Error Analysis

340-420-DW - 00001

ZiAn Gao, Yiyi Yao

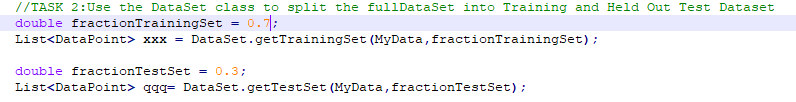
**Distributions of Model Accuracy**

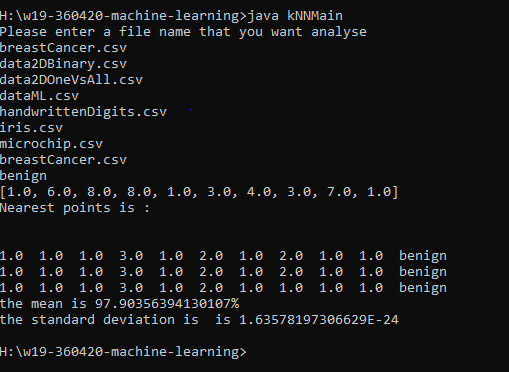
1. Each time you run the classification model, you should be getting a different accuracy. Why? (hint: lines 148-150 in DataSet.java)

At line 149, the command “**Collections. shuffle ()** ” means randomly permuting the elements in the list. It is to re-scramble the collection (random sorting).

1. Run the entire classification process 1000 times (load data, split into off 30% for a test set, evaluate model performance)

We get accuracy of 97.90% and a standard deviation 1.635\*10^-24, based on three nearest point.





1. What is a sensible baseline against which we should compare our model's performance? (hint: line 200 in DataSet.java)

Line 200: public static void printLabelFrequencies(List<DataPoint> fullDataSet) {

We analysis the “fullDataSet” and use the result of “printLabelFrequencies” as a baseline to predict the data. It is not accuracy because it only count the frequency of the specified element in the list rather than evaluating all the elements in the list. And baseline is used to improve our approach or solution to the problem.

**Analysis of different error types**

1. What is a False Positive:

False positive is an error in data reporting in which a test result improperly indicates presence of a condition, such as a disease (the result is positive), so in our breast cancer data sheet, a false positive error is the data show malignant but actually it is benign.

1. What is a False Negative:

False negative is an error in which a test result improperly indicates no presence of a condition (the result is negative), when in reality it is present. In our breast cancer data, a false negative means that the data show benign but it is malignant.

1. Extend your analysis in the previous step (with the 1000 runs) to keep track of Recall and Precision as well.

* What makes these two measures different?

Recall is the number of correct results divided by the number of results that should have been returned. Precision is the number of correct results divided by the number of all returned results.

* What are sensible baseline for each of these measures?

Precision is the percentage of the results which are relevant, and how many of the selected object were correct. Recall refers to the percentage of total relevant results correctly classifies by your algorithm.

1. How do the above results change with the hyperparameter k?

A hyperparameter is a parameter whose value is set before the learning process begins, the value of other parameters are derived via training. K is the number of closet neighbors that are used to determine the class of data point. The results will change when hyperparameter k change.