Introduction to Machine Learning (CSCI-UA.473): Homework 1

Instructor: Sumit Chopra

September 29, 2021

Submission Instructions

You must typeset the answers using IATFXand compile them into a single PDF file. Name the pdf file as: <Your-NetID>_hw1.pdf. For the programming part of the assignment, complete the Jupyter notebook named HW1.ipynb. Create a ZIP file containing both the PDF file and the completed Jupyter notebook. Name it \langleYour-NetID ⟩ hw1.zip. Submit the ZIP file on Brightspace. The due date is September 27th , **2021, 11 : 59 PM**.

Theory

Question T1: Empirical vs. Expected Cost (10 points)

We approximate the true cost function with the empirical cost function defined by:

$$\mathbb{E}_x[E(g(x), f(x))] = \frac{1}{N}\sum_{i=1}^{N} E\left(g\left(x^i\right), y^i\right)$$

where N is the number of training samples, f is the unknown function, g is the learnable function, y^2 is the label associated with the input x^2.

In the above equation, is it okay to give an equal weight to the cost associated with each training example? Given that we established that not every data x is equally likely, is taking the sum of all per-example costs and dividing by N reasonable? Should we weigh each per-example cost differently, depending on how likely each x is? Justify your answer.

Answer: Since the purpose of the cost function is to measures the discrepancy between the model prediction and the actual value on the training set, it is reasonable not to weigh each sample differently. By the assumption that training samples are independently and identically drawn from the same distribution, each label y^i has a one-to-one relation to a input x^i.

Question T2: Perceptron Learning Algorithm (10 points)

The weight update rule of the Perceptron Learning Algorithm (PLA) is given by:

$$w(t+1) \leftarrow w(t) + y(t)x(t).$$

Prove the following statements:

1. Show that $y(t)w^T(t)x(t) < 0$ (2 points)

   Answer: The Perceptron Learning Algorithm is defined by

   $$h(x) = \text{sign}\left(w^T x\right)$$

   In its weight update procedure, a random sample $(x(t), y(t))$ is picked from the set which is misclassified. Then the weight vector is updated with the following rule $w(t+1) \leftarrow w(t) + y(t)x(t)$.

   If $y(t) > 0$ and $w^T(t)x(t) > 0$, PLA correctly classifies the sample so no weight update should happen, vise versa for $y(t) < 0$ and $w^T(t)x(t) < 0$. Thus $y(t) \neq sign(w(t)^T x(t))$. Therefore, by multiplication, $y(t)w^T(t)x(t) < 0$.

2. Show that $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$ (4 points)

   Answer:

   $$\begin{aligned} y(t)w^T(t+1)x(t) &> y(t)w^T(t)x(t) \\ y(t)(w^T(t) + y(t)x(t))x(t) &> y(t)w^T(t)x(t) \\ y(t)w^T(t)x(t) + y(t)x(t)x(t) &> y(t)w^T(t)x(t) \end{aligned}$$

   For each wrongly-classified $x(t)$, we have $y(t) \neq sign(w(t)^T x(t))$. By convergence, $y(t)x(t)x(t)$. Therefore $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$.

3. Argue that the move from $w(t)$ to $w(t+1)$ is the right move as far as classifying $x(t)$ is concerned. (4 points)

   Answer: Assuming that the data is linearly-separable, a hyper-plane, or vector can be drawn between the two classes. Notice that the product $w^T x$ gives an angle information between w and x such that $\cos\alpha = \frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|\|\mathbf{x}\|}$ The PLA algorithm in its nature classifies x basing on this angle where $\mathbf{w}^T\mathbf{x} > 0 \Rightarrow \cos\alpha > 0 \Rightarrow \alpha < 90$. By adding wrongly-classified $y(t)x(t)$ to $w(t)$, the angle between a positive x and w decreases, vise versa.

Question T3: Gradient of Logistic Regression (10 points)

The logistic regression loss for a single sample (x, y) can be written as

$$\mathcal{L}_w(x, y) = -[y \cdot \log \sigma(wx) + (1 - y) \cdot \log(1 - \sigma(wx))],$$

where $\sigma(s)$ is the logistic function and w are the parameters of the model. Compute the gradient of the above loss function with respect to the parameter vector w. Show all the steps of the derivation.

Answer: Provided that $\nabla_w \sigma = \frac{\partial \sigma(wx)}{\partial w} = \frac{\partial}{\partial w} \frac{1}{1+e^{-wx}} = -\frac{xe^{-wx}}{(1+e^{-wx})^2}$

$$\begin{aligned}
\nabla \mathcal{L}_w(x, y) &= \frac{\partial(-[y \cdot \log \sigma(wx) + (1 - y) \cdot \log(1 - \sigma(wx))])}{\partial w} \\
&= -y \left(\frac{\partial}{\partial w} \log(\sigma(wx)) \right) - (1 - y) \left(\frac{\partial}{\partial w} \log(1 - \sigma(wx)) \right) \\
&= -y \frac{\frac{\partial}{\partial w}\sigma(wx)}{\sigma(wx)} - (1 - y) \left(\frac{\frac{\partial}{\partial w}(1 - \sigma(wx))}{1 - \sigma(wx)} \right) \\
&= -y \frac{\frac{\partial \sigma(wx)}{\partial w}}{\sigma(wx)} - (1 - y) \left(\frac{(1 - \frac{\partial \sigma(wx)}{\partial w})}{1 - \sigma(wx)} \right) \\
&= \boxed{-\frac{xye^{-wx}}{e^{-wx} + 1} + \frac{x(1 - y)e^{-wx}}{(e^{-wx} + 1)^2 \left(1 - \frac{1}{e^{-wx}+1}\right)}}
\end{aligned}$$

Practicum

See the accompanying Python notebook.

Question P1: Linear Regression (20 points)

Question P2: Gradient Descent (10 points)

Question P3: Logistic Regression (40 points)