COMP-4730/8740 Advanced AI/Machine Learning and Pattern Recognition - Fall 2019

Assignment 1

Deadline: **October 4, 2019, at 11:59pm**

This assignment must be done **individually**. The main goal of this assignment is that students understand how supervised classification techniques work. You will use the machine learning Python API, Scikit-learn.
You will work with the sample datasets posted on the BB tab called Resources. Each dataset contains two-dimensional samples that belong to one of the two classes: 0 or 1. The class labels (0 or 1) 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. Download the following datasets: circles0.3, moons1, spiral1, twogaussians33, twogaussians42, and halfkernel.
2. Run the following classifiers on all downloaded datasets:
    a. LDA (linear discriminant analysis)
    b. Quadratic (quadratic discriminant analysis)
    c. Naïve Bayes: Choose a distribution and explain why you chose it.
    d. SVM: Choose a kernel and explain why you chose it.

    For each classifier, provide the following five performance measures: PPV, NPV, specificity, sensitivity and accuracy.
3. Plot the samples of the all datasets (in separate plots) as points in the 2D space, using a different color and point shape for each class.
4. Briefly discuss the performance (accuracy only) for each classifier and dataset individually: Why do you think it is good/poor?

COMP-8740 only:

5. For each classifier: which classifier is the best for that particular dataset, why? Give very strong reasons, including mathematical formulas and/or arguments.
6. Explain mathematically (sketch of proof) how Naïve Bayes that uses Gaussian distributions is a simplified form of the quadratic (optimal Bayesian) classifier.

Submit the following:

1) A report in PDF showing the items as required:
    a) Your name and SID.
    b) Explain how you ran the classifiers on Scikit-learn (one or two paragraphs).
    c) Classification measures (accuracy and others) for each classifier.
    d) Discussions and comparisons as required.
2) Provide the source code used to run the classifiers.

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

Notes:

- Upload *a single* Zip file that includes all the files to the Blackboard system no later than the date/time specified as the deadline.
- No support will be given after **October 4 at 5pm**. The assignment can be submitted up to **two** days without incurring any penalty, but only at your own risk.
- Submissions after **October 6** will be penalized with 10% per day for up to 3 days – after the 3rd day, the mark will be **zero**.
- Missing explanations about your implementations or questions will imply marks deducted.