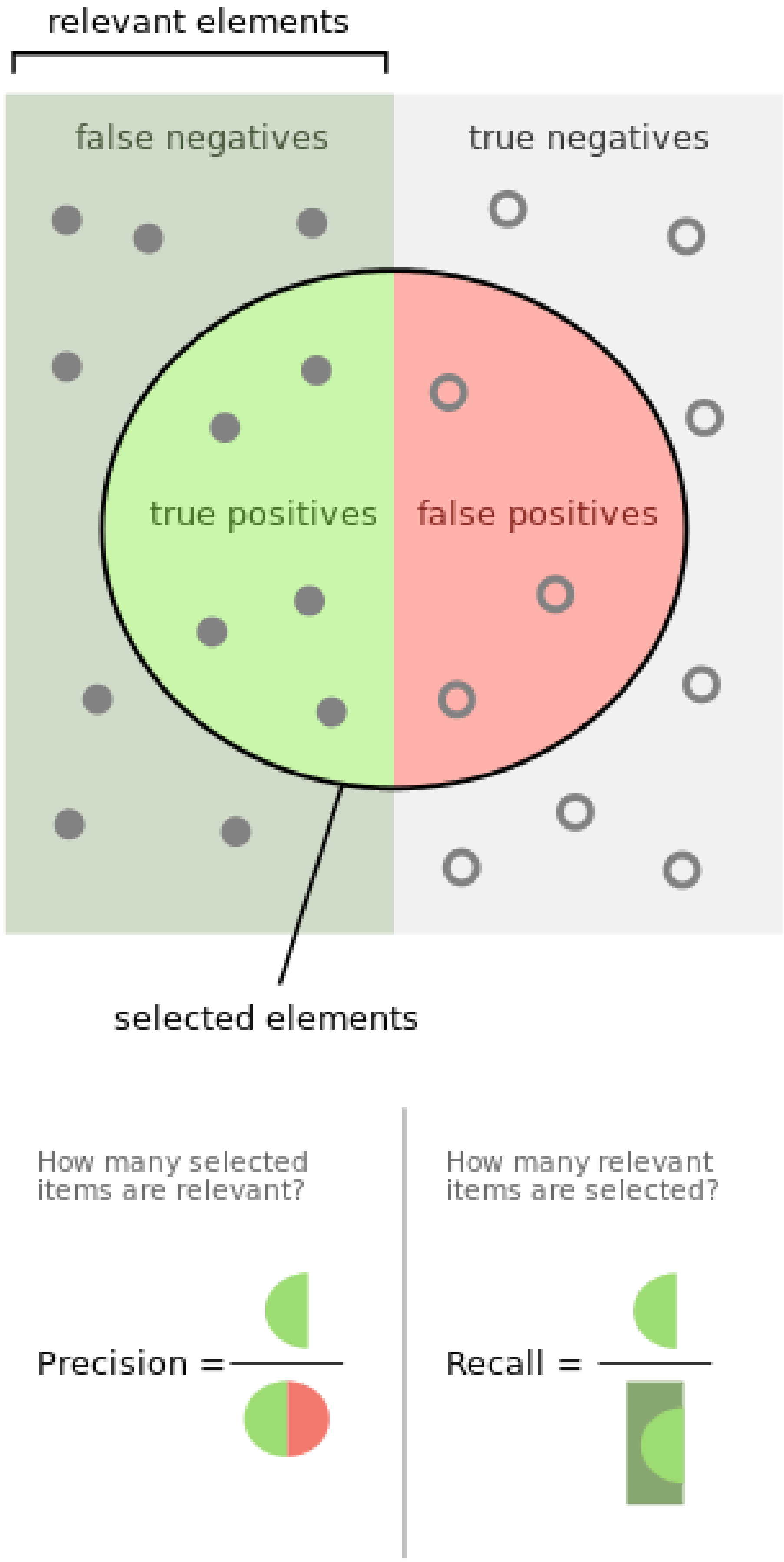
By definition a confusion matrix C is such that $C_{i,j}$ is equal to the number of observations known to be in group i but predicted to be in group j. Thus in binary classification, the count of true negatives is $C_{0,0}$, false negatives is $C_{1,0}$, true positives is $C_{1,1}$ and false positives is $C_{0,1}$

How to understand true positive, false positive, true negative, false negative, recall, precision?

Suppose a computer program for recognizing dogs in photographs identifies eight dogs in a picture containing 12 dogs and some cats. Of the eight dogs identified, five actually are dogs (true positives), while the rest are cats (false positives). The program's precision is 5/8 while its recall is 5/12. When a search engine returns 30 pages only 20 of which were relevant while failing to return 40 additional relevant pages, its precision is $20/30 = 2/3$ while its recall is $20/60 = 1/3$. So, in this case, precision is "how useful the search results are", and recall is "how complete the results are".

3.2 K-Nearest Neighbors (K-NN)

1. choose the number K of neighbours


relevant elements
false negatives
true negatives
true positives
false positives
selected elements
How many selected
items are relevant?
How many relevant
items are selected?
Precision =
Recall =

2. take the K nearest neighbours of the new data point, according to the Euclidean distance
3. among these K neighbours, count the number of data points in each category
4. assign the new data point to the category where you counted the most neighbours

3.3 Support Vector Machine (SVM)

3.4 Kernel SVM

3.5 Naive Bayes

3.6 Decision Tree Classification

3.7 Random Forest Classification

3.8 Evaluating Classification Models Performance

Chapter 4

Clustering

4.1 K-Means Clustering

4.2 Hierarchical Clustering