BigML Assignment 1: Streaming Naive Bayes

Runyun Zhang
runyunz@andrew.cmu.edu

1. **What changes could you make to reduce the amount of RAM required for the dictionary? (5 points)**
   One way is to compress the key, e.g. instead of saving “Y=CCAT,W=w1”, we can save the key as “CCAT,w1”.

2. **Right now we’re basically ignoring the fact that there are multi-labeled instances in the train/test sets. How would you extend your algorithm to enable it to predict multiple labels? (5 points)**
   One way is to count words several times for different labels in a document. And during the classification, one can set up a threshold to decide if a document belongs to one class or not.
   Another way is to treat a multiple labels as a new class (e.g. {a, b}) and perform the classification.

3. **Why should we use Laplace smoothing? What will happen if we don’t use any smoothing? (5 points)**
   Laplace smoothing helps mediate the strong effect of P=0, which may happens if the example in the training set is skewed or insufficient, and the word in the test never appears. If there is no smoothing, then the test documents with new words cannot be classified correctly due to the P=0 issue.

4. **What is the relationship between Laplace smoothing and Dirichlet prior? (5 points)**
   Laplace smoothing is equivalent with the expected value of the posterior distribution, using a Dirichlet prior with parameter α (set to 1 in our case).

5. **Your answers to collaboration policy (on the first page of this handout).**
   Did you receive any help whatsoever from anyone in solving this assignment? No.
   Did you give any help whatsoever to anyone in solving this assignment? No.