Brad Pershon  
Data Mining

**Proposal**

Background

Can we determine if a tumor is benign or malignant from a digitized image? That is the question we are attempting to answer by analyzing the Breast Cancer Wisconsin dataset. Our project will utilize an ensemble learner to determine if measurements taken from digital images can indeed determine tumor status. Our model will examine 30 attributes and use the Adaboost algorithm outlined in class.

Rationale

The goal of the final project in our Data Mining course will be to implement the ensemble classifier Adaboost. Since the algorithm was introduced I’ve wanted to experiment with my own implementation. Time permitting I would like to experiment with an Adaboost implementation built on top of Decision Trees and k-Nearest-Neighbors (k-NN) to see which performs better. If this is approved I may look to use an off the shelf decision tree model from Scikit-Learn so I can focus more on parameter experiments.

Dataset

We will use the Breast Cancer Wisconsin Dataset from the UCI Machine Learning repository to test our implementation. The dataset consist of 569 records with 30 numeric features. The set is broken up into 2 classes: 357 records are classified as benign tumors and 212 records are classified as malignant tumors. With the true class labels we can measure the classifiers performance using the performance metrics discussed in class such as precision and recall.

Software

For our Adaboost implementation we plan to use Python with Numpy and Pandas. For the base classifier we may use Scikit-learn’s decision tree implementation or design our own. We will also utilize the Weka Adaboost model to compare our results against an off the shelf model.

Collaborators

Unfortunately I don’t really know anyone in the class, so I will be the sole collaborator on this project. If for some reason another student needs a group I am open to the possibility, assuming they can pull their weight.