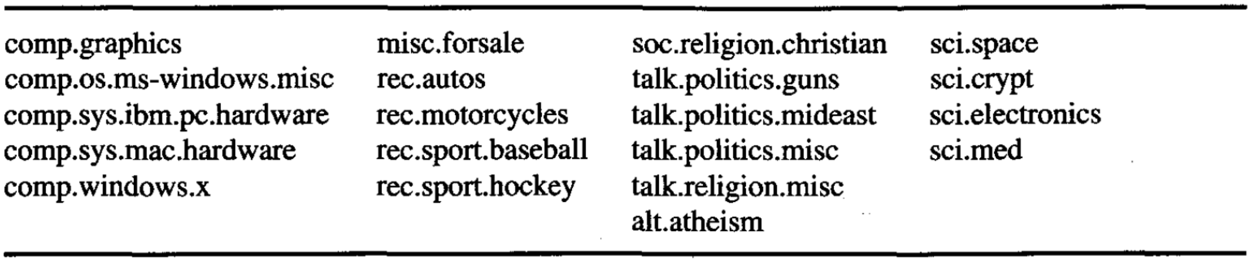
Bayesian Learning for Classifying Internet News Text Articles:

Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents. We will provide a dataset containing 20,000 newsgroup messages drawn from the 20 newsgroups. The dataset contains about 1000 documents from each of the 20 newsgroups (some groups have a little bit less documents), which are shown as following:



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

1. Download the data from http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html (Newsgroup Data)

2. For each data point, please remove the first four lines, i.e. the lines starting with Newsgroup, document_id, From, Subject.

3. Split data into two groups, and use half data as training data and the other half as testing data. Note: split the data of each class into two halves.

4. Please implement the Naive Bayes classifier by yourself (in Python). **Don't use any online code.** (You can consider the top 200 most frequent words as stop words and remove them from the vocabulary)

5. Use training data to compute the model parameters. Predict each testing data and compare the predicted label to the ground truth label. The average prediction accuracy is calculated.

6. Submit Jupyter notebook at ELMS