Machine Learning - Assignment 2 (Due: 25-Oct-2024)

This assignment consists of 4 parts. Undergraduate students should complete parts 1, 2 and 3, while graduate students should complete all 4 parts. In addition, the assignment should be completed in Python 3, and individually.

If you are an undergraduate student, please submit your python script (.py file) on D2L. Please note that if you submit your file in some other format besides .py, then your mark will at most be 60%. If you are a graduate student, please also include your redacted document (PDF file) in your submission D2L.

For this second assignment, several blanks (_____) have again been included to guide you. However, in no way these blanks are representative of the number of commands, parameters, length of what to input, or anything else. You may write as little or as much code as what is required to achieve the goals.

PART 1: basic linear regression
The goal is to predict the profit of a restaurant, based on the number of habitants where the restaurant is located. The chain already has several restaurants in different cities. Your goal is to model the relationship between the profit and the populations from the cities where they are located.

It is recommended you open the csv file in Excel, notepad++ or any other applications to have a rough overview of the data at hand. You will notice that there are several instances (rows), of 2 features (columns). The values to be predicted are reported in the 2nd column.

You will be asked to load the data from the file RegressionData.csv in a pandas dataframe, then to plot the data using a scatter plot to visualize the data. Using the training data, you will train a linear regression model, and eventually, apply it to predict the profit of a restaurant located in a city of 18 habitants.

PART 2: logistic regression
You are a recruiter and your goal is to predict whether an applicant is likely to get hired or rejected. You have gathered data over the years that you intend to use as a training set. Your task is to use logistic regression to build a model that predicts whether an applicant is likely to be hired or not, based on the results of a first round of interview (which consisted of two technical questions).
The training instances consist of the two exam scores of each applicant, as well as the hiring decision.

PART 3: multi-class classification using logistic regression
Not all classification algorithms can support multi-class classification (classification tasks with more than two classes). Logistic Regression was designed for binary classification.
One approach to alleviate this shortcoming, is to split the dataset into multiple binary classification datasets and fit a binary classification model on each. Two different examples of this approach are the One-vs-Rest and One-vs-One strategies.

You are asked to explain below how the One-vs-Rest and the One-Vs-One method work for multi-class classification.

PART 4 FOR GRADUATE STUDENTS ONLY: Multi-class classification using logistic regression
Please note that the grade for parts 1, 2, and 3 counts for 70% of your total grade. The following work requires you to work on a project of your own and will account for the remaining 30% of your grade.

Choose a multi-Class Classification problem with a dataset (with a reasonable size) from one of the following sources (other sources are also possible, e.g., Kaggle):

- UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.php.

- KDD Cup challenges, http://www.kdd.org/kdd-cup.

Download the data, read the description, and use a logistic regression approach to solve a classification problem as best as you can. Investigate how the One-vs-Rest and One-vs-One methods can help with solving your problem. Write up a report of approximately 2 pages, double spaced, in which you briefly describe the dataset (e.g., the size – number of instances and number of attributes, what type of data, source), the problem, the approaches that you tried and the results. You can use any appropriate libraries.

Marking: Part 4 accounts for 30% of your final grade. In the write-up, cite the sources of your data and ideas, and use your own words to express your thoughts. If you have to use someone else's words or close to them, use quotes and a citation. The citation is a number in brackets (like [1]) that refers to a similar number in the references section at the end of your paper or in a footnote, where the source is given as an author, title, URL or journal/conference/book reference. Grammar is important.

Submit the python script (.py file(s)) with your redacted document (PDF file) on the D2L site. If the dataset is not in the public domain, you also need to submit the data file.
Name your documents appropriately:
report_Firstname_LastName.pdf
script_ Firstname_LastName.py

Ensure you did not miss any step from the .py script. Here is how Parts 1, 2 and 3 are graded:

Line #	Points
27	5
29	5
30	5
32	5
35	5
36	5
40	5
41	5
42	5
48	8
52	8
67	2
70	2
71	2
79	2
83	2
84	2
88	2
95	2
110	12
115	11