CSE 250B Winter 2020 - Quiz 2

Feb 7, 2020 - 30 Minutes **Name:** **PID:**

---

(1) This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.
(2) Start writing when instructed. Stop writing when your time is up.
(3) Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

(1) Alice has collected a dataset of dependent and independent variables $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$; she does linear regression on it and obtains the weight vector w_{Alice}. Bob also collects the same dataset, but while recording the independent variables (the $x^{(i)}$'s) he uses different units. Specifically, Bob's dataset is $\{(z^{(1)}, y^{(1)}), \ldots, (z^{(n)}, y^{(n)})\}$, where for each i, $z^{(i)} = cx^{(i)}$, where $c > 0$ is a scalar. Bob does linear regression on this dataset and obtains a weight vector w_{Bob}.

(a) (5 points) Do Alice and Bob have the same training loss? Is $w_{Alice} = w_{Bob}$? In either case, justify your answer.

If X is Alice's data matrix, and Z is Bob's data matrix, then $Z = cX$. $w_{Bob} = (Z^\top Z)^{-1} Z^\top y = (c^2 X^\top X)^{-1} c X^\top y = w_{Alice}/c$; thus when $c \neq 1$, $w_{Alice} \neq w_{Bob}$.
The training loss of Bob is: $\|Z w_{Bob} - y\|^2 = \|cX(w_{Alice}/c) - y\|^2 = \|X w_{Alice} - y\|^2$, which by definition is equal to the training loss of Alice.

(b) (5 Points) Suppose now that Bob records each feature in a different unit; that is, for each coordinate j, $z_j^{(i)} = c_j x_j^{(i)}$ for all i, where $c_j > 0$ is a scalar, and the c_j's are not all equal. Do Alice and Bob still have the same training loss and is $w_{Alice} = w_{Bob}$? If yes, justify your answer. If no, provide an example of a dataset where this is not the case.

The answer will still be the same. Let C be a diagonal matrix whose j-th diagonal entry is c_j. Observe that Bob's data matrix $Z = XC$. By the closed form solution of linear regression, we have

$$\begin{aligned}
w_{Bob} &= (Z^\top Z)^{-1} Z^\top y \\
&= ((XC)^\top XC)^{-1} (XC)^\top y \\
&= (C^\top X^\top XC)^{-1} C^\top X^\top y \quad \text{(because } (AB)^\top = B^\top A^\top\text{)} \\
&= (CX^\top XC)^{-1} CX^\top y \quad \text{(because } C = C^\top \text{ for diagonal matrix } C\text{)} \\
&= (C(X^\top X)C)^{-1} CX^\top y \\
&= C^{-1}(X^\top X)^{-1} C^{-1} CX^\top y \quad \text{(because } (AB)^{-1} = B^{-1}A^{-1} \text{ when } A, B \text{ are invertible.)} \\
&= C^{-1}(X^\top X)^{-1} X^\top y = C^{-1} w_{Alice},
\end{aligned}$$

and the training loss for Bob is: $\|Z w_{Bob} - y\|^2 = \|XC(C^{-1} w_{Alice}) - y\|^2 = \|X(CC^{-1}) w_{Alice} - y\|^2 = \|X w_{Alice} - y\|^2$ which is equal to Alice's training loss.

(2) (5 points) Consider the following loss function:

$$L(w) = \sum_{i=1}^{n} \log(1 + w^\top x_i)$$

What is $\nabla L(w)$? Write down the update step for gradient descent.

Let $L_i(w) = \log(1 + w^\top x_i)$. Then

$$\begin{aligned} \nabla L(w) = & \sum_{i=1}^{n} \nabla L_i(w) \\ = & \sum_{i=1}^{n} \frac{x_i}{1 + w^\top x_i}. \end{aligned}$$

The gradient descent update step for w is then

$$w_{t+1} = w_t - \eta_t \nabla L(w) = w_t - \eta_t \sum_{i=1}^{n} \frac{x_i}{1 + w_t^\top x_i},$$

where η_t is the learning rate at the t-th descent step.