Machine Learning with Networks *ECEN765.600, Fall 2017*

Programming Assignment #1: Naive Bayes & K-Nearest Neighbor

Due date: October 13th, 2017 (Friday)

"On my honor, as an Aggie, I have neither given nor received unauthorized aid on this academic work."

Signature:

Name:

0. Reading Assignments: The relevant content in BRML Chapters 10, 14 & 15

1. Implement the naive Bayes classifier (NB) and k-nearest-neighbors (KNN) for digit recognition:

1. (20pts) Get familiar with the data set on eCampus, which was originally provided in "Machine Learning in Action" (`digits.zip`, in which it contains two folders with training and testing images. Each image has been transformed into texts with '0' and '1' representing the original foreground and background of digit images. From the training data, pick one sample for each digit, display them as binary images to visualize how they look like. Convert all the data (training and testing) to vectors so that they can be used later on.
2. (20pts) Implement naive Bayes classifier for digit recognition: Use the training data to learn conditional probabilities. Compute both training and testing error rates.
3. (40pts) Implement k-nearest-neighbor classifier for digit recognition: Use the data in the training folder as training samples, for which the actual digit can be obtained by the first character of the file names. Evaluate the testing error rates of k-nearest-neighbors using the data in the testing folder with varying k from 1 to 10. Plot the testing error rates with respect to k. For this dataset, which k works the best? Plot the training error rates using the data in the training folder for $k = 1$ to 10? How does it look different from the testing error rate plot? Can you briefly discuss the trends of two plots?

   Plot the testing error rates for model averaging of KNNs from $k = 1$ to 10. Discuss the differences from the above testing error rates.

   Compare them with the corresponding error rates from NBC. Briefly discuss pros and cons of KNN and NBC. Discuss possible ways to improve the performance of NBC.
4. (20pts) Apply principal component analysis (PCA) to the original data (both training and testing). Redo the previous KNN and NB training and evaluation. Compute the error rates correspondingly and discuss the advantages or disadvantages of PCA for this specfic application of digit recognition.

Note: If you are using **pmtk3** for this assignment, the relevant demos that you can refer include: *mnistKNNdemo, naiveBayesMnistSample, pcaDigitsVis10classes.*

2. Course Project

1. Please turn in your programming assignment together with your project proposal. Please check the course project guideline if needed.