University of Texas at Arlington

CSE 5334 – Data Mining
Final Project Report
Naïve Bayesian learning for classifying news text articles

Sindoora Ravikumar Murthy
1001862126

Overview:

Given the goal of training a Naïve bayes classifier, we must use the appropriate methods to train and test the model. This paper will explain my thought process in how I selected what kind of model I decided on, how I implemented it in Python and the experiments.

Data:

Naïve Bayes classifiers are one among the most successful known algorithms for learning to classify text documents. The dataset contains 20,000 newsgroup messages drawn from the 20 newsgroups. The dataset contains 1000 documents from each of the 20 newsgroups.

Download dataset: http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naivebayes.html

The 20 news groups:

alt.atheism	rec.sport.hockey
comp.graphics	sci.crypt
comp.os.ms-windows.misc	sci.electronics
comp.sys.ibm.pc.hardware	sci.med
comp.sys.mac.hardware	sci.space
comp.windows.x	soc.religion.christian
misc.forsale	talk.politics.guns
rec.autos	talk.politics.mideast
rec.motorcycles	talk.politics.misc
rec.sport.baseball	talk.religion.misc

Need for this application:

With this text classification model, I was able to automate the entire process of picking relevant articles and tagging them under their respective categories, saving 80% of time. Manual text classification often results in human error from distraction, fatigue, or inconsistent criteria. Our machine learning model eliminates potential for this error, and creates a centralized and accurate way to categorize texts.