Question 1: Midterm Correction Choose the question part (e.g. 1a or 2c) on the midterm you performed most poorly on. Please type up a solution guide that explains the solution and steps needed to arrive at this solution. Please show your work (or if it is a conceptual question, details on how you analyze the concept and evaluate the importance). Then include a section in which you detail your mistakes and explain your new understanding of the problem. Finally, attach an image of the question you are correcting to show the points taken off and the adjustments made. This question will count for both homework credit and give you the possibility of gaining up to 10 points back on the midterm question. If you do not have a question that you lost 10 points on, you will receive full credit for the question and the remaining points will be considered extra credit on top of your overall exam score.

Q1.b) The snapshot of the question paper is as below:

b) (15 points) Please describe gradient descent. What are the key parameters and how do they impact performance?

Gradient descent is the process of finding the local minima of a convex function. There are 3 types: Stochastic, Batch & mini batch. The hyper parameter of gradient descent is α which is the step size. This parameter helps providing how much or how little steps we take to change the weight.

Small α results in smaller steps and longer time to reach minima.

Large α results in overshooting the minima.

We determine the best α by cross validation

how? -5

stopping? -5

c) (20 points) Please give the loss optimization function for Lasso and for Ridge Regression. Describe the differences in impact each would have on a trained logistic regression model.

Ridge:

$$\sum_{i=1}^{n} \left(y_i - \beta_{0i} - \beta_{1i}x_i\right)^2 + \sum_{i=1}^{n} \lambda\left(\beta_{0i}^2 + \beta_{1i}^2\right) \quad \text{where}$$

What it more than 2?

Lasso:

$$\sum_{i=1}^{n} \left(y_i - \beta_{0i} - \beta_{1i}x_i\right)^2 + \sum_{i=1}^{n} \lambda\left(|\beta_{0i}| + |\beta_{1i}|\right) \quad \text{where}$$

$$\beta_0^2 + \beta_1^2 \leq 3$$

$$|\beta_0| + |\beta_1| \leq 3$$ -3

Differences in impact:-

Ridge regression has the L^2 norm which results in the predictors shrinking in value and values of predictors coming close to 0 when $\lambda \to \infty$ and same as RSS when $\lambda = 0$

Lasso Regression has the L^1 norm which results in some predictors actually reaching value 0 when $\lambda \to \infty$ and RSS when $\lambda = 0$.

Thus Lasso performs feature selection as well and performs better than Ridge when $P > n$ and all predictors are not important

The solution is as below:

I have addressed the missing parts of how gradient descent is applied and what are the stopping criteria for gradient descent.

Q.1.b) We can perform ordinary least square regression by using ~~either~~ the following approaches:

a) Solving the model parameters

b) Using optimization algorithm such as gradient descent.

Using a gradient descent algorithm, weights are updated incrementally after each iteration

The cost function $J(\cdot)$ is given as

$$J(w) = \frac{1}{2} \sum_i (y^i - \hat{y}^i)^2$$

The magnitude and direction of the weight update is computed by taking a step in the opposite direction of the cost function

$$\Delta w_j := -\eta \frac{\partial J}{\partial w_j}$$

where η is the learning rate. The weights are updated by

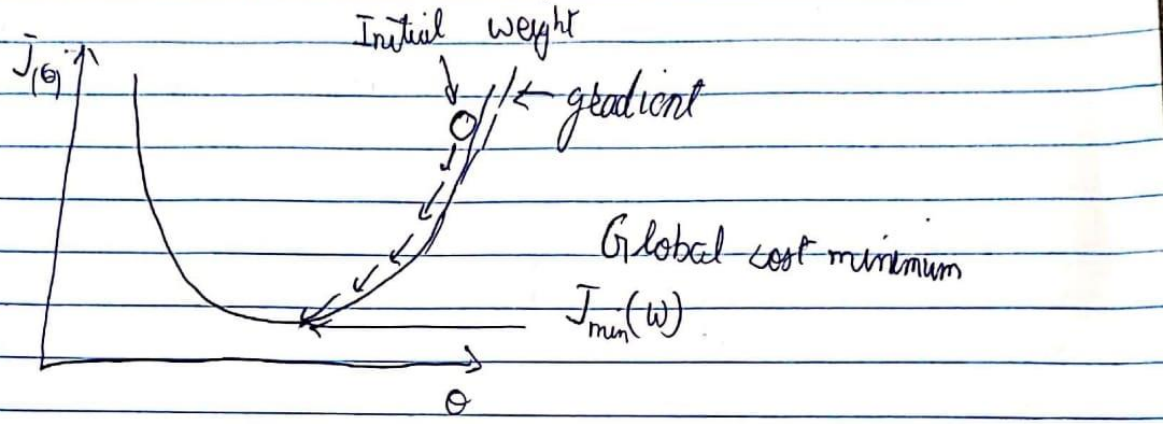
$$w := w + \Delta w$$

where Δw is a vector that contains the weight updates of each w

$$\Delta w_j = -\eta \frac{\partial J}{\partial w_j}$$

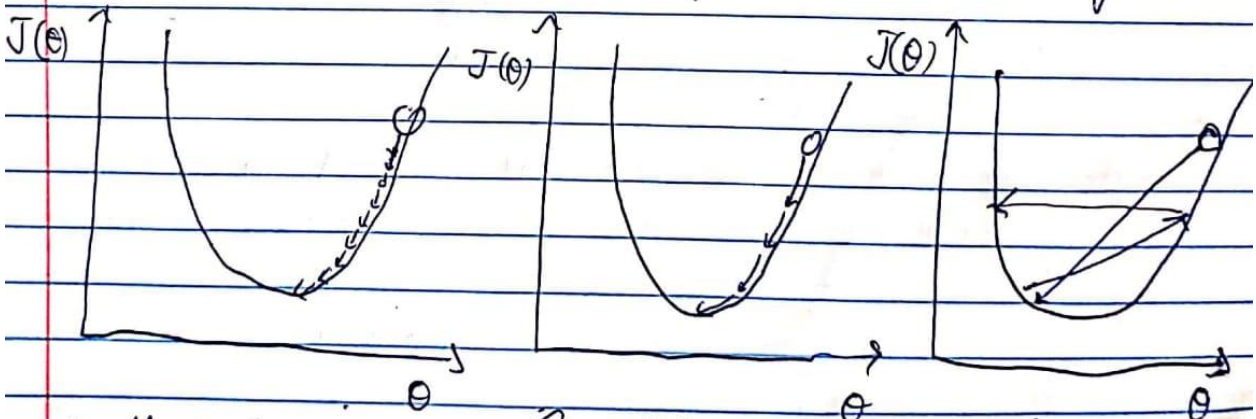
$$= -\sum_i \left(y^{(i)} - \hat{y}_i^{(i)}\right)\left(-x_j^{(i)}\right)$$

$$= \sum_i \left(y^{(i)} - \hat{y}_i^{(i)}\right) x_j^{(i)}$$



We have 3 conditions of learning rate

Too low | Just right | Too high



small η requires many updates before reaching minima

The optimum η swiftly reaches minimum

Too large η leads to drastic updates and we can miss the minima

Stopping criteria

There are two broad ways to stop gradient descent

1) Set the maximum number of iterations.

2) Terminate the algorithm when the value is below a threshold

$$|f(x_i) - f(x_{i-1})| \leq \epsilon$$

Q2.b) The snapshot of the paper is as below:

I lost complete marks in this question because I did not understand the question. Upon reviewing the homework1, I recollected what had been done there and wrote the solution accordingly addressing all parts asked in the question.

b) (15 points) Assume you have polynomial regression of degree d. Please provide the RSS formulation, then provide the formulas that would optimize each beta coefficient.

$$RSS = \sum_{i=1}^{n} \sum_{j=1}^{d} \beta_0 + \beta_{ij} x_{ij}^{j}$$

$$\beta_0 = (\bar{y} + \beta_1 \bar{x})^d .$$

15

Q 2b)

The polynomial regression is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_m x_i^d + \varepsilon_i$$

This can be expressed in matrix form as :-

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_d \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ 1 & x_3 & x_3^2 & \dots & x_3^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_n^d \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{n-d} \end{bmatrix}$$

or

$$y = X\beta + \varepsilon$$

The estimate of β is

$$\hat{\beta} = (X^T X)^{-1} X y$$

where $X^T X$ is a $(d+1) \times (d+1)$ matrix

Rss is given by the following :

$$\varepsilon_i = y_i - \hat{y}_i$$

For the case of exponents d, the least square method is applied by minimizing the SSE which yields the following optimization problem:

$$\min_{\beta} C(\beta) = \sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - X_{ij}\beta_j\right)^2$$

where j goes from 1 to d.

The optimal regression coefficient β^* are obtained from stationary point of the problem above. In matrix notation, this is represented as:

$$\frac{\partial C(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} (y - X\beta)^T (y - X\beta) = 2X^T(y - X\beta) = 0$$

$$\Rightarrow \boxed{\beta^* = (X^T X)^{-1} X^T y}$$