LINGI2262: Assignment 4
Performance Assessment

Group 8: Maxime Dimidschstein, Doriane Olewicki

April 26, 2018

Q1

Q1.1

Let a decision tree model M be used to classify 1000 independent test examples. We consider the events :

A : M classifies the third test example as positive;

B : M classifies the third test example as negative.

Those events are :

Random with the test examples. The model being fixed and deterministic, if we always use the same test example, we always get the same result for the third one. However, if we consider the test set random, then the outcome for the third example is also random.

Not independent because to have independence, $P(A|B)$ must be equal to $P(A)$ and $P(B|A) = P(B)$. But in this case, B is actually equal to $\neg A$, the probability of $P(A|B)$ and $P(B|A)$ is thus null.

Mutually exclusive because it is not possible for the model to classify the third test example at the same time as positive and negative. It will only choose one.

Q1.2

A pair of mutually exclusive events can also be independent if the events respect the two following properties.

$$A \cap B = \emptyset => P(A \cap B) = 0 \tag{1}$$
$$P(A \cap B) = P(A)P(B) \tag{2}$$

To do so, $P(A) = 0$ or $P(B) = 0$. As an example, we can respect this with $A = E$, the set of all possible outcomes, and $B = \emptyset$.
But in the general case, when considering events A and B with non-zero probability : in that case, they cannot be both mutually exclusive and independent at the same time.

Q1.3

The probability of failing at example x follows the geometric distribution $P(X = x) = p^{x-1}(p-1)$, with $p = 0.75$ (the probability to classify an example with success).
The expected number of examples processed before the first fail is thus computed as:

$$E[X] = \frac{p}{1-p} = \frac{0.75}{0.25} = 3$$

The failure happens on the fourth example, so we expect to need a total of 4 examples.

The variance corresponding to this distribution is

$$V(X) = \frac{p}{(1-p)^2} = \frac{0.75}{0.25^2} = 12$$

The probability to have to get the third fail on the y^{th} example is $P(Y = y) = \binom{y-2}{2} p^{y-3}(p-1)^3$. This negative binomial distribution has the following expected value for the third fail:

$$E[Y] = \frac{3}{1-p} = \frac{3}{0.25} = 12$$

With the third fail on the 12^{th} example.

Q2

Q2.1

The RF model classifies correctly the $n = 100$ test examples with a probability of $\hat{p}_{RF} = 0.78$. The SVM model classifies correctly the $n = 100$ independent test examples with a probability of $\hat{p}_{SVM} = 0.75$.

We know that $\hat{p} \sim N(p, \frac{p(1-p)}{n})$ and the The 95% confidence interval is given by $\hat{p} \pm Z_N \sqrt{\frac{p(1-p)}{n}}$ with $Z_N = 1.95$. We do not know p but we estimate it at $\hat{p}$ so the interval is given by $\hat{p} \pm Z_N \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
This gives for the RF model a 95% confidence interval of $\hat{p}_{RF} = [0.699; 0.861]$ and for the SVM model a 95% confidence interval of $\hat{p}_{SVM} = [0.665; 0.835]$.

To check if the RF model is equivalent to the SVM model, we compute $\hat{p}_{SVM} - \hat{p}_{RF}$ and check if zero is in the interval. The outcome interval is $[-0.034; -0.026]$ and does not contain zero : we can thus say that the two models are not equivalent. The values on the interval being negative, the RF model may be better.

Q2.2

We define the null hypothesis H_0 and the alternative hypothesis H_a :

$$H_0 : p = p_{RF} - p_{SVM} = 0 \qquad\qquad H_a : p \neq 0$$

We have $\alpha = 5\%$.
The test statistic is $T = \hat{p}_{RF} - \hat{p}_{SVM} = 0.78 - 0.75 = 0.03$. T is normally distributed :

$$T \sim N(p, \sigma_p^2) \Rightarrow \frac{T-p}{\sigma_p} \sim N(0, 1)$$

with $\sigma_p = \sqrt{\frac{\hat{p}_{RF}(1-\hat{p}_{RF})}{n} + \frac{\hat{p}_{SVM}(1-\hat{p}_{SVM})}{n}} = 0.0599$.
Under the null hypothesis $p = 0$, $t = \frac{T-p}{\sigma_p} = \frac{0.03-0}{0.0599} = 0.501$.
We now need to check if $t < -z_{\alpha/2}$ or if $t > z_{\alpha/2}$ with here $z_{\alpha/2} = z_{0.025} = 1.96$. We have $-z_{\alpha/2} < t < z_{\alpha/2}$. We thus accept, with 95% confidence, that the true error rates are not significantly different.

To compute the p-value, we must check for which limit value the relation $t = z_{\alpha/2}$.

$$\begin{aligned} t &= Z_{\alpha/2} \\ \Rightarrow \alpha/2 &\approx 0.308 \\ \Rightarrow \alpha &\approx 0.617 \text{ (the p-value)} \\ \Rightarrow 1-\alpha &\approx 0.383 \end{aligned}$$

Q2.3

To have a p-value below 1%, we at least need a z_n value of 2.58. We thus have the following equation :

$$\begin{aligned} z_n &= t \\ z_n &= \frac{T-p}{\sigma_p} \\ 2.58 &= \frac{0.03-0}{\sqrt{\frac{0.78(1-0.78)+0.75(1-0.75)}{n}}} \\ n &= (\frac{2.58}{0.03})^2 \cdot (0.78(1-0.78)+0.75(1-0.75)) \\ n &= 2655.904 \approx 2656 \end{aligned}$$

Q3

Q3.1

Based on the information we have, we can expect the quality measures to approach 50% the higher k becomes. The classification accuracies being uniformly distributed over $[0\%; 100\%]$, their sample mean should be normally distributed around 50%, with variance $\frac{\sigma^2}{k}$, with $\sigma = 8.333\%$ being the variance of the classification accuracies. So when k increases, the variance of the quality measures decreases.

Q3.2

The boxplots on figure 1 show the spread of the quality measures over 100 runs for increasing values of k. It appears clearly that our expectations are met: the larger k, the closer to 50% the quality measures and the lower the variance.

Q3.3

If the classification test accuracies were distributed non-uniformly, the Central Limit Theorem tells us we could still consider their sample mean tends to be normally distributed with greater k. And the variance of this distribution would still become smaller with increasing k.

Q4

Q4.1

When selecting randomly one test sample containing 50 distinct examples from `BostonHouseTest.csv`, we build a decision tree by CART. We use this tree to see the classification rate with 10% of the

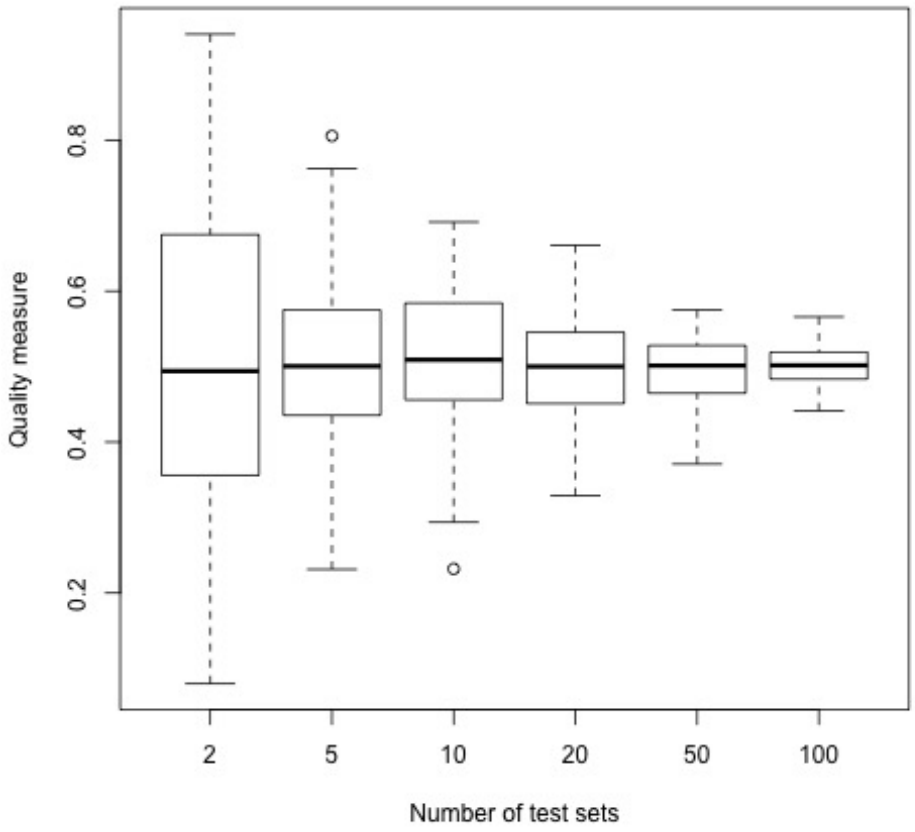

Figure 1: Quality measure as a function of number of test sets

total training examples. The result was that $\hat{p} = 0.76$ of the sample was well classified. We can thus derive a 95% confidence interval for the classification rate with the formula $\hat{p} \pm Z_N \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$:

$$interval_1 = [0.642; 0.878]$$

Q4.2

For this question, we need to sample 50 distinct test samples and repeat this random sample 100 times. The mean classification rate is $\hat{p} = 0.77$. The lower and upper bound observed are $interval_2 = [0.66; 0.88]$.
When computing a 95% interval on this mean classification rate, we obtain the interval $interval_3 = [0.655; 0.888]$.
Those three intervals are quite similar.

Q4.3

When recomputing several times the above intervals, we observe that the bounds and means vary. The random sampling of the initial single test set and the initial training set influence the results. We even got an interval from $[0.4, 0.6]$ where the sampling was really bad. Getting a "good" interval depends on the random sampling and is thus, by definition, random.