Aliya Tyshkanbayeva
Proposal

In this project, my main goal is to compare Naive Bayes Classifier and Logistic regression on two large data sets and visualize their performance errors. I was inspired to work on this topic after reading a paper by Andrew Y. Ng and Michael I. Jordan "*Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes*". My best friend George who is at the moment doing his bachelors in computer science and mathematics at Oxford University was the one who got me extremely interested in machine learning and data science. So, after reading the introduction and experiments sections, i decided that I want to qualitatively produce some of the experimental results in this paper.

The data sets I am using in discovering the performance errors of NBC and LG are following:

1. Congressional Voting Data Set, which is basically voting records in US House of Representatives in 1984. I am aiming to predict whether the representative is a republican or democratic.

http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

2. Iris Data Set, which contains information of three varieties of iris. There are four features, all real-valued, measurements of sepal length and width and petal length and width

https://archive.ics.uci.edu/ml/datasets/iris

For both dataset, I am going to compare the classification errors of the NBC and LR trained on large training datasets. If I have time, I also want to recap the experiments on missing values for voting dataset.