CALIFORNIA STATE SCIENCE FAIR
2012 PROJECT SUMMARY

Name(s)
Abraham P. Karplus

Project Number
S1413

Project Title
Machine Learning Algorithms for Cancer Diagnosis

Abstract

Objectives/Goals
Machine learning algorithms can be used for cancer diagnosis, but which are best? This project compares four algorithms: Decision Tree, Majority, Nearest Neighbors, and Best Z-Score (my own design; a slight variant of the Naive Bayes algorithm) at diagnosing cancer type in two datasets: breast cancer and colorectal cancer.

Methods/Materials
Both datasets were gene expression levels from tumor cells. For breast cancer, the algorithms predict basal or luminal (breast cancer types). For colorectal cancer, they predict the p53 mutation. A cross-fold validation program split the data into training and testing sets: the algorithms trained on 80% of the samples and were tested on the rest for performance and time. This train-test split was done 150 times for each algorithm on each dataset.

Results
For the breast cancer dataset, the Best Z-Score algorithm did best. All three levels of Decision Tree were adequate but slow. Majority was fast but did terribly. Nearest Neighbors was perfect using few neighbors, but as bad as Majority when using many neighbors. For the colorectal dataset, the Best Z-Score algorithm again performed best. All three Decision Trees performed comparably and only slightly worse than Best Z-Score, but took 50-140x longer. Majority was again terrible. Nearest Neighbors was 4x slower than Best Z-Score and had a performance about that of Decision Tree (a bit worse than Best Z-Score); its best performance used a medium number of neighbors.

Conclusions/Discussion
In summary, Best Z-Score did very well on all tests. Nearest Neighbors did extremely well on easy tasks and acceptably on hard ones. My guess that Decision Tree would work best was not vindicated, as it performed reasonably but took over an order of magnitude longer to train than the others.
I learned a lot about the field of machine learning, especially as I implemented all of the algorithms myself, and even designed the most effective of them.
To continue the project, I will implement the algorithms Random Forest, Support Vector Machine, and Naive Bayes, some of the most popular currently in use. I am especially curious how Naive Bayes would perform, since my current best, Best-Z Score, was based on Naive Bayes. Another expansion is to implement feature selectors and other dimensionality reducers, which could greatly speed up the algorithms (especially Decision Tree) and improve their performance.

Summary Statement
How well do several different machine learning algorithms do at diagnosing cancer from gene expression levels?

Help Received
Father suggested this project and connected me to Cancer Machine Learning group at UCSC. Cancer Machine Learning group provided cancer data and project suggestions. Mother helped with the writing and time management.