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.

$$skew = n * \hat{a}(1 - \hat{a}) = 9.064095 > 5$$

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 $\mu \pm 1.96\sigma$.

Let's then compute the standard deviation σ.

$$\sigma = \sqrt{\frac{\hat{a}(1 - \hat{a})}{n}} = 0.02071999$$

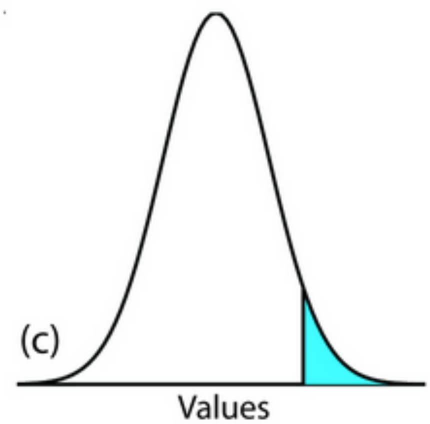
The interval will be $[0.9333 - 1.96\sigma;\ 0.9333 + 1.96\sigma]$
Therefore, the 95% interval for the expected error is: [0.8926888; 0.9739112].
We are 95% confident that the expected error falls in the interval [0.89; 0.97].

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:
$H_0: \mu_0 \leq 0$ and if it is not rejected, then the algorithm 1 will outperform the algorithm 2 at the confidence level.
$H_1: \mu_0 > 0$ 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:

$$\bar{x} = \frac{0.41 - 0.04 + 0.99 - 0.35 - 0.14 + 0.78 + 0.46 - 0.56 + 0 - 0.03}{10} = \frac{1.52}{10} = 0.1520$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{9}[(0.41 - 0.152)^2 + (-0.04 - 0.152)^2 + \cdots + (-0.03 - 0.152)^2]}$$

$$= \sqrt{\frac{1}{9}[0.0666 + 0.0369 + \cdots + 0.0331]}$$

$$= \sqrt{\frac{1}{9}\times 2.1954}$$

$$= \sqrt{0.243929}$$

$$= \ 0.4938916$$

The t-statistic value is computed below:

$$t = \frac{avg - \ \mu_0}{(\frac{stdev}{\sqrt{n}})} = \frac{0.152}{(\frac{0.4938916}{\sqrt{10}})} = \ 0.9732221$$

μ_0 equals 0 here because of our null hypothesis.

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

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437

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.