UCLA CS 145 Homework #1

DUE DATE: Thursday 10/19/2017 11:59 pm

Note:

- You are expected to submit both a report and code. For your code, please include clear README files. When not specifically mentioned, use the default data provided in our program package.
- "`########## Please Fill Missing Lines Here ##########`" is used where input from you is needed.

1. Linear Regression

1.1 In the table below, we have Height (in inches) and Weight (in pounds) for 5 students in a class. We want to build a linear regression model to use Height (x) to predict Weight (y).

Student ID	Height	Weight
1	60	130
2	70	155
3	62	125
4	72	162
5	65	150

(a) Write down the linear regression model, and calculate the weights β using the closed-form solution for ordinary least square method. Please clearly show the steps involved in your calculation.

(b) Calculate the predicted Weight for each student.

1.2 In LinearRegression\linearRegression.py, fill in the missing lines in the python code to estimate MSE, and β using (1) closed-form solution, (2) batch gradient descent, and (3) stochastic gradient descent.

(a) Report weights β and MSE (Mean Square Error) in the test dataset for each version, are they the same and why?

(b) Apply z-score normalization (https://en.wikipedia.org/wiki/Standard_score) for each feature x (i.e. use μ and σ of each feature/column to normalize the corresponding feature/column) by filling the missing lines of code, and report whether the normalization affect the weights β and MSE (Mean Square Error) in the test dataset, for all the three versions of the algorithm, and why?

2. Logistic Regression and Model Selection

Much of the equations discussed in class are given in the form of vectors and matrices. This is done for computation advantages, and compactness. But we don't need them. For the following exercises (except for the Hessian matrix in part C and all of part D) please write out the full equations without using vectors, matrices, or summations.

We are given a data set consisting of the following experiment. The height and weight of 3 people were recorded at the beginning of each person's 65th birthday. At exactly one year after each person's 65th birthday the vital status was recorded to be either alive or deceased. Our end goal is to use logistic regression to predict the probability that a person's life expectancy is at least 66 years given their age of 65, initial vital status of alive, height, and weight (but we won't go that far here). The data is given in the following table:

Height (inches)	Weight (lbs)	Vital Status
60	155	Deceased
64	135	Alive
73	170	Alive

(a). State the log-likelihood function

(b). State the gradients for each parameter

(c). Give the Hessian Matrix

(d). Assuming an initial guess of 0.25 for each parameter, write python code for finding the values of the parameters after 2 iterations using the Newton-Raphson method.

3. Decision Tree

3.1. Construct a decision tree for samples from the congressional voting records dataset with the first three attributes in the UCI machine learning repository. Information gain is used to select the attributes. Please write down the major steps in the construction process, i.e., you need to show the information gain for each candidate attribute when a new node is created in the tree.

Class	Vote for handicapped-infants?	Vote for water-project-cost-sharing	Vote for budget-resolution-adoption
Democrat	Y	N	Y
Republican	N	Y	N
Democrat	Y	Y	Y
Republican	N	Y	N
Democrat	N	Y	N
Democrat	N	Y	Y
Democrat	Y	N	Y
Democrat	Y	Y	Y
Republican	N	Y	Y
Republican	Y	Y	N
Democrat	N	N	Y
Republican	N	Y	N
Republican	N	N	N
Democrat	N	N	Y
Republican	N	N	N
Republican	N	Y	N
Democrat	Y	N	Y
Democrat	Y	N	Y
Republican	N	N	N
Republican	Y	N	Y

3.2 In DecisionTree\DecisionTree.py, fill in the missing lines for building a decision tree model.

(a) Find the datasets in the files “tic-tac-toe.data” and “house-votes-84.data”, and calculate the average accuracy based on 5-fold cross-validation. To handle missing values in the dataset, “?” is treated as a separate value in the dataset instead of missing values. More detailed information about datasets are shown in https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records and https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame.

(b) Implement a gain ratio-based decision tree, and run on the same dataset using the same 5-fold cross-validation. Output the average accuracy and compare the two versions of decision tree. Which attribute selection measure do you want to choose for each dataset and why.