CS3390/CS5590/AI5000 Foundations of Machine Learning
Assignment 1

Marks : 120, Submission Deadline: Oct 8, 2023

Instructions:

- The submission should be a single ZIP file with the name **ROLL_NUMBER-foml-assignment-1.zip**
- Datasets used should be present in the same zip file
- Before submitting, students should ensure that the code works and there are no issues with file paths of the dataset
- Every programming question should be in a separate ipynb notebook named by question number
- Subjective questions should be in PDF format (either picture of handwritten work or typed up). If it is handwritten, please ensure the image is clear and readable. We suggest keeping each answer in a separate file named by question number

1. Please collect data on how long you will take to finish the food in the mess (including waiting time) for breakfast, lunch, tea, dinner etc. for next 2 weeks. Prepare the data by considering various features like day, time, holiday, category (breakfast, lunch…), and other features that you think could be helpful for accurate prediction of time you will take to finish the food.
    a. Prepare the data and perform an exploratory data analysis, prepare a plot of the time taken against different days for each of meal type.
    b. Due to the nature of the output, we know that linear regression may not be suitable for our problem. Suggest a model to fit the data and learn parameters using Maximum Likelihood estimation using 80% of the data. Code the model and learn the parameters from the data. Analyze the parameter values and figure which parameter is important for prediction.
    c. Use the learned model to do prediction. For prediction, you may consider 20% of the collected data. Provide an appropriate evaluation metric and compare the results of the prediction of your model with a linear regression model (you may use existing packages like scikit-learn for linear regression).

    Provide explanations on the choice of your features and model. Code has to be in Python with sparse use of packages. Provide a Readme file explaining how to run your code and with proper comments.
    Marks : 10 (a) + 15 (b) + 5 (c) = 30 + 10 (readable code) = 40

2. Read Regression Models for Ordinal Data by Peter McCullagh (please find attached).
    a. Provide a brief summary of the paper. Explain how the likelihood and odds ratio for ordinal regression is different from multi-class classification. How is it different from the regression problems?

b. Explain and derive the parameter estimation technique for the ordinal likelihood described in the paper.
c. Consider the WINE data set (https://archive.ics.uci.edu/dataset/186/wine+quality). Develop an ordinal regression model in Python to predict the rating of the wine based on the input features. You may consider 20% of the data for testing purpose. Also, propose an appropriate evaluation metric and compare the result with a linear regression model. You may use existing ordinal regression and linear regression packages.

Marks : 7.5 (a) + 7.5 (b) + 15 (c) = 30 + 10 (readable code) = 40

3. The method of ordinary least squares assumes that there is constant variance in the errors (which is called homoscedasticity). The method of weighted least squares can be used when the ordinary least squares assumption of constant variance in the errors is violated (which is called heteroscedasticity).
   a. Derive the expression of likelihood and prior for a heteroscedastic setting for a single data point with input x_n and output t_n.
   b. Provide the expression for the objective function that you will consider for the ML and MAP estimation of the parameters considering a data set of size N.
   c. Show that the ML objective will result in a data set in which each data point t_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n \left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2.$$

   Find an expression for the solution w that minimizes this error function.

Marks : 5 (a) + 10 (b) + 5 (c) = 20

4. For logistic regression, there is no longer a closed-form solution, due to the nonlinearity of the logistic sigmoid function. the error function can be minimized by an efficient iterative technique based on the Newton-Raphson iterative optimization scheme
   a. Provide the expressions of the gradient, Hessian, and update equations for the Newton-Raphson optimization technique used to obtain the parameters in the logistic regression model. Provide an algorithm describing the methodology.
   b. Show that the Newton-Raphson update scheme is related to the weighted least squares problem described in question 3 (c) and explain why it is called the iterative reweighted least squares method.
   c. Show that the error function of the logistic regression is a concave function of w and hence has a unique minimum with the help of the Hessian matrix.

Marks : 10 (a) + 5 (b) + 5 (c) = 20