Assignment 1

09/18/2020

I pledge that this assignment has been completed in compliance with the Graduate Honor Code and that I have neither given nor received any unauthorized aid on this assignment

Name: Yusuf Elnady

Signature: Yusuf Elnady

Q3.a

First, we have the following loss function for linear regression, and we want to minimize it to find the estimates of w and b

$$L(w,b) = \sum_{n=1}^{m} r_n \left(y_n - wx_n + b\right)^2$$

In order to estimate the values of w and b, I will need to do the partial derivatives for each of them, that means I need to calculate $\frac{\partial L(w,b)}{\partial w}$ and $\frac{\partial L(w,b)}{\partial b}$. This is basically the approach used in gradient descent (Iterative approach). I will define a vector of thetas/coefficients called

$$\theta = \begin{bmatrix} -b \\ w \end{bmatrix}$$

I have also the vector y_n which has the actual y values. There's also the matrix

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix}$$

The goal is to fit the data points $(x_n; y_n)$ in proportion to their weights r_n, so we define the matrix

$$R = \begin{bmatrix} r_1 & 0 & 0 & 0 \\ 0 & r_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & r_m \end{bmatrix}$$

Now after defining these values, we can rewrite the objective function to be

$$L(w,b) = \sum_{n=1}^{m} r_n \left(y_n - wx_n + b\right)^2$$

$$= (y - \theta^T X) R (y - \theta^T X)^T$$

$$= yRy^T - 2yRX^T\theta + \theta^T XRX^T\theta$$

How I converted the objective function to the matrix form? Basically, I rewrote $wx_n + b$ as $\theta^T x$ using the vector θ I made, then as the difference between the actual y value and the predicted value is squared, so we can multiply this difference by its transpose to get the same effect, what is new here is that I added the R matrix (the weights).

Following is some matrix derivative formulas that I will use to get the derivative of $L(w, b)$:

$$\frac{\partial AX}{\partial X} = A^T$$

$$\frac{\partial X^T A}{\partial X} = A$$

$$\frac{\partial X^T X}{\partial X} = 2X$$

$$\frac{\partial X^T AX}{\partial X} = AX + A^T X$$

To get the minimum we need to get the values of w and b when the objective function is at the zero gradient (slope = 0), so using the above formulas:

$$0 = -2XRy^T + 2XRX^T\theta$$

we can simplify it as:

$$XRy^T = XRX^T\theta$$

Thus we compute θ as

$$\boldsymbol{\theta = (XRX^T)^{-1}XRy^T}$$

And as mentioned previously $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{-b} \\ \boldsymbol{w} \end{bmatrix}$

Q3.b Given the original $\boldsymbol{y_n}$, we will add the Gaussian random variable (Normal distribution) $\boldsymbol{\varepsilon}$ that has variance of $\boldsymbol{\sigma^2}$ and a zero mean, that is $\boldsymbol{\varepsilon = \mathcal{N}(0, \sigma^2)}$. So, the new $\boldsymbol{y_n}$ is $\boldsymbol{y_n = w^T x - b + \varepsilon}$.

The probability density function of $\boldsymbol{\varepsilon}$ is given by

$$\boldsymbol{Pr(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\varepsilon^2}{2\sigma^2})}$$

and that implies that

$$\boldsymbol{Pr(y_n|X_n) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y - wx + b)^2}{2\sigma^2})}$$

The negative log-likelihood is a cost function that is used as a loss for machine learning models that describes the performance of the model.

Using the formula of the negative log-likelihood we get

$$-l(w, b) = -\log \prod_{n=1}^{M} Pr(y_n|X_n)$$

$$= -\sum_{n=1}^{M} \log Pr(y_n|X_n)$$

$$= -\sum_{n=1}^{M} \log \sqrt{2\pi\sigma_n^2} + \sum_{n=1}^{M} \frac{(y_n - wx_n + b)^2}{2\sigma^2}$$

$$= M \log \sqrt{2\pi\sigma_n^2} - \frac{1}{\sigma^2} . \frac{1}{2} \sum_{n=1}^{M} (y_n - wx_n + b)^2$$

Now we want to minimize the output of the negative log likelihood. So, for the first part of the last equation we got, we cannot minimize anything on it because it doesn't contain $\boldsymbol{w}$ or $\boldsymbol{b}$, so it will be neglected. For the second part by minimizing the output of the negative log-likelihood we get

$$\frac{1}{2} \sum_{n=1}^{M} (y_n - wx_n + b)^2$$

So, finally we can say that the objective we minimized in Q3.a is same as the negative log-likelihood for liner regression in Q3.b.**and $r_n = \frac{1}{2\sigma^2}$ and the variance of each measurement noise in this model is related inversely to the weight such that $r_n \propto \frac{1}{\sigma^2}$**