In this assignment, we will use the mathematical notations from the following reference books:

- **Reference books:**

  1. *An Introduction to Statistical Learning with Applications in R* **(ISLR)**. Authors: James, G., Witten, D., Hastie, T., Tibshirani, R.

  2. *The Elements of Statistical Learning Data Mining, Inference, and Prediction* **(ESL)**. Authors: Hastie, T., Tibshirani, R., Friedman, J.

1 Bias-Variance Trade-Off

(a) **(PhD: 5 points, MS: 10 points)** What does **variance** and **bias** of a statistical learning method mean? Explain in your own words.

(b) **(PhD: 5 points, MS: 10 points)** What is **bias-variance trade-off**?

(c) **(PhD: 10 points, MS: Extra credit 10 points)** Show that the expected test MSE, for a given value x_0 can be decomposed into the sum of three quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$, and the variance of the error ϵ. [Equation 2.7, page 34, **ISLR** book]

2 Linear Regression

(**PhD: 15 points, MS: 15 points**) Please give short (1-3 sentence) answers. You may discuss these questions with others, but write the answers in your own words.

(a) Does linear regression assume uncertainty in the measurement of the independent variables ($\mathbf{X}$), the dependent variable ($\mathbf{y}$), or both?

(b) Linear regression is sensitive to outliers. Briefly describe one solution to make it more robust to outliers.

(c) It is often assumed that the magnitude of the regression coefficients (abs(β)) indicate the importance of the corresponding features. Describe a situation where this may not be true.

(d) What problem arises if one independent variable is a perfect linear combination of other independent variables?

(e) Suppose you are using logistic regression for binary classification, predicting the positive class whenever the posterior probability is greater than 0.5. Would you get the same predictive accuracy by running linear regression instead, predicting the positive class whenever $\hat{y}$ is greater than 0.5? Whether you answer yes or no, describe an advantage of using logistic regression over linear regression for binary classification.

3 K-Nearest Neighbors

(**PhD: 25 points, MS: 25 points**)
Suppose we know the locations (co-ordinates) and the disease types of 10 patients.
Disease type 1: $\{(10, 49), (-12, 38), (-9, 47)\}$
Disease type 2: $\{(29, 19), (32, 31), (37, 38)\}$
Disease type 3: $\{(8, 9), (30, -28), (-18, -19), (-21, 12)\}$

- (**5 points**) Normalize the data (follow the formula mentioned in the lecture).
- (**20 points**) Suppose we have a new patient whose location is $(9, 18)$. Using K-Nearest Neighbor with L_2 distance metric, predict the new patient's disease type if we are using $K = 1$ (**4 points**) and if we are using $K = 3$ (**4 points**). Similarly, what are the disease predictions for the patient if we use L_1 distance metric with $K = 1$ (**4 points**) and with $K = 3$ (**4 points**). For $K = 3$, if there is a tie, please choose the label of the data point with closer distance. Please compare the results between these 4 different predictions (**4 points**). Do not forget to normalize/standardize the coordinate of the new patient using the mean and standard deviation of the known patients. Provide intermediate computations of how do you arrive at your predictions.

4 Ridge Regression

(PhD Extra credit: 20 points, MS Extra credit: 20 points)
Ridge regression shrinks the regression coefficients by imposing a l_2 penalty. The ridge coefficients minimize a penalized residual sum of squares, [Equation 3.41, page 82, **ESL** book]

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \Big(\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \Big)$$

- Prove $\hat{\beta}^{ridge} = (\mathbf{X^T X} + \lambda \mathbf{I})^{-1} \mathbf{X^T y}$. **(10 points)**
- Derive the relation between ridge regression estimator and the classical ordinary least squares estimator. List any assumptions you make. **(10 points)**

5 Programming

5.1 Data Analytics

(PhD : 25 points, MS: 25 points)
Please write programs for the exercise problem 10 of the chapter 2, **ISLR** book (page 56).

5.2 Classification

(PhD : 65 points, MS: 65 points)
In this assignment, you will experiment with three classification algorithms on real-world datasets. Below, we describe the steps that you need to take to accomplish this programming assignment.

Dataset: We will use the *Tic-Tac-Toe Endgame Dataset* (for all classification algorithms) and the *Nursery Dataset* (ONLY for Naive Bayes) from UCI's machine learning data repository. The training/validation/test sets are provided along with the assignment in Blackboard as `hw1ttt_train.data`, `hw1ttt_valid.data`, and `hw1ttt_test.data` for the *Tic-Tac-Toe Endgame Dataset*, and `hw1nursery_train.data`, `hw1nursery_valid.data`, and `hw1nursery_test.data` for the *Nursery Dataset*. For description of the datasets, please refer to https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame and https://archive.ics.uci.edu/ml/datasets/Nursery.

Please follow the steps below:

(a) Data Pre-processing (10 points) The first step in every data analysis experiment is to inspect the datasets and make sure that the data has the appropriate format. You will find that the features in the provided dataset are categorical. However, some of the algorithms require the features to be real-valued numbers. To convert a categorical feature with K categories to real-valued number, you can create K new binary features. The ith binary feature indicates whether the original feature belongs to the ith category or not. For example, if you have a feature called 'height' with possible values 'low', 'medium', and 'high', then you will have three (3) binary features: if in a single data instance, you have a value of 'low', then your binary features will be 1-0-0; if value is 'medium', then your binary features will be 0-1-0; if value is 'high', then your binary features will be 0-0-1. About the training labels, if there are M labels in the dataset, then convert each of them into a number between 1,...,M. For example if there are 3 possible labels: 'positive', 'neutral', and 'negative', then 'positive' label will be converted to 1, 'neutral' label will be converted to 2, and 'negative' label will be converted to 3. So, for a training dataset (`*_train.data`) of size N, then `train_data` will be a matrix of size $N \times D$ (with D is the total number of binary features, as a result of this pre-processing) and `train_label` will be a column vector of dimension $N \times 1$. This also applies to validation dataset (`*_valid.data`) and testing dataset (`*_test.data`). This pre-processed matrices and vectors will be an input to the classification algorithms.

(b) Performance Comparison (35 points) Compare the three algorithms (K-NN, Naive Bayes, and Logistic Regression) on the provided dataset.

***K*-NN:** Consider $K = 1, 3, 5, \cdots, 15$. For each K, report the training, validation and test accuracy. When computing the training accuracy of K-NN, we use leave-one-out strategy, i.e. classifying each training point using the remaining training points. Note that we use this strategy only for K-NN in this assignment. Operate ONLY on the *Tic-Tac-Toe Endgame Dataset*.

Logistic Regression: Report the training, validation and test accuracy for different parameter settings of logistic regression. Vary penalty and solver parameters only. Operate ONLY on the *Tic-Tac-Toe Endgame Dataset*.

Naive Bayes: Report and compare the training, validation, and test accuracy, both on the *Tic-Tac-Toe Endgame Dataset* and the *Nursery Dataset*. If there is a significant difference between the two datasets' classification accuracy, could you guess why?

(c) Decision Boundary (20 points) In this step, you need to apply K-NN on the `hw1boundary` dataset (provided in Blackboard as `hw1boundary_features.csv` and `hw1boundary_labels.csv`

files) which is a binary classification dataset with only two features. You need to run K-NN with $K = 1, 5, 15, 25$ and examine the decision boundary. A simple way to visualize the decision boundary, is to draw 10000 data points on a uniform 100×100 grid in the square $(x, y) \in [0, 1] \times [0, 1]$ and classify them using the K-NN classifier. Then, plot the data points with different markers corresponding to different classes. Repeat this process for all k and discuss the smoothness of the decision boundaries as K increases.

Submission Instruction: You need to provide the followings:

- Provide your answers to the problems in a PDF file, named as `IS777_hw1_fa19_YourLastName.pdf`. You need to submit the homework in electronic version as pdf file on Blackboard. If you choose handwriting instead of typing all the answers, you will get 40% points deducted.
- Submit ALL the code and report via Blackboard. Recommend to use Python programming language. For your program, you MUST include the main function called `IS777_hw1_fa19.py` in the root of your folder. After running this main file, your program should be able to generate all of the results needed for this programming assignment, either as plots or console outputs. You can have multiple files (i.e your sub-functions), however, the only requirement is that once we unzip your folder and execute your main file, your program should execute correctly. Please double-check your program before submitting. You should only submit one `.zip` file. No other formats are allowed except `.zip` file. Also, please name it as `[lastname]_[firstname]_hw1_spring19.zip`.

Collaboration: You may collaborate. However, collaboration has to be limited to discussion only and you need to write your own solution and submit separately. You also need to list with whom you have discussed.