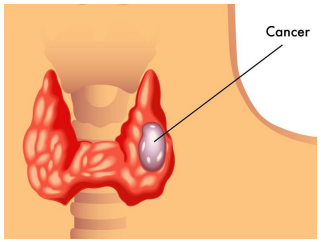
Machine Learning Classifier for Preoperative Diagnosis of Benign Thyroid Nodules

Blanca Villanueva[1], Joshua Yoon[2]

[1]Department of Computer Science, [2]Applied Physics

Introduction

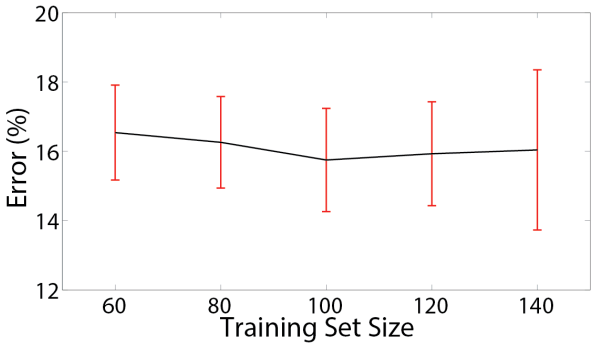
Although thyroid nodules are common and usually benign, they prove to be malignant in 5-15% of cases. The high rates of **fine-needle aspiration biopsies** (FNAB) indeterminacy coupled with this large variance in malignancy percentage mean that most individuals with thyroid nodules larger than 1cm in diameter are referred for surgical consultation for a next step towards diagnosis. 15-30% of results obtained via this common testing method, are labeled indeterminate. These individuals referred for diagnostic surgery are exposed to a 2-10% risk of serious surgical complications.



These biopsies contain gene expression data on **167 genes**. Each of these gene expression data is represented as a numerical value in the data set.

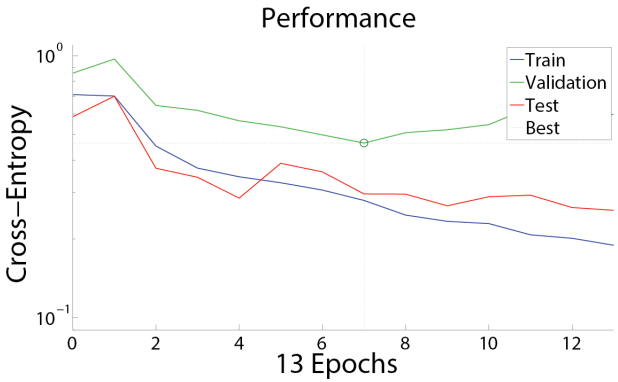
The data were collected from a study conducted by scientists from multiple centers over 19 months 1; the data set itself consists of **367 FNAB specimens** (47 benign, 55 malignant, and 265 indeterminate).

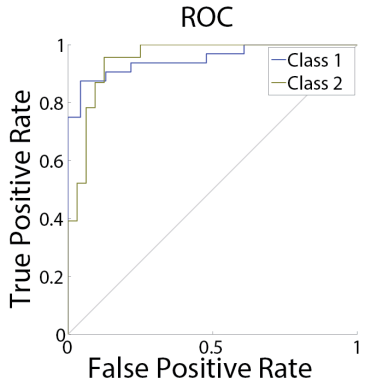
Naïve Bayes



We implemented Naïve Bayes (NB) as a baseline model to see what range of errors we would obtain by taking into account all features presented in the data as a function of training set size. We opted for NB over Logistic Regression as NB has shown better performance for smaller sample sizes. Each point on the plot is a mean of errors over a run 50 test runs with error bars representing the respective standard deviations.

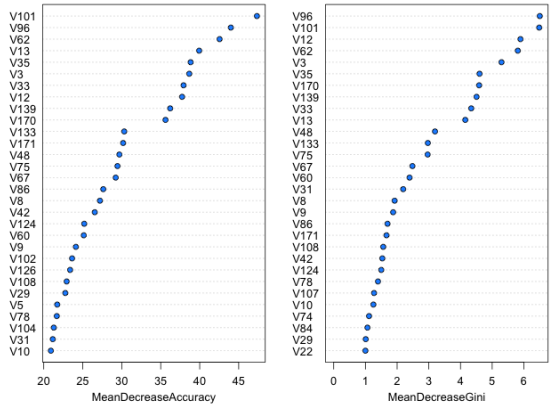
Neural Network





Using neural networks, we were able to obtain a test classification error of about 11% (55 test samples), the best results out of all the algorithms implemented.

Variable Selection with Random Forests



Overall Analysis

For evaluation of generalization error, we used a held-out set of 55 test samples out of the original data set of 366 samples (one sample proved invalid).

Due to the large feature space compared to sample size, we decided to try to reduce our feature space using supervised learning methods. We chose to implement variable selection methods instead of dimensionality reduction due to that identifying which specific gene expressions were correlated with the response (if any) could serve useful for future studies. The variables selected by our Random Forests model complemented our implementation of Neural Networks quite well. We were able to effectively reduce our feature space by an order of magnitude without significantly affecting our generalization error on the held-out set. On a larger data-set, this order of magnitude reduction in the feature space could significantly reduce computational complexity.