Aaron Zhang
CS 189
April 9, 2015

CS189 HW5

Decision tree techniques implemented: Stopping criteria was done by stopping if a split caused everything to go to one side, which meant that the segmentor could not find a split that can better classify the data, or if all the data was correctly classified, or if the tree reached the maximum depth allowed. Usually the tree stopping criteria was due to the first two cases and rarely the third. Decision forests utilized both data bagging and attribute bagging. I picked data (with replacement) for each tree and gave 1500 samples to each tree. Attribute bagging was done by restricting randomly 10 features to choose for each decision node.

Features added: Some common key words found in spam and a measure of how many non-ascii characters found in text.

Results: I got about 90% on the training data and 81% on the Kaggle score.

Most common splits: on 25 trees, with random bagging of 10 attributes at each node:

FEATURE	SPLIT_VALUE	#OCCURENCES
19 (path)	0.0	10
28(#)	0.0	8
3(bank)	0.0	2
16(volumes)	0.0	2
33(viagra)	0.0	1
31([])	0.0	2

As the data shows, bag of words models are pretty sparse and also a complete lack of a particular word turns out to be a good split value for predicting spam/ham. A count of non-alphanumeric features was also used quite often but not shown in this random forest.