INGI2262 – Machine Learning

Assignment 4

Performance Assessment

1 Questions

Q 1 Review quiz on probability and statistics.

Q 1.1 Let a decision tree model M be used to classify 1000 independent test examples. We consider the event *A: M classifies the third test example as positive* and *B: M classifies the third test example as negative*. Are those events **random** considering that the model is fixed and deterministic? **independent**? **mutually exclusive**? Explain **why**.

Q 1.2 In general, can a pair of mutually exclusive events also be independent? If so, or not so, under which condition(s)? **Argue.**

Q 1.3 A RBF kernel SVM is estimated on a validation set to have a 75% classification accuracy. When this model is run to classify a stream of new test examples, what is the expected number[1] of examples to be processed until the ***first*** classification error appears? How do you expect this number to ***vary*** according to the actual set of test examples? What is the expected number of examples to be processed until the ***third*** classification error is observed[2]?

Q 2 A random forest classifies correctly 78 out of 100 test examples. An SVM with a RBF kernel classifies correctly 75 out of 100 independent test examples.

Q 2.1 Compute 95% confidence intervals for the performances of each of the two models. Do you conclude from these intervals that the RF model is better than the SVM model?

Q 2.2 Apply a statistical test to decide whether the performances of the two models differ. Use a 5% level of significance. Specify which statistical test you are using and why. What is the p-value of your test?

Q 2.3 What is the minimal number of test examples (instead of the 2×100 examples originally used) that would be needed in the previous test to get a p-value below 1%, assuming the test classification rates do not change (78% versus 75%)?

[1] Include the first error in the total number of examples required.
[2] Include all 3 errors in the total number of examples required.

Q 3 A well informed data analyst observes that a machine learning package is apparently bugged as it produces models that, once tested on independent test examples, have classification accuracies distributed uniformly in the interval $[0\%, 100\%]$. To support his hypothesis, the data analyst implements the following protocol. He repetitively learns models with the package under study and he observes the accuracies on independent test samples. More specifically, he reproduces this experiment (learn and test) over several independent training/test sets and he reports as quality measure the ***average*** of the test accuracies observed on k such test sets. As his results could depend on particular tests, he repeats the whole protocol over 100 distinct runs. He then plots the distribution of this quality measure over the 100 runs and he monitors how this distribution evolves as a function of the number $k = 2, 5, 10, 20, 50, 100$ of test sets considered.

Q 3.1 From what you know about the problem at hand, how do you expect the distribution of the quality measure to behave as a function of k? Why?

Q 3.2 Simulate numerically the protocol of the data analyst and check whether these simulations satisfy the expected results from your theoretical analysis of the problem. Report plot(s) to support your claims.

Q 3.3 What would differ in the previous analysis if the classification test accuracies would be distributed **non-**uniformly? Why?

Q 4 In a previous assignment you were asked to compute a learning curve for CART with 10 different random selections of various fractions (5%, 10%, 20%, 50%, 99%) of the ***BostonHouseTrain.csv*** training set. We study here the variations of test classification rates observed on **various test samples**.

Q 4.1 Select randomly one test sample containing 50 distinct examples from ***BostonHouseTest.csv***. Compute a 95% confidence interval for the classification rate obtained with a decision tree built by CART with only 10% of the total training size.

Q 4.2 Sample at random 50 distinct test examples from ***BostonHouseTest.csv***, and repeat this random sampling 100 times to form 100 test folds (each containing 50 test examples[3]). Consider the same decision tree as in question 4.1 (and built on a 10% fraction of the total training set) and compute the classification rates obtained with this model on the 100 different test folds. Compute the mean classification rate and observed lower and upper bounds. The observed lower (respectively upper) bound is such that 2.5 % of the 100 classification rates are below it (respectively above it). Turn in a table comparing the lower and upper bounds of the confidence interval (as computed in question 4.1) and your observed bounds.

Q 4.3 The above sub-questions refer to the comparison between a confidence interval computed on a single test set and the variability observed across 100 test sets. Could your conclusion be affected by the random sampling of the initial single test set and/or the initial training set? If you repeat the whole protocol (questions 4.1 and 4.2) several times, does it change your conclusions?

Submit a **PDF** report with answers to the questions of this assignment. Add any comment you consider relevant about the use of the R software for this task.
Submit as well on Moodle the **R code** that you wrote for this assignment. Put your **PDF** report and your **R code** all together in a zip archive.

[3] Since ***BostonHouseTest.csv*** does not include 100×50 examples, the various test sets are expected to overlap partially.