LUMS
A Not-for-Profit University

Department of Electrical Engineering
School of Science and Engineering

EE514/CS535 Machine Learning

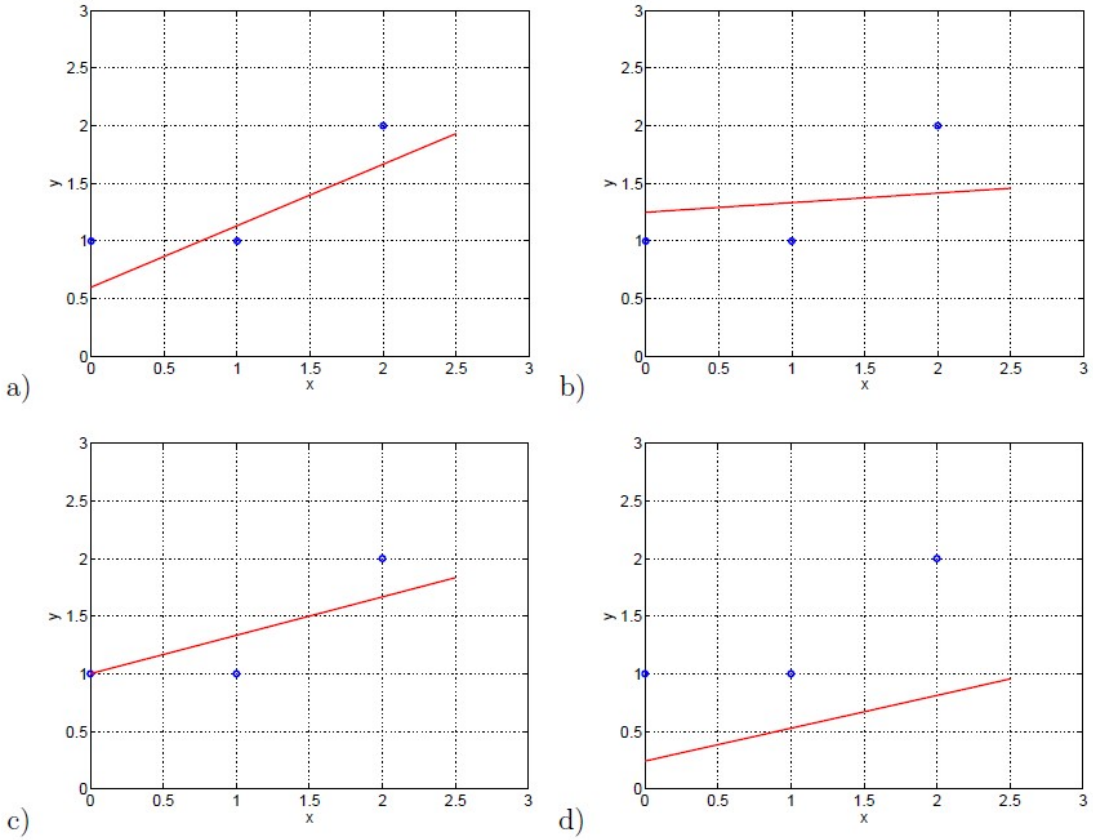
ASSIGNMENT 1

Due Date: 11 am, Tuesday, February 25, 2025.
Format: 9 problems, for a total of 100 marks
Instructions:

- You are allowed to collaborate with your peers but copying your colleague's solution is strictly prohibited. This is not a group assignment. Each student must submit his/her own assignment.
- Solve the assignment on blank A4 sheets and staple them before submitting.
- Submit in-class or in the dropbox labeled EE-514 outside the instructor's office.
- Write your name and roll no. on the first page.
- Feel free to contact the instructor or the teaching assistants if you have any concerns.

- You represent the most competent individuals in the country, do not let plagiarism come in between your learning. In case any instance of plagiarism is detected, the disciplinary case will be dealt with according to the university's rules and regulations.

Figure 1: Plots of linear regression results with different types of regularization



Problem 1 (10 marks)

Consider a linear regression model with single feature where input x_t and output y_t are related by $y_t = \theta\, x_t + \theta_0$. The plots in Figure 1 illustrate linear regression results using only three data points. Various types of regularization were applied to generate these plots (as shown below), but we are uncertain about which plot corresponds to each regularization method. Please match each plot to one, and only one, of the following regularization methods and present your calculations/explanation for each of the matches.

- $\sum_{t=1}^{3}(y_t - \theta x_t - \theta_0)^2 + \lambda(\theta^2 + \theta_0^2)$
- $\sum_{t=1}^{3}(y_t - \theta x_t - \theta_0)^2 + \lambda\theta^2$

for either $\lambda = 1$ **or** $\lambda = 10$.

Problem 2 (10 marks)

Polynomial regression is a form of regression analysis in which the relationship between the independent variable and the dependent variable is modeled as an M-th degree polynomial. The model equation that relates the input to the output is of the form:

$$y_i(x_i) = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \ \ldots \ + \theta_M x_i^M$$

(a) Assuming we are using a degree $M = 3$ for our polynomial regression model and the input data points are $x = [1, 2, 3]$:

- Write down the design matrix A filled with the data points such that the equation for the model is given by $y = A\theta$ in vectorized form.
- Explain how the structure of A changes as the polynomial degree M increases.

(b) Increasing M often leads to an increase in the risk of overfitting. Regularization is often applied to address this. You need to compute the gradient $\nabla J(\theta)$ and derive the gradient descent update rule $\theta^{(t+1)}$ for the regularized objective:

$$J(\theta) = \frac{1}{2n}\|y - A\theta\|^2 + \frac{\lambda}{2}\|\theta\|^2$$

(c) For the regularized objective function, what is the impact of λ values (large vs small) in terms of bias and variance.

Problem 3 (10 marks)

For a linear model $\mathbf{y} = \mathbf{X}\theta$, we can find the parameter vector θ given the data matrix $\mathbf{X}$ and the output vector $\mathbf{y}$ by formulating the following optimization problem

$$\text{minimize} \quad f(\theta) = \|\mathbf{X}\theta - \mathbf{y}\|_2^2 \,,$$

which minimizes the objective function $f(\theta)$. The solution can be obtained using ordinary least squares as

$$\theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Now, if the loss function is modified as

$$f(\theta) = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - x_i^T\theta)^2 + \sum_{j=1}^{M} \beta_j \theta_j^2$$

for α_i,β_j some positive constants, derive a closed-form solution to the weight vector that minimizes this loss function.

Problem 4 (10 marks)

A fitness researcher is studying the relationship between height and weight among a group of individuals to understand patterns in body composition. The researcher collects data from 10 participants, recording their height (in cm) and weight (in kg). The dataset is represented as a matrix $\mathbf{X}$, where each row corresponds to an individual, and the two columns represent height and weight, respectively.

The recorded data is as follows: Given the following matrix $\mathbf{X}$,

$$\mathbf{X} = \begin{bmatrix} 156 & 44 \\ 153 & 44 \\ 157 & 46 \\ 178 & 62 \\ 170 & 57 \\ 170 & 56 \\ 163 & 66 \\ 175 & 96 \\ 171 & 94 \\ 170 & 90 \end{bmatrix}$$

for $n = 10$ and $d = 2$.

We will make use of Principal Component Analysis (PCA) to reduce the dimensions of the matrix $\mathbf{X}$ from $d = 2$ to $d = 1$ by carrying out the following steps:

(a) Plot the data points on a 2-dimensional plane.

(b) Compute the principal components using the procedure taught in class (refer to the slides) and plot them as well.

(c) Now, project the original data matrix X onto its first principal component and plot on a 1-dimensional number line.

(d) how much of the total variance in the data is captured by the first principal component.

(e) Can we use this 1D representation to compare individuals in terms of body size more efficiently than using both height and weight, give your justification for both cases (a) when to use (b) when not to use.

Problem 5 (12 marks)

In PCA (Principal Component Analysis), the covariance matrix $\sum$ is defined as:

$$\sum = \frac{1}{n} X^T X,$$

where $X \in \mathbb{R}^{n \times d}$ is the data matrix (with n samples and d features), assumed to be mean-centered (each column has zero mean).

(a) Prove that the covariance matrix $\sum$ is positive semi-definite.

(b) Show that if $\sum$ is positive semi-definite, all eigenvalues of $\sum$ are non-negative.

Problem 6 (20 marks)

In class, we discussed Principal Component Analysis (PCA) in detail for dimensionality reduction. Here we will discuss another method for dimensionality reduction, Linear Discriminant Analysis (LDA).

LDA tries to find a linear combination of features that achieves maximum separation for samples between classes and the minimum separation of samples within each class. Here we will assume only 2 classes, but this can easily be generalized to more classes. We will use LDA to project our data onto a line.

LDA achieves this by

1. Maximizing the distance between the mean of the two classes.
2. Minimizing the scatter (variation) within each class.

Mathematically, We want to find a projection vector $\mathbf{w}$ which we can use to obtain the one-dimensional approximation (projection) of each data-point $\mathbf{x}_i$ as $z_i = \mathbf{w}^T\mathbf{x}_i$, such that the following objective function is maximized

$$J(\mathbf{w}) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2},$$

where the numerator is the difference between the **projected class means**, and the denominator is the within class scatter of the **projected samples** defined as

$$\tilde{s_i}^2 = \sum_{z \in \text{ Class}_i} (z - \tilde{\mu}_i)^2$$

Here $z = \mathbf{w}^T\mathbf{x}$ is the projected sample, and $\tilde{\mu}_i$ is the projected class mean for i-th class. In *simple words*, we want a projection such that samples of the same class are projected close to each other and the class means of the projected samples are far from each other.

The objective function formulated above can be expressed in terms of projection vector $\mathbf{w} \in \mathbf{R}^d$ as

$$J(W) = \frac{\mathbf{W}^T\mathbf{S_B}\mathbf{W}}{\mathbf{W}^T\mathbf{S_W}\mathbf{W}},$$

where

- $\mathbf{S_B}$ is the between-class scatter matrix of the samples in the original space
$$\mathbf{S_B} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$$
- $\mathbf{S_W} = \mathbf{S}_1 + \mathbf{S}_2$ is the within-class scatter matrix, where $\mathbf{S_i}$ is the covariance matrix of class i given by
$$\mathbf{S}_i = \sum_{\mathbf{x} \in \text{ Class}_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T$$
- $\boldsymbol{\mu}_i$ denotes the mean of samples for i-th class.
- $\mathbf{S_W}$ and $\mathbf{S_B}$ are symmetric and positive semi-definite.

(a) We want to determine $\mathbf{w}$ as a solution to the following optimization problem:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{maximize}} \quad J(\mathbf{w}).$$

Assuming that $\mathbf{S_W}$ is non-singular, show that the solution is the eigenvector of $\mathbf{S_W}^{-1}\mathbf{S_B}$ corresponding to the largest eigenvalue.

(b) Now we have a closed-form solution of LDA, we will implement it on a simple data set for a binary classification problem. We will be Using the same data of height and weight, from the previous problem.

i. Assign labels to data using BMI (Body mass Index) formula given below to define healthy and unhealthy individuals. You need to set the lower bound of 19 and upper bound of 30 for healthy individuals. i.e., All BMI values falling between 19-30 belong to the healthy class (or 1) and vice versa.

$$\mathbf{BMI} = \frac{\mathbf{weight(kg)}}{\mathbf{height^2(m)}} \text{ (height is in meters)}$$

ii. Project the data onto a line using LDA, visualize and mark an appropriate boundary (straight line) to separate the two classes.
You can do the visualizations, the matrix multiplications, and the eigenvalue decomposition using Matlab/Python. But you must implement LDA using the closed-form solution derived above, you can not use any libraries for it.

(c) Explain the geometric interpretation of LDA (a) in terms of the direction of projection of data in contrast to PCA (b) in terms of distance minimization or maximization for numerator and denominator in the objective function

(d) LDA requires that the within-class scatter matrix $\mathbf{S_W}$ be invertible. What happens if it is singular?

(e) Suggest a regularization technique that avoids matrix $\mathbf{S_W}$ singularity in LDA.

Problem 7 (8 marks)

In this problem, we will use gradient descent algorithm for linear regression. A university research team is studying how study hours impact exam performance. They collect data from 4 students, tracking their study hours under a given duration (of days) and their corresponding exam scores. The team decides to use linear regression with gradient descent to model this relationship. Given a data matrix $\mathbf{X}$, a parameter vector $\mathbf{w}$, a bias term b, and the output vector $\mathbf{y}$. Following is a linear model:

$$\tilde{y} = \mathbf{w^T x} + b$$

The data collected by team for the students is given in matrix X, the output vector $\mathbf{y}$, contains their exam score. To run the model they have used an initial estimates of the parameters $\mathbf{w_0}$.

$$\mathbf{X} = \begin{bmatrix} 1 & 5 \\ 1 & 2 \\ 1 & 7 \\ 1 & 4 \end{bmatrix}, \mathbf{w_0} = \begin{bmatrix} 25.0 \\ 1.5 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 65 \\ 55 \\ 75 \\ 60 \end{bmatrix}$$

The column of 1 in the data matrix incorporates the bias term. Please note that you must use the MSE loss function to compute the loss after each iteration. Take the value for the learning rate $\alpha = 0.01$.

Note : You need to show complete work, and not just the answers for the following parts

(a) Compute the predictions for the first iteration.

(b) Calculate the Mean Squared Error (MSE) loss after the first iteration.

(c) Compute the gradient and update the weight vector

(d) Repeat for the second iteration and present the loss value and updated weights

(e) use the updated weights of second iteration to predict the exam score for a student who studies for 3 hours per day

(f) the prediction of the previous part, makes sense? (based on data provided) if No, how would you proceed with your gradient descent algorithm ?

Problem 8 (8 marks)

The Government of Punjab, after growing tired of watching its Chief Minister being ridiculed online by fake news, has decided to take action. They have approached the CITY at LUMS to develop a fake news detection system. The CITY, working with their competent Machine Learning class, has built two competing AI models (classifiers), **Model A** and **Model B**, and tested them on 800 news articles, each labeled as either **Fake** (misinformation) or **Real** (authentic journalism). The following are the results they have obtained for the two models:

Model A:

- Correctly identified 310 fake news articles as fake
- Incorrectly identified 140 real news articles as fake
- Incorrectly classified 130 fake news articles as real
- Correctly classified 220 real news articles as real

Model B:

- Correctly identified 305 fake news articles as fake
- Incorrectly identified 130 real news articles as fake
- Incorrectly classified 145 fake news articles as real
- Correctly classified 225 real news articles as real

The goal of this problem is to determine which model should be deployed to help filter out misleading content while protecting freedom of speech.

(a) construct a confusion matrix based on provided information.

(b) How well do the two models correctly identify fake news out of all actual fake news?

(c) How often do the two models correctly classify fake news, i.e., correctly identified fake news out of all predicted fake news?

(d) Which model makes fewer total classification errors?

(e) Which model provides a better trade-off between detecting fake news and minimizing false positives?

(f) After evaluating the model's performance (b-e), which model do you think should be officially deployed? Do give a thought over catching fake news vs. mislabeling real news, i.e., free speech restrictions.

Show your working for all of the parts above.

Problem 9 (12 marks)

The curse of dimensionality refers to the phenomenon where, as the number of dimensions d increases, the behavior of distance-based algorithms like k-Nearest Neighbors (k-NN) changes significantly. Specifically, distances between points become less meaningful, making nearest-neighbor-based methods less effective.

Let's assume we have a unit d-dimensional hypercube $[0, 1]^d$, where N points are uniformly distributed. You have to prove that in d-dimensional space, the expected Euclidean distance d_{NN} to the nearest neighbor follows approximately.

$$d_{NN} \approx (\frac{lnN}{N})^{\frac{1}{d}}$$

where N is the number of points.

Also comment on the interpretation of this result/ approximation.

— End of Assignment —