ECE421 - Winter 2022
Homework Problems - Tutorial #2

Theme: Perceptron Learning Algorithm and Linear Regression

Due: January 30, 2022 11:59 PM

In the following, LFD refers to the course textbook "Learning from Data".
For all questions we denote the weight vector by

$$w = \begin{bmatrix} b & w_1 & w_2 & \dots & w_d \end{bmatrix}^\top \in \mathbb{R}^{d+1},$$

where $b \in \mathbb{R}$ is the bias term, and we denote the example vectors by

$$x = \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_d \end{bmatrix}^\top \in \mathbb{R}^{d+1}.$$

Question 1 (Perceptron Learning Algorithm)

Given a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$, where $x_n \in \mathbb{R}^d$ and $y_n \in \{+1, -1\}$, we wish to train a perceptron model

$$h(x) = \text{sign}(b + \textstyle\sum_{i=1}^d w_i x_i) = \text{sign}(w^\top x), \tag{1}$$

that correctly classifies *all* examples in $\mathcal{D}$. Consider the perceptron weight update rule (1.3) on Page 7 of LFD, i.e.,

$$w(t+1) = w(t) + y(t)x(t). \tag{2}$$

This weight update rule moves the weights in the direction of classifying examples correctly. To see this, show the following.

(a) If $x(t)$ is misclassified by $w(t)$, show that $y(t)w^\top(t)x(t) < 0$.

(b) Use equation for $w(t+1)$ to show that $y(t)w(t+1)x(t) > y(t)w(t)^\top x(t)$.

(c) Argue that the weight update from $w(t)$ to $w(t+1)$ is a move "in the right direction".

Remark: Problem 1.3 on page 33 shows the steps toward a rigorous proof of the convergence of the perceptron algorithm. You may wish to look ahead and see if you can partially solve this problem on your own. The solution will be explained in detail in tutorials.

Question 2 (Linear Regression)

Given a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, where $x_n \in \mathbb{R}^d$ and $y_n \in \mathbb{R}$, we wish to train a linear regression model

$$h(x) = b + \sum_{i=1}^{d} w_i x_i = w^\top x \tag{3}$$

that fits the examples in $\mathcal{D}$. In-sample error associated with linear regression model is

$$E_{\text{in}}(w) = \frac{1}{2N} \sum_{n=1}^{N} (w^\top x_n - y_n)^2. \tag{4}$$

(Note: The $\frac{1}{2N}$ coefficient sometimes appear as $\frac{1}{N}$. This doesn't make a difference because positive scaling does not change the optimum weight vector.)
Define the data matrix X and the target vector y as

$$X = \begin{bmatrix} x_{10} & x_{11} & \dots & x_{1d} \\ \vdots & \vdots & & \vdots \\ x_{N0} & x_{N1} & \dots & x_{Nd} \end{bmatrix} = \begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix} \in \mathbb{R}^{N \times (d+1)} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N,$$

where $x_i = \begin{bmatrix} x_{i0} & x_{i1} & \dots & x_{id} \end{bmatrix}^\top$ and $x_{i0} = 1$ for all $i \in \{1, \dots, N\}$ (similar to Problem 3.9 in LFD).

(a) Show that in-sample error in (4) can be written as

$$E_{\text{in}}(w) = \frac{1}{2N} \|Xw - y\|_2^2 = \frac{1}{2N} (w^\top X^\top X w - 2 w^\top X^\top y + \|y\|_2^2). \tag{5}$$

(b) Find the expressions for the gradient of (4) and (5) with respect to w. (You may wish to start with a low-dimensional example, e.g., set $d = 2$ and $N = 2$). Verify that the gradient expressions for (4) and (5) are equivalent.

(c) Suppose $X^\top X$ is invertible. Let $w^\star = (X^\top X)^{-1} X^\top y$ and $y_{\text{ls}} = X w^\star = X (X^\top X)^{-1} X^\top y$. Show that $E_{\text{in}}(w)$ can be decomposed into sum of two sub-parts as

$$E_{\text{in}}(w) = \frac{1}{2N} (\|Xw - y_{\text{ls}}\|_2^2 + \|y - y_{\text{ls}}\|_2^2). \tag{6}$$

(Hint: Show that the inner product $(Xw - y_{\text{ls}})^T (y - y_{\text{ls}})$ is zero.)

(d) Use the result in (c) to show that the solution to the least-squares problem, $\text{argmin}_w E_{\text{in}}(w)$, is

$$w^\star = (X^\top X)^{-1} X^\top y.$$

Knowing that $w^\star$ is the minimizer, can you connect Xw and y_{ls} to range of X and explain geometrically why $(Xw - y_{\text{ls}})^T (y - y_{\text{ls}})$ is zero?