

# CARLETON UNIVERSITY

## SCHOOL OF COMPUTER SCIENCE

### WINTER 2020

**COMP 5107**

**Project**

**Due April 17, 2020**

This is the only project for your course. It involves the pattern recognition of real life data. In one sense, it is almost exactly a repetition of what you have done for the assignments, but for real life data.

**Important:** Before March 19, 2020, each of you must report the data set you intend to use so that no two students use the same.

#### **The Classes**

The problem involves PR on samples from a number of real-life problems that are stored at <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

The classification problem is this: You are to take the data from any one data file stored at this site. From this data set, you are to extract a *maximum* of 6 features. Please do not consider data sets where there are less than 6 features. Also, if the PR involves multiple classes, concentrate on any two classes.

These features are to be used in the training and testing of the classifiers below.

#### **Classifiers: Training and Testing**

1. You must test the system with at least two traditional classifiers:
  - a. Quadratic, and
  - b. A Nearest Neighbor Method.
2. You must test the system with at least two *linear* classifiers:
  - a. The Ho-Kashyap Rule
  - b. Fisher's Discriminant Method
3. You must verify the effect for diagonalized and non-diagonalized versions of the data.
4. You must do your training using a maximum-likelihood and a Bayesian method.
5. You must also do your training using a Parzen Window method for a single feature. *But you do not need to do the testing using this.*
6. In each case, plot the classifier for any two-dimensional projection of the data.
7. All your testing must be done with a five-fold cross validation.

#### **Report**

Submit a *short* (maximum of 5 pages) well-written report describing each module of your project.

1. Explain *how* you have tested the quality of your classifiers.
2. Also comment on the accuracy of your classification, the reason why it is (or is not) good.
3. You must explain how you can extend your system to be a complete statistical pattern recognition system.

#### **Remarks**

I am not interested in the actual accuracy of your classifier. I am more interested in the fact that you actually implement them.