Assignment 1
Due date: October 1st, 2019

COMP 5413-FC/FD

Rules

Please refrain from using any unnecessary libraries, toolkits or other pices of software to implement to algorithms required. Implementations should not use high level functions from from pre-implemented classifiers or frameworks.

If you decide to implement the algorithms using Matlab make sure they can be executed in Octave.

Create a PDF file for your graphs and results. Add the PDF and source code files with instructions to reproduce your results to a .zip file and upload to the system.

–

Problem 1 *(0.5pt) What is the main disadivantage when applying kNN to large datasets or datasets involving a large number of features?*

Problem 2 *(0.5pt) Condensing is a technique to reduce the complexity of the kNN by reducing which term of the $O(nd)$ complexity?*

Problem 3 *(2pts) Implement the exact solution knn for the* ***dataset 1****. Find the best k using 5 fold cross validation. Don't forget to shuffle your dataset. Create a table with results for k from 1 to 11, for the validation fold and test set.*

There are two folders in your dataset 1: training_validation should be used to find the best k; test should be used to test the classifier after your find the best k using the training_validation set.

Learning outcomes:
Author(s): Thiago E A Oliveira

Problem 4 *(0.5pt) Given the following loss matrix:*

$$\Lambda = \begin{bmatrix} 0 & 20 \\ 10 & 0 \end{bmatrix} \quad (1)$$

Assume that the rows of the matrix represent actions of treating a patient with radiotherapy (row 1) or medication (row 2). The columns of the matrix represent the classes of patients, column 1 indicates that the patient may have cancer, column 2 indicate that the patient has a pulmonar infection. Assume that the probability of a patient x having cancer is 0.6 ($P(y_1|x) = 0.6$), what is the risk of applying each one of the possible treatments?

Problem 5 *(2pts) Implement a naive bayes classifier for newsgroup messages for the* ***dataset 2****. The classes in the dataset are in the folders "sic.electronics" and "comp.sys.ibm.pc.hardware". To implement the classifier you will need to implement a tokeninzer to break the messages into tokens (in this case you may use a tokenizer). Since we are only interested in words, make sure your tokenizer filters out special characters, symbols and other undesired features. Your preprocessing step should also remove the header and signatures from the meassages. Use the test set to display the final performance of the classifier trainned with the messages from the training set.*

Problem 6 *(2pts) Implement gradient descent for the following functions:*

1. $f(\beta) = \beta^2$, *with* $\beta \in [-5, 5]$ *and* $\beta_{initial} = 4.5$
2. $f(\beta) = \dfrac{sin(10\pi\beta)}{2\beta} + (\beta - 1)^4$ *with* $\beta \in [0.5, 2.5]$ *and* $\beta_{initial} = 2.4$

Initialize beta with the values $\beta_{initial}$. Plot the graphs displaying $f(\beta)$ and the points for β and $f(\beta)$ found during the descent. What is the effect of the learning rate in both problems?

Problem 7 *(0.5pt) Why applying linear regression to a classification problem may not be adequate? What the the effect of outliers in this type of classifier?*

Problem 8 *(2pts) Implement a classifier based on logistic regression for digits 0 and 6 of* ***dataset 1****. Use the test set to display the results of your classifier.*