ASSIGNMENT 1 – WRITTEN PART

Note: The solution is needed for the programming part.

P1. Suppose you are given a training data set $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. Each x_i is a scalar real-valued input; each y_i is a scalar real-valued output. We will consider using a linear function to fit the data, i.e., linear regression. In other words, we will consider the family of linear functions

$$y = \theta_0 + \theta_1 x,$$

where θ_0 and θ_1 are unknown (scalar) parameters. We wish to learn θ_0 and θ_1 from the training data. Here, we will not worry about generalization. We will all N training instances for learning.

As usual, our objective is to minimize the residue sum of squares, sometimes known as square error. That is,

$$\min_{\theta_0,\ \theta_1} \sum_{i=1}^{N} \left(y_i - (\theta_0 + \theta_1 x_i)\right)^2 . \tag{1}$$

Please solve the above minimization problem. That is, express the optimal θ_0 and θ_1 in terms of the training data. You will use the result for your programming assignment.

Let h denote the objective function for the minimization problem in (1):

$$h(\theta_0, \theta_1) \triangleq \sum_{i=1}^{N} \left(y_i - (\theta_0 + \theta_1 x_i)\right)^2 .$$

You will find the following notations useful for simplifying your expressions.

$$\overline{x} \triangleq \frac{1}{N} \sum_{i=1}^{N} x_i, \qquad \overline{y} \triangleq \frac{1}{N} \sum_{i=1}^{N} y_i$$

$$\overline{x^2} \triangleq \frac{1}{N} \sum_{i=1}^{N} x_i^2, \qquad \overline{xy} \triangleq \frac{1}{N} \sum_{i=1}^{N} x_i y_i.$$

$$\overline{y^2} \triangleq \frac{1}{N} \sum_{i=1}^{N} y_i^2.$$

While solving the problem, please also answer the following:

(a) Express the objective function h in terms of $\overline{x}$, $\overline{y}$, $\overline{x^2}$ and $\overline{xy}$. Show that the critical points are:

$$\theta_0 = \frac{\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2}$$

$$\theta_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}.$$

(b) When does the critical point exist?
(c) You see that when a critical point exists, it is unique. Argue that, in this case, the critical point is the minimum.
(d) When a critical point does not exist, what must the training data set look like? You should conclude that it is very unlikely that a critical point does not exist.

Hint: The objective function in (1) $h(\theta_0, \theta_1)$ is a quadratic function in θ_0 and θ_1. Expand $h(\theta_0, \theta_1)$ and find its gradient with respect to the variables θ_0 and θ_1. The gradient of h is denoted by ∇h, and it is a vector

$$\nabla h = \begin{pmatrix} \frac{\partial h}{\partial \theta_0} \\ \frac{\partial h}{\partial \theta_1} \end{pmatrix}. \tag{2}$$

Therefore, you will need to compute the partial derivatives.

For an unconstrained optimization problem where the objective function is differentiable, a necessary condition for the point (θ_0, θ_1) to be an optimal (either a minimum or a maximum) is that it is a critical point, which means (θ_0, θ_1) satisfies $\nabla h(\theta_0, \theta_1) = 0$. Once you find the expression for ∇h, you will set it to zero and this gives you two equations to solve.

For part (c), in an unconstrained differentiable optimization problem, a point (θ_0, θ_1) that satisfies $\nabla h(\theta_0, \theta_1) = 0$ must be a minimum, a maximum or a saddle point. You have to rule out the possibilities of a maximum or a saddle point. You can reason by computing the Hessian and show it is positive definite. We will review Hessian in future lectures.

Alternatively, you can follow that line of argument outlined below. You first argue that h is unbounded from above. Therefore, it is easy to rule out the maximum case.

A saddle point is a critical point (i.e., a point where the gradient is equal to 0) at which the function rises in some directions and falls in others. You can argue that if a quadratic function falls in a direction, it falls towards negative infinity. To formalize the argument, you evaluate the function h in a falling direction and the function becomes a single-variable quadratic function. The function must go to negative infinity in that direction.

You will plot function like $\theta_0^2 - \theta_1^2$ which has a saddle point $(0, 0)$ in your programming part of the assignment.

For part (d), you will apply the Cauchy-Schwarz inequality to vectors $e = (1, 1, \ldots, 1)^T$ and $x = (x_1, \ldots, x_N)^T$.