LAHORE UNIVERSITY OF MANAGEMENT SCIENCES
Department of Electrical Engineering

EE514/CS535 Machine Learning
Quiz 06 Solutions

Name: ____________________

Campus ID: ____________________

Total Marks: 10
Time Duration: 15 minutes

Question 1 (4 marks)

Choose the correct answer.

(a) If a function is convex, gradient descent is guaranteed to converge to the global minimum. (T/F)

Solution: F: does not converge if the learning rate is not suitable.

(b) Since the learning rate α in the gradient descent algorithm is a hyperparameter, it can be used to quantify the trade-off between variance and bias in the model. (T/F)

Solution: F: it can effect the bias variance tradeoff i.e., by letting the model converge to suboptimal minima or not converge at all. But it can not be used to 'quantify' the tradeoff.

(c) Stochastic gradient descent can be used to effectively tackle outliers in the dataset. (T/F)

Solution:
T: The outliers are very few, stochastically sampling the dataset will result in the outliers not being selected as often.
F: SGD is severely impacted by outliers, they are balanced by other points in the batch in the case of GD.

(d) We are using stochastic gradient descent to minimize the loss function $\mathcal{L}(\mathbf{w})$, where $\mathbf{w}$ denotes the model parameters. If α denotes the learning rate and $\mathcal{L}_i(\mathbf{w})$ denotes the loss function for i-th training input, we carry out the following update in each iteration:

i. $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla \sum_{i=1}^{n} \mathcal{L}_i(\mathbf{w})$, where n is the number of training data points

ii. $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla \mathcal{L}_i(\mathbf{w})$

iii. $\mathbf{w} \leftarrow \mathbf{w} + \alpha \nabla \mathcal{L}_i(\mathbf{w})$

iv. $\mathbf{w} \leftarrow \mathbf{w} + \alpha \nabla \sum_{i=1}^{n} \mathcal{L}_i(\mathbf{w})$, where n is the number of training data points

Solution: ii. we use only the loss for the i-th training input.

Question 2 (6 marks)

Consider the linear model, $\mathbf{y} = \mathbf{w}^T\mathbf{x}$, we get the optimal value of the parameter vector $\mathbf{w}$ by minimizing the following cost function:

$$J(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x_i} - y)^2 = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

(a) [2 marks] What is the update step of the gradient descent for which $J(\mathbf{w})$ is minimized? You can use any of the formulations of the cost function written above.

(b) [1 mark] The loss function $J(\mathbf{w})$ weights all examples equally. If some examples in the training data are more important than others, we would like to give weightage β_i to the i-th data point. Formulate a cost function that incorporates this weightage.

(c) [3 marks] What is the update step for the updated cost function?

Solution:

(a) the update step of batch gradient descent for the standard cost function J is,

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{1}{n} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x_i} - y) \mathbf{x_i}$$

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha (\mathbf{X^T X w} - \mathbf{X^T y})$$

(b)

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} (\beta_i (\mathbf{w}^T \mathbf{x}_i - y))^2$$

β is the vector of length n, β_i is the weight of x_i.

$$J(w) = \frac{1}{n} ||\beta(\mathbf{Xw} - \mathbf{y})||^2$$

β is a $n \times n$ diagonal matrix, $\beta_{i,i}$ is the weight of x_i.

(c)

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{1}{n} \sum_{i=1}^{n} \beta_i (\mathbf{w}^T \mathbf{x_i} - y) \mathbf{x_i}$$

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha (\mathbf{X^T} \beta \beta^{\mathbf{T}} \mathbf{X w} - \mathbf{X^T} \beta \beta^{\mathbf{T}} \mathbf{y})$$