ECE421: Introduction to Machine Learning — Fall 2024

Worksheet 1: Pocket Algorithm and Linear Regression

Notation

(a) We use a **underline** to represents **column vectors**, *e.g.*, $\underline{p} \in \mathbb{R}^k$ represents a column vector with k elements. We adopt the following notations to list the elements of a **column vector**

$$\underline{p} = (p_1, p_2, \ldots, p_k) = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{bmatrix}.$$

Note the usage of parentheses and brackets. The notation with parentheses provides a more compact representation of vectors and optimizes space usage.
Additionally, **row vectors** can be represented by $\underline{q}^\top = [p_1, p_2, \ldots, p_k]$. Note the use of transpose and brackets.
Finally, the context and notation should make it clear whether a vector is a column vector or a row vector.

(b) For all questions we denote the weight vector by $\underline{w} = (b, w_1, \ldots, w_d) \in \mathbb{R}^{d+1}$, where $b \in \mathbb{R}$ is the bias term, and we denote the example vectors by $\underline{x} = (1, x_1, x_2, \ldots, x_d) \in \mathbb{R}^{d+1}$.

(c) In the following, LFD refers to the textbook "Learning from Data."

Q0 Linear Algebra Review

0.a (The ℓ_p-norm) For a real number $p \geq 1$, define the ℓ_p-norm of a vector $\underline{x} \in \mathbb{R}^n$.

0.b (The ℓ_1, ℓ_2, and ℓ_∞-norm) Consider the vector $\underline{x} = (5, 2, -3)$. Find the ℓ_1, ℓ_2, and ℓ_∞-norm of $\underline{x}$.

0.c (Matrix Multiplication) Let $\underline{w} = (w_0, w_1, \ldots, w_d)$ and $\underline{x}_i = (x_{i0}, x_{i1}, \ldots, x_{id})$ for $i \in \{1, 2, \ldots, N\}$. Let

$$X = \begin{bmatrix} x_{10} & x_{11} & \ldots & x_{1d} \\ x_{20} & x_{21} & \ldots & x_{2d} \\ \vdots & \vdots & \ldots & \vdots \\ x_{N0} & x_{N1} & \ldots & x_{Nd} \end{bmatrix} = \begin{bmatrix} \underline{x}_1^\top \\ \underline{x}_2^\top \\ \vdots \\ \underline{x}_N^\top \end{bmatrix},$$

$$\underline{\hat{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} \underline{w}^\top \underline{x}_1 \\ \underline{w}^\top \underline{x}_2 \\ \vdots \\ \underline{w}^\top \underline{x}_N \end{bmatrix}.$$

Show that $\underline{\hat{y}} = X\underline{w}$.

Q1 Gradient and Optimization Fundamentals

1.a (Gradient) Prove that $\nabla_{\underline{x}}(\underline{a}^\top \underline{x}) = \underline{a}$, and $\nabla_{\underline{x}}(\underline{x}^\top \underline{a}) = \underline{a}$ and $\nabla_{\underline{x}}(\underline{x}^\top A \underline{x}) = 2A\underline{x}$, where $\underline{a}$ and $\underline{x}$ are vectors with k entries and A is a symmetric squared matrix.

1.b (Exercise 3.17 (a),(b) in LFD) Recall that for a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$ and a vector $\underline{p} \in \mathbb{R}^n$, the first-order Taylor series approximation of $f(\underline{x}+\underline{p})$ is $f(\underline{x}+\underline{p}) \approx f(\underline{x}) + \nabla f(\underline{x})^\top \underline{p}$. Consider the function $E(u, v) = e^u + e^{2v} + e^{uv} + u^2 - 3uv + 4v^2 - 3u - 5v$, where u and v are scalars.

1.b. i Denote by $\hat{E}_1(\Delta u, \Delta v)$ the first-order Taylor series approximation of E at $(u, v) = (0, 0)$. We know that $\hat{E}_1(\Delta u, \Delta v)$ is of the form $\hat{E}_1(\Delta u, \Delta v) = a_u \Delta u + a_v \Delta v + a$. What are the values of a_u, a_v, and a?

1.b. ii Minimize $\hat{E}_1$ over all possible $(\Delta u, \Delta v)$ such that $\|(\Delta u, \Delta v)\|_2 = 0.5$, *i.e.*,

$$\begin{aligned} \min_{\Delta u, \Delta v} \quad & \hat{E}_1(\Delta u, \Delta v) \\ \text{s.t.} \quad & \|(\Delta u, \Delta v)\|_2 = 0.5. \end{aligned}$$

Recall that the column vector $\begin{bmatrix} \Delta u^* \\ \Delta v^* \end{bmatrix}$ that minimizes $\hat{E}_1$ is in the direction of $-\nabla E(u, v)$, *i.e.*, the negative gradient direction. Compute $\begin{bmatrix} \Delta u^* \\ \Delta v^* \end{bmatrix}$ that minimizes $\hat{E}_1$, and the resulting $\hat{E}_1(\Delta u^*, \Delta v^*)$.

Q2 **(Perceptron Learning Algorithm)**

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

$$h(\underline{x}) = \text{sign}\left(b + \sum_{i=1}^d w_i x_i\right) = \text{sign}(\underline{w}^\top \underline{x})$$

that correctly classifies *all* examples in $\mathcal{D}$. Consider the perceptron weight update rule

$$\underline{w}(t+1) = \underline{w}(t) + y_n \underline{x}_n,$$

where $(\underline{x}_n, y_n)$ is the misclassified datapoint after iteration t. This weight update rule moves the weights in the direction of classifying examples correctly. To see this, show the following.

2.a If $\underline{x}(t)$ is misclassified by $\underline{w}(t)$, show that $y_n \underline{w}^\top(t) \underline{x}_n < 0$.

2.b Use the equation for $\underline{w}(t+1)$ to show that $y_n \underline{w}^\top(t+1) \underline{x}_n > y_n \underline{w}^\top(t) \underline{x}_n$.

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

[**REMARK:** Problem 1.3 in LFD, page 33, shows steps towards a rigorous proof of convergence of the Perceptron algorithm. Feel free to attempt solving this problem on your own. This is an optional exercise.]

Q3 **(Linear Regression)**

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

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

The in-sample error associated with the linear regression model is

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

Define the data matrix X and target vector $\underline{y}$ as:

$$X = \begin{bmatrix} x_{10} & x_{11} & \dots & x_{1d} \\ \vdots & \vdots & \dots & \vdots \\ x_{N0} & x_{N1} & \dots & x_{Nd} \end{bmatrix} = \begin{bmatrix} \underline{x}_1^\top \\ \vdots \\ \underline{x}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times (d+1)},$$

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N.$$

where $\underline{x}_i = (x_{i0}, x_{i1}, \dots, x_{id})$ and $x_{i0} = 1$ for all $i \in \{1, 2, \dots, N\}$.

3.a Show that the in-sample error can be written as:

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

3.b Find the expressions for the gradient of (1) and (2) with respect to $\underline{w}$. Verify that the gradients of the two forms are equivalent.

3.c Suppose $X^\top X$ is invertible. Let $\underline{w}^\star = (X^\top X)^{-1}X^\top \underline{y}$. Show that $E_{\text{in}}(\underline{w})$ can be decomposed as:

$$E_{\text{in}}(\underline{w}) = \frac{1}{2N}\left(\|X\underline{w} - \underline{y}_{\text{ls}}\|_2^2 + \|\underline{y} - \underline{y}_{\text{ls}}\|_2^2\right),$$

where $\underline{y}_{\text{ls}} = X\underline{w}^\star$.

3.d Use the result in (c) to show that the least-squares solution is $\underline{w}^\star = (X^\top X)^{-1}X^\top \underline{y}$. Explain geometrically why

$$(X\underline{w} - \underline{y}_{\text{ls}})^\top(\underline{y} - \underline{y}_{\text{ls}}) = 0.$$