CS 1678/2078 Homework 1

Abstract

This assignment is the first introduction to doing gradient based optimization. Specifically, it focuses on solving regression and classification problems using linear function approximation with and without a basis function. In this assignment, you will solve analytically for the partial derivative of a loss function with respect to the model parameters (weights), and you will begin the basic setup of a program to perform linear regression on a provided data set. To submit this assignment, upload a `.pdf` to Gradescope containing your responses to the questions below. You are required to use LaTeX for your write up. To submit the assignment's coding portion, upload a zip folder to gradescope containing all python files. This code will be used to verify your answers and check for plagiarism. We will be using cheating detection software, so, as a reminder, you are allowed to discuss the homework with other students, but you must write your code on your own. You may also not use ChatGPT, Co-Pilot, or any other AI software to write your answers or code.

1 Written Responses

Consider approximating the function

$$f_*(x) = 6x + 4\cos(3x+2) - x^2 + 10\ln\left(\frac{|x|}{10} + 1\right) + 7$$

with the function approximator $f(x, w) = \phi(x)^\top w$, where $\phi\colon \mathbb{R} \to \mathbb{R}^n$ and $w \in \mathbb{R}^n$, i.e., $\phi(x)$ creates a feature vector of length n and the weights w are a vector length n.

1. (3 points) What are features that allow f to exactly represent f_*? Specify the features $\phi(x)$.
   Answer:
   $$\phi(x) = \left[x, \cos(3x+2), x^2, \ln\left(\frac{|x|}{10} + 1\right), 1\right]^\top$$

2. (2 points) What are the optimal weights for this approximation?
   Answer:
   $$w^* = [6, 4, -1, 10, 7]^\top$$

3. (3 points) What would be the basis function and weights to perfectly represent the function
   $$f_*(x) = 6x \times 4\cos(3x+2) \times x^2 \times 10\ln\left(\frac{|x|}{10} + 1\right) \times 7?$$
   Answer:
   $$\phi(x) = \left[x\cos(3x+2)x^2\ln\left(\frac{|x|}{10} + 1\right)\right]^\top$$
   $$w^* = [6 \times 4 \times 10 \times 7]^\top$$

4. (3 points) In the programming homework below we will be making predictions for multiple data points at a time. More specifically, we will consider linear function approximation of the form $f\colon \mathbb{R}^{m\times n} \times \mathbb{R}^n \to \mathbb{R}^m$, where f takes as input a matrix $X \in \mathbb{R}^{m\times n}$ that represents a m data points each being a row vector of n features, a vector $w \in \mathbb{R}^n$ that represents the parameters of the function, and maps these to a vector $\hat{y} \in \mathbb{R}^m$ that represents a prediction for m data points, i.e.,
   $$\hat{y} = f(X, w) = Xw.$$

Let $y \in \mathbb{R}^m$ represent the target (or label) for data point, i.e. y_i and $\hat{y}_i$ are the i^{th} labels and prediction, respectively. Consider the mean squared error loss function $g \colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ for each batch of predictions, e.g.,

$$g(\hat{y}, y) = \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_i - y_i\right)^2.$$

What is the partial derivative of g with respect to the predictions? Note the derivative should be a vector in $\mathbb{R}^m$.

Answer:

$$\begin{aligned}
\frac{\partial}{\partial \hat{y}} g(\hat{y}, y) &= \frac{\partial}{\partial \hat{y}}\frac{1}{2}\frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_i - y_i\right)^2 \\
&= \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial \hat{y}}\left(\hat{y}_i - y_i\right)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_i - y_i\right)\frac{\partial(\hat{y}_i - y_i)}{\partial \hat{y}} \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_i - y_i\right)\frac{\partial \hat{y}_i}{\partial \hat{y}} \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_i - y_i\right)\begin{bmatrix}\mathbf{1}_{i=1} \\ \mathbf{1}_{i=2} \\ \vdots \\ \mathbf{1}_{i=m}\end{bmatrix} \\
&= \frac{1}{m}\begin{bmatrix}\hat{y}_1 - y_1 \\ \hat{y}_2 - y_2 \\ \vdots \\ \hat{y}_m - y_m\end{bmatrix} \\
&= \frac{1}{m}(\hat{y} - y)
\end{aligned}$$

5. (3 points) What is the partial derivative of $f(X, w)$ with respect to w. Write your answer using matrices and/or vectors.

Answer:

$$\begin{aligned}
\frac{\partial}{\partial w} f(X, w) &= \begin{bmatrix}\frac{\partial f(X,w)_1}{\partial w} & \frac{\partial f(X,w)_2}{\partial w} & \cdots & \frac{\partial f(X,w)_m}{\partial w}\end{bmatrix} \\
&= \begin{bmatrix}\frac{\partial X_{1,\cdot} w}{\partial w} & \frac{\partial X_{2,\cdot} w}{\partial w} & \cdots & \frac{\partial X_{m,\cdot} w}{\partial w}\end{bmatrix} \\
&= \begin{bmatrix}\frac{\partial X_{1,\cdot} w}{\partial w_1} & \frac{\partial X_{2,\cdot} w}{\partial w_1} & \cdots & \frac{\partial X_{m,\cdot} w}{\partial w_1} \\ \frac{\partial X_{1,\cdot} w}{\partial w_2} & \frac{\partial X_{2,\cdot} w}{\partial w_2} & \cdots & \frac{\partial X_{m,\cdot} w}{\partial w_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_{1,\cdot} w}{\partial w_m} & \frac{\partial X_{2,\cdot} w}{\partial w_m} & \cdots & \frac{\partial X_{m,\cdot} w}{\partial w_m}\end{bmatrix} \\
&= \begin{bmatrix}X_{1,1} & X_{1,2} & \cdots & X_{1,m} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,m}\end{bmatrix} \\
&= X^\top
\end{aligned}$$

6. (2 points) Consider the loss function $l(w) = g\left(f(X, w), y\right)$. What is the gradient of l? Express your answer using matrices/vectors without summations. Note that we did this derivation in class but used summations.

Answer:

$$
\begin{aligned}
\nabla l(w) &= \frac{\partial}{\partial w} g\left(f(X,w),y\right) \\
&= \frac{\partial g\left(f(X,w),y\right)}{\partial f(X,w)} \frac{\partial f(X,w)}{\partial w} \\
&= \frac{1}{m}\left(f(X,w)-y\right) X^\top
\end{aligned}
$$

7. (4 points) Consider the NLL loss function for classification, $g(\hat{y}, y) = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i \ln \sigma(\hat{y}_i) + (1-y_i)\ln\left(1-\sigma(\hat{y}_i)\right)\right]$, where σ is the sigmoid (also called the logistic function). What is the gradient of the loss function $l(w) = g\left(f(X,w),y\right)$? Express your answer using matrices/vectors without summations. Note that we did this derivation in class but used summations.

Answer:

$$
\begin{aligned}
\nabla l(w) &= \frac{\partial}{\partial w} g\left(f(X,w),y\right) \\
&= \frac{\partial g\left(f(X,w),y\right)}{\partial f(X,w)} \frac{\partial f(X,w)}{\partial w} \\
&= -\frac{1}{m}\left(y-\sigma(\hat{y})\right) X^\top
\end{aligned}
$$