COMS4721 - HW 3

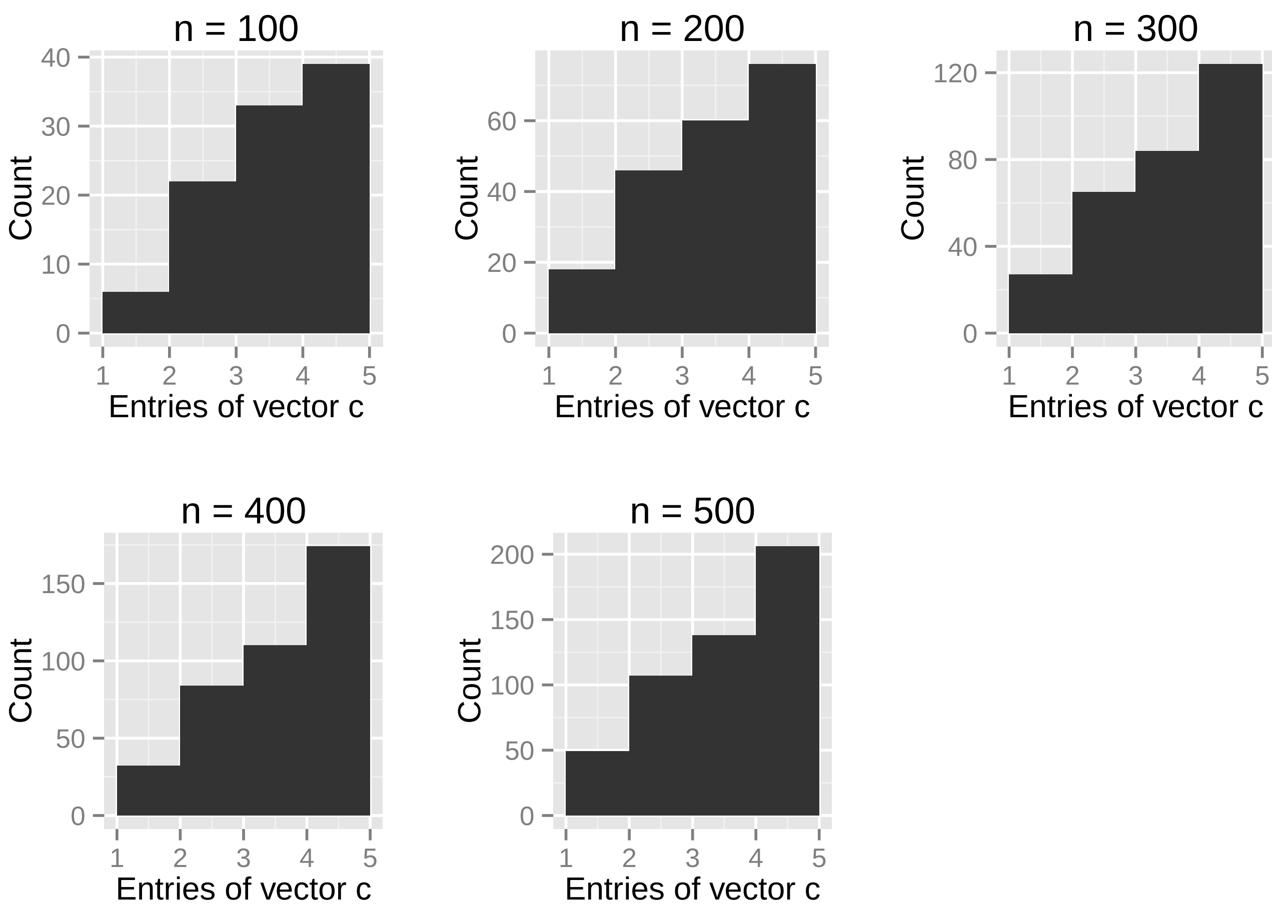
Jeff Hudson (jdh2182)

Tuesday, March 31, 2015

Part 1

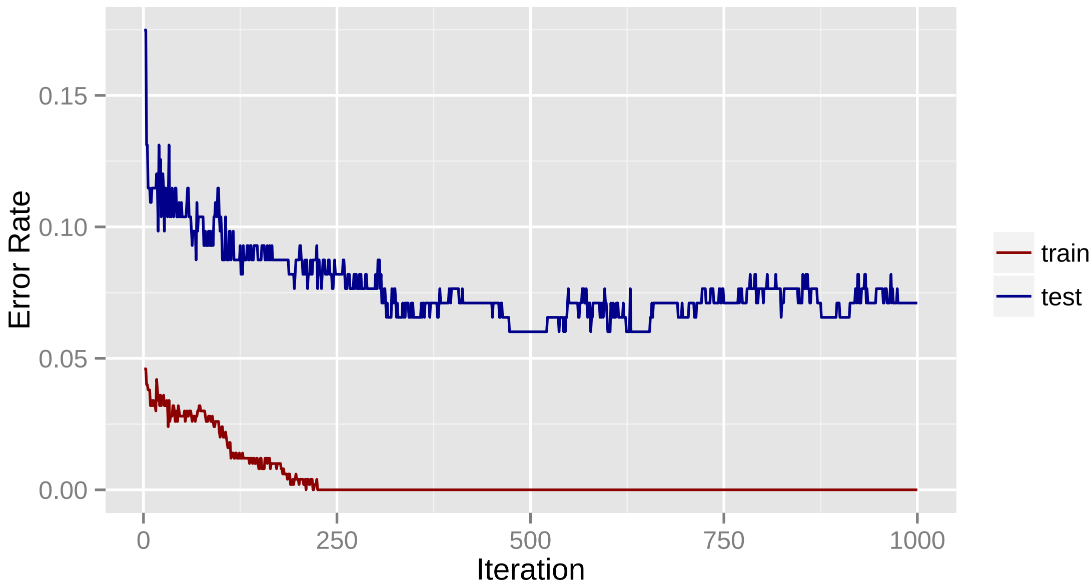
Write a function that samples discrete random variables. You will use this function to implement Step 1 of the boosting algorithm given above. The function should take in a positive integer n and a discrete, k-dimensional probability distribution w, and return a $1 \times n$ vector c, where $c_i \in 1, \ldots, k$, $\text{Prob}(c_i = j|w) = w(j)$ and the entries of c are independent. For a distribution $w = [0.1, 0.2, 0.3, 0.4]$, show the histogram of a sample vector c when $n = 100, 200, 300, 400, 500$.

```
weightedsample <- function(n,w){
  cdf <- cumsum(w)
  c <- runif(n)
  return(sapply(c, function(x) which(x < cdf)[1]))
}
```



Part 2

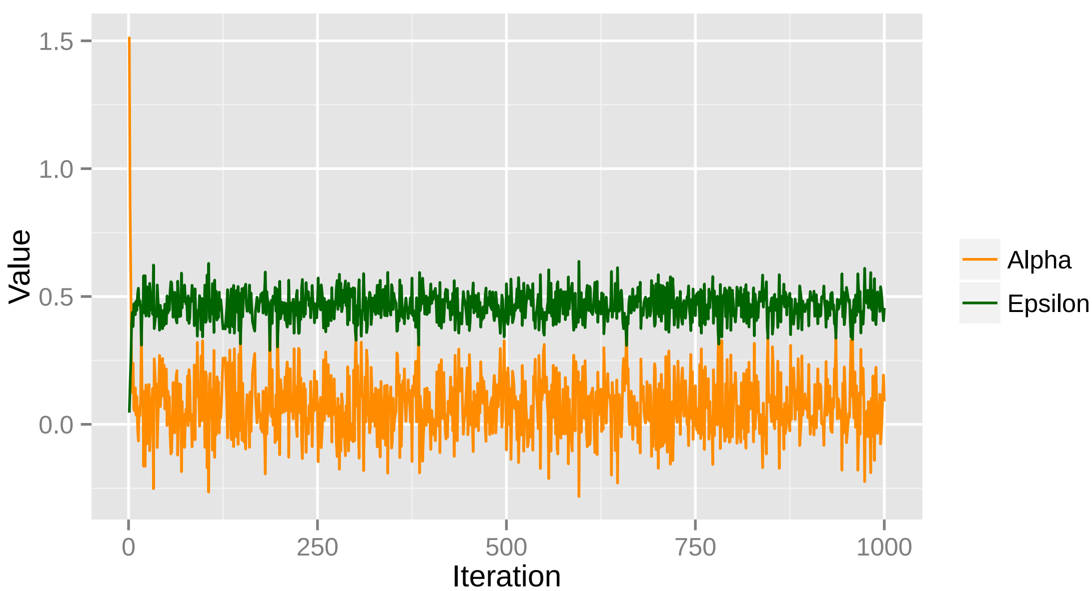
1. Implement a boosted version of this Bayes classifier, where class-specific π and μ, and shared $Sigma$ are learned on the bootstrap set B_t . Notice that you only need to store w_0 and w for this problem, as indicated in the equation above. Since the data already contains a bias dimension, you can store a single "augmented" vector where w_0 and w are combined.
2. On a single plot, show the training and testing error as a function of iteration t.



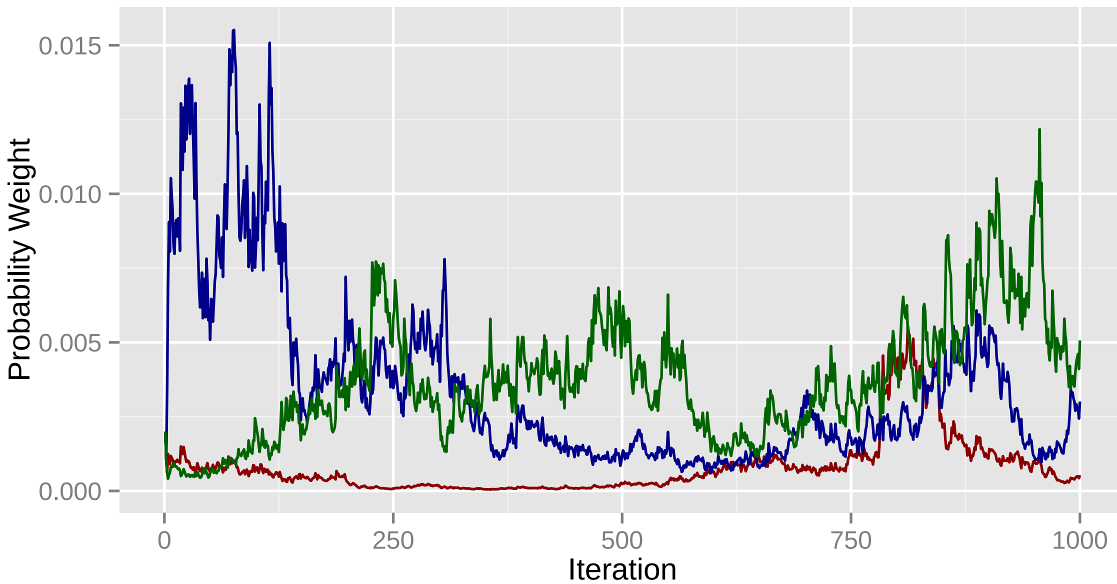
3. Indicate the testing accuracy by learning the Bayes classifier on the training set without boosting.

```
## [1] "Unboosted Bayes Classifier Accuracy: 84.15"
```

4. Plot α_t and ϵ_t as a function of t.



5. Pick 3 data points and plot their corresponding $p_t(i)$ as a function of t. Select the points such that there is some variation in these values.



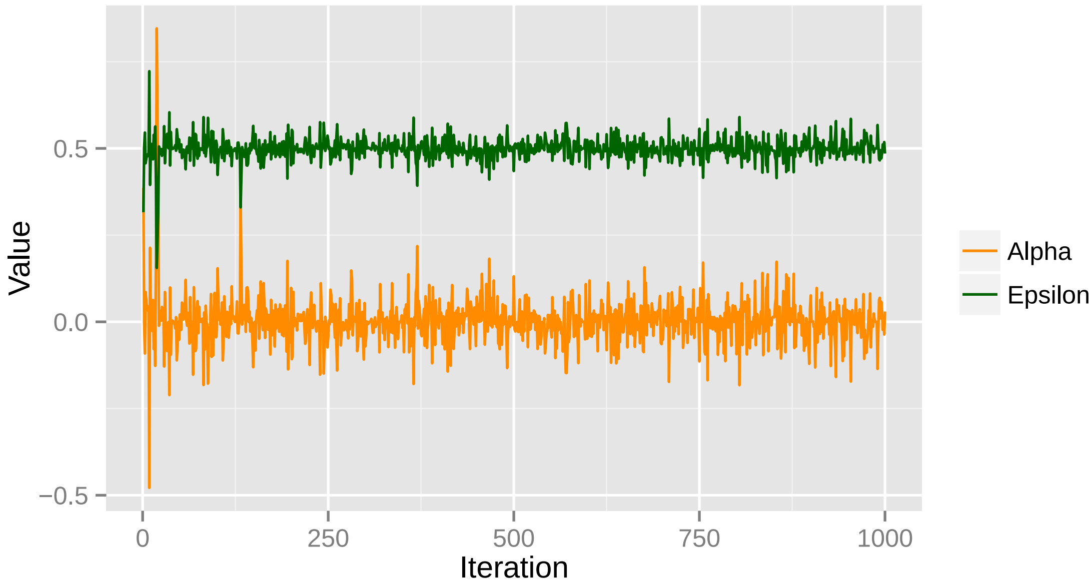
Part 3

1. Implement the online logistic classifier.
2. On a single plot, show the training and testing error as a function of iteration t.

3. Indicate the testing accuracy by learning logistic regression model on the training set **without** boosting. You can use the two-class version of your softmax logistic regression code from Homework 2 to do this, or your own implementation of binary logistic regression.

```
## [1] "Unboosted Binary Logistic Regression Accuracy: 96.17"
```

4. Plot α_t and ϵ_t as a function of t.



5. Pick 3 data points and plot their corresponding $p_t(i)$ as a function of t. Select the points such that there is some variation in these values.