Report on Error Analysis

Introduction to Computer Programming – Section 3

Chulin Liang, Lily Bai

Distributions of Model Accuracy

1. Each time you run the classification model, you should be getting a different accuracy. Why?

According to the line 148-150 in DataSet, a different accuracy occurs since every time the model runs, the order of data are changed.

1. Run the entire classification process 1000 times.

The model has an accuracy of 97.63%, and the standard deviation is 8.021\*10^-25.

1. What is a sensible baseline against which we should compare our model's performance?

The baseline should be 65.20%, which is (the frequency of benign)444/681(the total number of cells).

1. False Positive?

A false positive is an error in data reporting in which a test result improperly indicates presence of a condition, when in reality it is not present.

1. False Positive?

A false negative is an error in which a test result improperly indicates no presence of a condition, when in reality is present.

1. Recall & Precision

What makes these two measures different?

The false positive and false negative make the precision and recall different. Recall bases on the percentage of the amount of true positive(e.g If the target of a search is benign, then we say that the benign elements selected are true positive ) in the sum of true positive and false negative(elements benign but considered as malignant and not selected).

Precision bases on the percentage of the amount of true positive in the sum of true positive and false positive(elements malignant but considered and selected as benign).

What are sensible baseline for each of these measures?

The baseline is percentage of the number of true positive can be selected in the all positive during the searching.

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

Hyperparameter is already set instead of the parameter obtained after the process of learning .