RAZAVET Mael 06/03/2014

**Data Mining – Exercise 7**

# Question 1:

We have:

â = 1-0.0667=0.9333

n = 145 instances.

Before calculating the 95% interval for the expected error, we need to make sure the normal distribution is a good approximation for the binomial one (distribution of the estimated accuracy of a single test set).

If na(1-a) < 5, then this would lead to asymmetric confidence intervals. Otherwise, we can assume the normal distribution is a good approximation and we can construct the confidence intervals.

So, according to the skew of the sampling distribution, the normal distribution is a good approximation to construct symmetric confidence intervals.

According to the normal density function used to determine the 95% confidence interval for the expected error, the 95% of area lies in .

Let’s then compute the standard deviation .

The interval will be []

Therefore, the 95% interval for the expected error is: [0.8926888; 0.9739112].

We are 95% confident that the expected error falls in the interval [0.89; 0.97].

# Question 2:

In this question, you can assume that each fold would have at least 30 instances so that the accuracy follows a normal distribution.

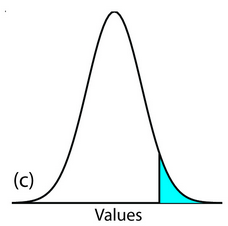
Our statistical hypothesis is that algorithm 1 will outperform algorithm 2.

Therefore, our null hypothesis is:

and if it is not rejected, then the algorithm 1 will outperform the algorithm 2 at the confidence level.

and we will assume the algorithm 1 will outperform the algorithm 2.

Here is a representation of our one-tailed test:



What is the confidence level that will allow us to accept this hypothesis?

To do so, we need to use the paired t-test.

The following table provides the accuracies for the 10-fold cross validation method over two different algorithms. I also computed the average and the standard deviation of the accuracies.

|  |  |  |
| --- | --- | --- |
| CV Fold | Algorithm 1 | Algorithm 2 |
| 1 | 91.11 | 90.7 |
| 2 | 90.48 | 90.52 |
| 3 | 91.87 | 90.88 |
| 4 | 90.52 | 90.87 |
| 5 | 89.88 | 90.02 |
| 6 | 89.77 | 88.99 |
| 7 | 91.44 | 90.98 |
| 8 | 90.88 | 91.44 |
| 9 | 90.77 | 90.77 |
| 10 | 90.89 | 90.92 |
| Avg | 90.761 | 90.609 |
| Standard deviation | 0.6445403 | 0.6730272 |

Let’s compute the t-test:

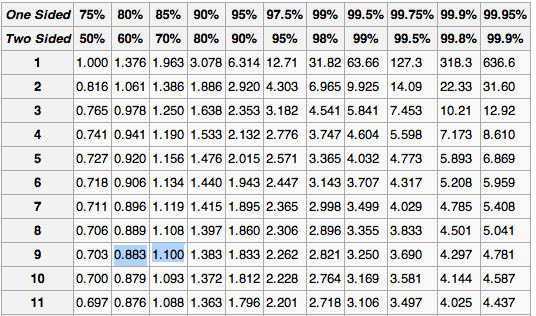
|  |  |
| --- | --- |
| Fold | Algorithm 1 – Algorithm 2 |
| 1 | 0.41 |
| 2 | -0.04 |
| 3 | 0.99 |
| 4 | -0.35 |
| 5 | -0.14 |
| 6 | 0.78 |
| 7 | 0.46 |
| 8 | -0.56 |
| 9 | 0 |
| 10 | -0.03 |
| Avg | 0.152 |
| Stdev | 0.4938916 |

The mean and the sample standard deviation are calculated like the following:

The t-statistic value is computed below:

equals 0 here because of our null hypothesis.

Then, we compare to the values in the t-distribution table. The degree of freedom to use here is 9 (because we have 10 folds).



According to the table, the confidence level of that hypothesis would be between 80 and 85%. So, we are between 80 and 85% confident that the algorithm 1 will outperform the algorithm 2. Therefore, we are between 15 and 20% not confident about this assumption.

# Question 3:

The question 3 has been generated via the pandoc package in R, to produce a pdf of my code with my working and the interpretations that I made.