Machine Learning
CAP 6610
Homework 2
Due: Monday, Feb. 5

(1) You are given 5 text files:

T1.txt, T2.txt, T3.txt, T4.txt, T5.txt.

The words in T1 represent documents of from Class 1 and the words in T2 represent documents from Class 2.

The words in 2 of the 3 files T3, T4, and T5 represent Class1 and Class2 but we don't know which. The other file contains words from an unknown class.

Rank the features according to their mutual information with the classes. Build a Naïve Bayes Classifier to classify the documents using the most informative features.

(2) Generalize the formula discussed in exercise 3.21 to include multinomials.

(a) Try your formula on the data set:

```
X      = [1,2,2,3,2,1,1,1,1,1,2,2,2,3,3,3,3,3,3,3];
Class1 = [1,1,1,1,1,2,3,3,3,3];
Class2 = [1,2,2,2,2,2,3,3,3 3];
```

(b) Use the formulas to rank the quantized, 60-dimensional Bar Features (BFs) calculated on MNIST digits according to mutual information. The BFs are described in:

P. D. Gader, M. Mohamed, and J. Chiang, "Comparison of Crisp and Fuzzy Character Neural Networks in Handwritten Word Recognition," *IEEE Trans. Fuzzy Systems*, Vol. 3, No. 3, pp. 357-364, August 1995.

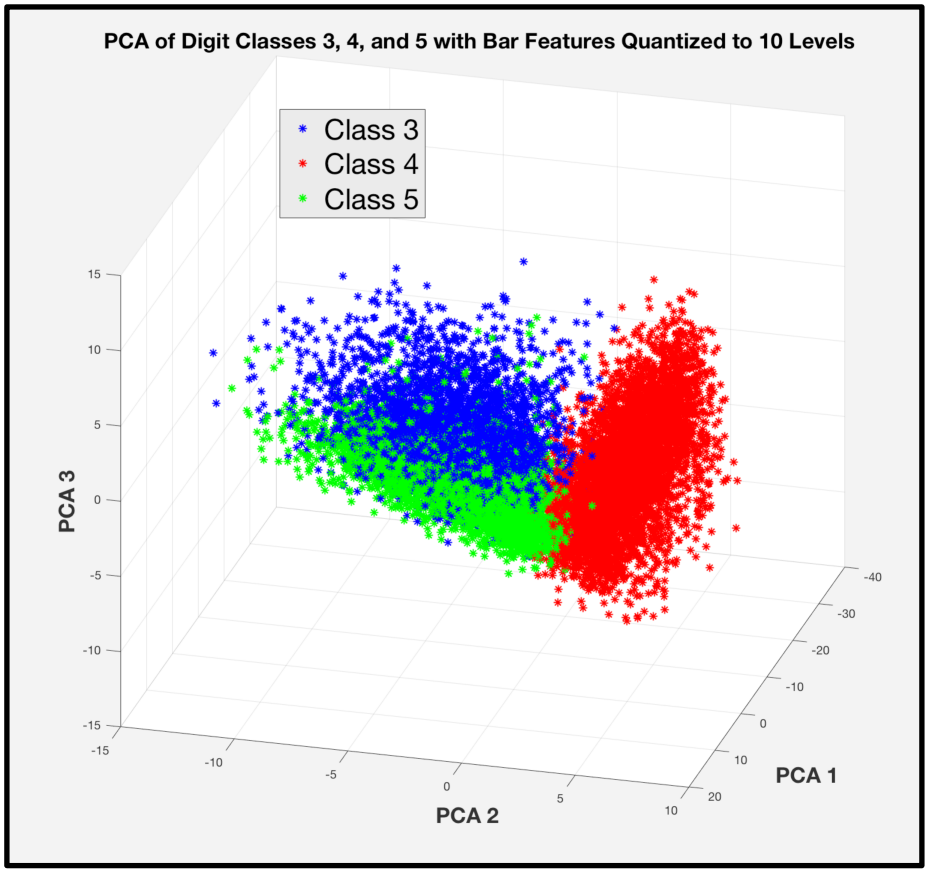
which is in the file: **`CrispFuzzyNeuralNetworks.pdf`**

Use the most informative features to classify digits in classes '3', '4', and '5'. The unquantized & quantized BFs from the classes are stored in `60 x 15000` matrices are stored in the files **`BF345.mat`** and **`BF345Quant10.mat`** .

Each column is one feature vector. They are stored as follows:

```
Digit Class 3:   Columns     1- 5000
Digit Class 4:   Columns  5001-10000
Digit Class 5:   Columns 10001-15000.
```

The first 3 components of the PCA of the quantized BFs are depicted in Figure 2.1.



(3) Write a program that displays the posterior of the Beta-Binomial given different sample sizes and initial Binomial probabilities. Do the same for the Dirichlet-Multinomial model.

(4) Run the program `naiveBayesBowDemo` found on the textbook website.