Instructions

1. The lab must be implemented in Matlab or Python. Each question should have a readme file.

2. A readme file should precisely tell how to compile and run your program. Give the exact commands with respect to the datasets provided.

3. You should create a small video (hard limit max:12 mins), showing how you have solved the problems. A video will not be watched beyond the provided limit. The marks will be deduced accordingly.

4. You should start the video explaining the overall code like what functions it has and the overall logic. After explaining the video goes through the conceptual topics as mentioned specifically in each question.

5. You can use the software like free cam to record the video. Create different versions of the program for each variation and have the all the results ready before creating the video.

6. You should use the handbrake tool to reduce the overall size of the video. With handbreak the size of the video would be at max 50MB. That will help you and us to overcome the speed limitations.

7. Upload the video file separately with the zipped code files.

8. The marks will be given on the basis of quality of code, use of innovative data structures, scalability, correctness, and your video explanations.

9. You are supposed to submit both the videos as well as code (zipped file) on google classroom no later than 7^{th} **February 2021**. This is a strict deadline and any assignment submitted later will not be consider for evaluation.

10. Name the zip file as rollnumber-assignmentno.zip and the video file as rollnumber-assignmentno.mp4

11. You are not allowed to use the direct libraries and implement the code from the scratch.

12. Students are expected to follow the honor code of the class. We will follow a strict anti-plagiarism policy.

1 Finding the most specific/general hypothesis [Warm-UP]

Create a dataset $\mathcal{D} = (x_i, y_i)_{i=1}^N$ where each x_i is given in two dimension $x_i = (x_i^1, x_i^2)$ and each y_i is a binary label i.e. $y_i = \{0, 1\}$. First generate the vector Y where each example will take $y_i = 1$ with probability $1/2$ and $y_i = 0$ with probability $1/2$ with $N = 30$. Now fixing the y_i, sample the X matrix as follows: [20 Marks]

- If $y_i = 1$ then $x_i^1 \sim \mathcal{U}(2, 7)$ and $x_i^2 \sim \mathcal{U}(4, 6)$. where $\mathcal{U}(a, b)$ represent the uniform distribution between a, b.
- If $y_i = 0$ then $x_i^1 \sim \mathcal{U}(0, 2) \cup \mathcal{U}(7, 9)$ and $x_i^2 \sim \mathcal{U}(1, 3) \cup \mathcal{U}(6, 8)$.

Implement the following with respect to the above dataset:

1. Color code the examples with $y_i = 1$ as red and $y_i = 0$ as green and plot the dataset.
2. Write a program to find most specific and most general hypothesis when hypothesis class is considered as all possible Rectangles. Plot both the obtained hypothesis along with the dataset.
3. Write a program to find most specific and most general hypothesis when hypothesis class is considered as all possible Circles. Plot both the hypothesis obtained along with the dataset.
4. Mention any observations corresponding to second and third points.

2 Polynomial Regression in One Dimension [Easy]

In this question, we will repeat the experiments discussed in the class with respect to polynomial regression but with a different function. [30 Marks]

1. Generate 20 data points from function $f(x) = cos(2\pi x) + \frac{x}{2\pi}$ + noise where noise$\sim \mathcal{N}(0, 0.004)$ with x ranging from 0 to 2π.
2. Fit a polynomial regression with optimal weight vector w^* and plot the curves for different degree of polynomials $M = 1, 2, 3, 5, 7, 10$. Explain your observations by plotting the data points generated and the curve obtained for different values of M.
3. Repeat the previous experiments with more number of data points and report your findings.

3 Linear Ridge Regression in Multiple Dimension [Medium]

We will be understanding the concept of linear regression along with the regularization parameter on the dataset given below which has multiple attributes. Medical Dataset (https://www.kaggle.com/sudhirnl7/linear-regression-tutorial/data): The medical cost dataset comprises of independent attributes like age, sex, BMI (body mass index), children, smoker, and region. The charge/cost is a dependent feature. Our goal is to predict the individual medical costs billed by the health insurance. [50 Marks]

1. **Feature Normalization:** As discussed in the class, we first have to standardize all the features by subtracting with the mean and dividing by the standard deviation. Verify your technique by computing the mean and variance of the transformed data and check if the mean is 0 and variance is 1.

2. $K-$**Fold Cross Validation** Randomly partition the data into a training, validation, and test set. Fix 20% of the instances into the test set. For the remaining data perform the below experiments with $K-$fold cross validation. You can take the value of K to be 10.

3. **Ridge-Regression:** Here, implement your own function ridgereg(X, Y, λ) that calculates the linear least square solution with the ridge regression penalty parameter λ and return regression weights. Use gradient descent technique to find these weights. Implement predridgereg(X, weights) that returns Y given the input X with learnt weights.

4. Plot the mean square error for each of the dataset obtained from $K-$fold cross validation with respect to different λ values. Explain your finding and suggest what value of λ will you choose based on the obtained plot.

5. Plot the training error, variance and test error against different values of λ. Explain your finding and suggest what value of λ will you choose based on the obtained plot. Explain your result in the context of bias variance trade off. Does this value coincide with the previous question?