Q&A session – Week 40: Practice Questions with solutions

Q1: A dataset of 1000 examples was partitioned into a training set of 600 examples and a test set of 400 examples. A 1-NN model has an accuracy of 80% on the test set. It was subsequently found that the partitioning had been done incorrectly and that 100 examples from the training set had been accidentally duplicated and had overwritten 100 examples in the test set. What is the accuracy for the 300 cases that were truly part of the test set?

80% on the test set means that 320 examples are classified correctly. If we remove 100 examples from the test set, then only 220 examples were classified correctly. Hence, the accuracy is 220/300.

Q2: When generating forests of classification trees, it is often recommended that the individual trees should not be pruned to maximize the predictive performance of the forest. Explain why pruning may have a detrimental effect on the forest, even in cases when it improves the predictive performance of the individual trees in the forest.

In order to satisfy the Condorcet's jury theorem each individual classifier should be as independent as possible from the rest. Hence each individual tree should be as independent as possible from the rest. One way of achieving this is by letting each tree overlearn the training set.

Q3: Mary and Bob are given a classification problem on a dataset containing 1000 real-valued features (among which, the last feature is the class label) and 1 million examples (records), with no missing values. Their task is to compare the performance of 1-NN and random forests on the given dataset. Bob decides to build both classifiers using all data examples and the first 999 features. To his surprise, when he evaluates the performance of the built classifiers on the same data examples and same features, he observes that both classifiers achieve close to 99.9% precision and recall. Mary decides to take an alternative approach and instead performs 10-fold cross-validation using all 1000 features. Again, both classifiers achieve recall and precision close to 99.9%. When they both present their approaches and results to their boss, they instantly get fired. Explain what was wrong with their approaches.

Bob trained and tested on the same dataset. This is a major flaw as the test set should always be independent of the corresponding training set. Mary on the other hand trained and tested correctly by using 10-fold cross-validation. However, she used the class label both as feature and class label. This is a major flow.

Q4: Why do we want to use "weak" learners when *boosting*? Explain briefly.

There are three main benefits: (1) you get speedier classifiers since weak learners tend to be faster than strong learners, (2) weak learners tend to avoid overfitting, and (3) the improvement in accuracy is substantial when using several weak learners compared to using many strong learners.

Q5: Explain the difference between the concepts of *bias* and *variance* in predictive modeling. Give an example of a predictive model with high bias and low variance, as well as an example of a model with low bias and high variance.

Refer to lecture 10.

High bias + low variance: Perceptron, decision stump, linear SVM

Low bias + high variance: random forest, neural networks

Q6: Suppose you are given a dataset with 4 attributes (F1, F2, F3, and F4). The class label is contained in attribute F4. Suppose that you build a random forest classification model and you test its performance using 10-fold cross-validation. For building the model you have used all four attributes (F1, F2, F3, and F4). The precision and recall of your experiment are both close to 100%. Is there anything that went wrong? Would you obtain similar performance if you used a decision tree instead? Motivate your answers carefully.

You have used the class label as a feature. This is independent of the type of classifier you use.

Q7: According to the curse of dimensionality, as the number of data dimensions increases the notion of similarity becomes meaningless. Explain the rationale behind this and discuss at least one way to go about solving the problem.

As the number of data dimensions increase then distance measures tend to use many features to come up with differencing. The more the features the higher the probability of ending up with examples with a similar distance distribution. Hence, the sense of distance becomes meaningless. One solution to this problem is dimensionality reduction, e.g., via PCA.