Summer of ML 2021: Week 1 Day 2 Quiz
Multivariable Calculus & Linear Regression

by Nirjhar Das

Wednesday 16$^{\text{th}}$ June, 2021

Questions

1. Let Loss function be $J(\theta_0, \theta_1) = \sum_{i=1}^{m} |\theta_0 + \theta_1 \cdot x_i - y_i|$. Then $\frac{\partial J}{\partial \theta_0} =$

   A. 0

   B. m

   C. $\sum_{i=1}^{m} sgn(\theta_0 + \theta_1 \cdot x_i - y_i)$

2. Which of the following is TRUE?

   A. Gradient Descent will always converge.

   B. Gradient Descent will stop as soon as Loss function = 0

   C. Gradient Descent can overshoot the minima if learning rate is too large.

   D. Gradient Descent doesn't depend on feature scaling.

3. Suppose $J(x, y) = x^2y + y^2x$. What is the Hessian at (0, 1)?

   A. $\begin{bmatrix} 2 & 2 \\ 2 & 0 \end{bmatrix}$

   B. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

   C. $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

4. Which of the following are TRUE?

   A. Linear Regression always finds the best fit to a data.

   B. Linear Regression can fit polynomial with degree > 1, when higher dimensional features are allowed.

   C. The line produced by Linear Regression is independent of the loss function used.

   D. Linear Regression models has high interpretability.

Answers

1. **Answer: C**

$$\frac{\partial J}{\partial \theta_0} = \frac{\partial}{\partial \theta_0}(\sum_{i=1}^{m} |\theta_0 + \theta_1 \cdot x_i - y_i|) = \sum_{i=1}^{m} \frac{\partial}{\partial \theta_0} |\theta_0 + \theta_1 \cdot x_i - y_i|$$

Now apply Chain Rule. We know $\frac{\partial |f(x)|}{\partial x} = sgn(f(x)) \frac{\partial f(x)}{\partial x}$
Thus,

$$\begin{aligned} \frac{\partial}{\partial \theta_0} |\theta_0 + \theta_1 \cdot x_i - y_i| &= sgn(\theta_0 + \theta_1 \cdot x_i - y_i) \cdot \frac{\partial}{\partial \theta_0}(\theta_0 + \theta_1 \cdot x_i - y_i) \\ &= sgn(\theta_0 + \theta_1 \cdot x_i - y_i) \cdot 1 \end{aligned}$$

Hence, option C is correct.

2. **Answer: C**

A. Gradient Descent may not converge as it is an iterative algorithm that tries to reach the local minima by taking greedy steps. The convergence depends on number of iterations and also on the learning rate.

B. Gradient Descent stops when a local minima is achieved (where derivative is zero w.r.t all variables). However, it is not necessary that the point where Loss function = 0, is the minima. For different loss functions, the value at the minima can be negative.

C. If learning rate is too large, the steps taken in the direction of decrease may be too large, thus skipping the minima. And when the minima is skipped, the gradient descent algorithm will diverge (try it out in notebooks) as shown in fig.1. Hence, this is correct.

D. Gradient Descent depends on feature scaling in terms of rate of convergence. When the features are scaled down to a narrow range, gradient descent quickly attains minima with smaller number of iteration. More details can be found here.

3. **Answer: A**

$$H = \begin{bmatrix} \frac{\partial^2 J}{\partial x^2} & \frac{\partial^2 J}{\partial x \partial y} \\ \frac{\partial^2 J}{\partial y \partial x} & \frac{\partial^2 J}{\partial y^2} \end{bmatrix}$$

$$\frac{\partial J}{\partial x} = 2xy + y^2, \quad \frac{\partial J}{\partial y} = x^2 + 2xy, \quad \frac{\partial^2 J}{\partial x^2} = 2y, \quad \frac{\partial^2 J}{\partial y^2} = 2x, \quad \frac{\partial^2 J}{\partial x \partial y} = \frac{\partial^2 J}{\partial y \partial x} = 2x + 2y$$

Now put $(x, y) = (0, 1)$ to evaluate the Hessian H.

4. **Answer: B, D**

A. Linear Regression *assumes* that the given distribution can be fit by a (multivariable) polynomial (usually of degree 1). This is the **inductive bias** of Linear Regression. For more on inductive bias, read this cool thread on StackOverflow.

B. Recall that all n-times differentiable function $f(x)$ can be expanded into a polynomial of degree n using Taylor's theorem. Thus, using linear regression, we can find the suitable coeffiecient of this polynomial to approximate $f(x)$.

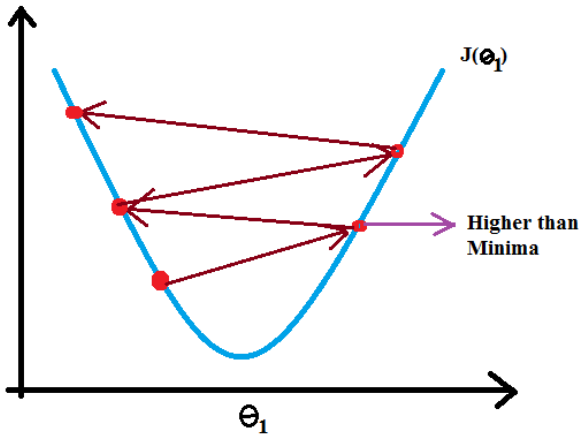

Figure 1: Gradient Descent Overshoot

C. Line produced very much depends on the loss function as the objective of solving the linear regression problem is to minimize the loss function. Hence, on changing the loss function, the objective is changed and subsequently, the solution obtained. Here is a beautiful write-up to refer to for understanding more about choices of loss functions and their effects.

D. This is true since, from a linear regression model, we can directly understand the contribution made by each feature on the output by looking at the learned weights $\overrightarrow{\theta} = (\theta_0, \theta_1 \dots \theta_n)$. The greater the value of θ_i, the greater the contribution of x_i.