60-473/574 Advanced AI/Pattern Recognition - Fall 2017
Assignment 1

Deadline: **October 6, 2017, at 11:59pm**

This assignment must be done **individually**. The main goal of this assignment is that students obtain hands-on experience in using tools for Naïve Bayes and *k*-Nearest Neighbor classifiers. For Naïve Bayes and *k*-NN you will use Weka or Scikit.
You will work with the 4 datasets posted on Resources (clusterincluster, halfkernel, twogaussians and twosprials). Each dataset contains two-dimensional samples that belong to one of the two classes: 1 or 2. The class labels (1 or 2) for each sample are located in the last column. The first and second columns contain the coordinates of the 2D points that represent the samples.

1. Using Weka or Scikit, run the following classifiers on all 4 datasets:
   a) *k*-NN with the Euclidean distance function, where $k = 1$. When do you think we need to consider a tie-resolution scheme for *k*-NN? Give clear reasons for it.
   b) The Naïve Bayes classifier that uses normal distributions, with default parameters.
   c) Specify how you have used the tool for running these classifiers. Also, provide two sample screenshots (before/after classifying) for each classifier on *one* dataset
2. Evaluate the two classification methods using 10-fold cross validation, obtaining for each classifier the five measures of efficiency studied in class: PPV, NPV, specificity, sensitivity, accuracy, where class 1 corresponds to "positive" and class 2 to "negative". Explain (in 5-6 lines) how 10-fold cross validation works.
3. For each classifier show the results obtained for the 4 datasets.
4. For the *k*-NN find the best value of *k*, and compare it with those of 1-NN and Naïve Bayes classifiers. What is the best value of *k* for each dataset? Why?
5. Plot the samples of the 4 datasets (in 4 separate plots) as points in the 2D space, using a different color and point shape for each class.
6. Briefly discuss the following for each dataset individually:
   a) Explain the behavior of each classifier. Why do you think it is good/poor?
   c) In your opinion, which classifier is the best for that particular dataset, and why?

Submit the following:

1) A report in PDF showing the items as required (this may not be needed):
   a) Explain how you ran the classifiers using Weka or Scikit.
   b) Classification measures (accuracy and others) for each classifier.
   c) Discussions and comparisons as required.
2) Provide the source code used to run the classifiers or if used the Weka GUI, two screenshots that show how you ran the classifiers in Weka.

Note: Missing explanations or results of your implementations/experiments will imply marks deducted!

Upload *a single* Zip file that includes all the files to the Blackboard system no later than the date/time specified as deadline. Late submissions will be penalized with 10% per day for up to 3 days – after the 3rd day the mark will be **zero**.