ENPM 808R Exam 1

Total of 100 points.
Questions 1-6 are 10 points each (a, b are 5 points each).
Question 7 is 40 points (a, b are 20 points each).
Exam due on the ***Vernal Equinox, March 20th, at 11:59pm.***

1) What are the three primary types of machine learning, i.e., what type of learning do the classifiers do?
    a. Describe the three types **concisely**.
    b. Give an example of an application for two of those types.
2) Give a **short** description of Adaline and perceptron models.
    a. What are the primary components of each?
    b. Compare and contrast the two models.
3) Give a **short** description of one method for extending a binary classifier to many classes.
    a. What primary algebraic component is used to represent features in data?
    b. How is this algebraic component used in other classifiers such as SVMs?
4) Briefly describe what is an SVM?
    a. What are the key concepts?
    b. How does it operate in higher dimensions - what common method does it use?
5) Compare briefly Adaline and Logistic regression
    a. When would you use one over the other method?
    b. What single step or "exchange" would convert Adaline into logistic regression?
6) What is overfitting and underfitting?
    a. What could you do to each to avoid these situations?
    b. Are linear or nonlinear separable data classes more prone to these?
7) Credit card fraud detection using machine learning follows the same approach that you would use to detect malware with ML. Each transaction is analogous to an executable file, where most are benign, but a few are malicious.

The creditcard.csv.zip file contains anonymized data of actual credit card transactions collected in an open-source community. For each transaction, there are 30 categories, 28 of which are principal components. The last category is a 1 or 0, and represents whether the transaction is fraudulent or valid, while the second to last category is the amount of each transaction in dollars. Since most transactions are valid and the fraudulent are outliers, i.e., there are only a small number of them, this justifies the need for a large data set of over 284,000 samples.

An example of a classifier being trained and applied to detection of fraudulent transactions is given by the link in the QR code attached to the bottom of this exam, and in the .png file: *"qrcode_www.kaggle.com"* in the same directory as this test and the *.csv fraud data file*. Please take a look at how this is implemented, including dividing training and classification data similar to the IRIS data. The 28 features are also part of this just to see if there are any noticeable trends or "fingerprints" within the transactions.

a) You will use parts of Lab1a to classify the provided credit card data as either fraudulent or valid. You may use either ADALINE or the SVM. Keep in mind that you will need to replace the IRIS dataset with the credit card dataset described above. Show the resulting graph of fraudulent and valid transactions separated out.

b) You will use parts of Lab1b to classify the provided credit card data into clusters that define each of the feature categories. You will justify your choice of clustering algorithm with a sentence or two that describes why it is better suited than the other clustering approaches, what methods you use to measure similarity and distance. Show graphically, and also show if there is any relationship between some clusters and fraudulent transactions.

Hint: you may want to start out by trimming down the dataset temporarily: this means that you select 1 out of each 100 transactions for training and classification, just to test your code. If it seems ok, then you should also be able to estimate just how long (e.g., 100 times as long if initial training was on 1 of 100) it will take you to run the entire set. Keep in mind that a small sample set may not have sufficient data to define all the features that you need for the correct results.