IE643 Deep Learning : Theory and Practice **July-Dec 2022**

Assignment 1: Due On 25^{th} August 2022 (11:59 PM IST)

1 Instructions

We make the following assumptions for all questions: a vector $u \in \mathbb{R}^d$ for some $d \geq 1$, is represented as $\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{pmatrix}$; a vector $u \in \mathbb{R}^d$ with $u_i = 0 \; \forall i = 1, \ldots, d$ is called a zero vector and is represented by $\mathbf{0}$; the transpose of a vector u is denoted by $u^\top$; the notation $|\alpha|$ denotes the absolute value of some $\alpha \in \mathbb{R}$.

Answer all questions. Write your answers clearly. You can score a maximum of 50 marks in this assignment.

Please make sure that all your answers are present in a single pdf document. If you use python notebook (.ipynb) files, make sure that your answers and plots are visible in .ipynb file. Upload on moodle, the python code, plots and pdf document as a single zip file named as "IE643_rollno_assignment1.zip". All your files within the zip file should follow similar naming convention. There will be no extensions to the submission deadline.

Note: The questions which need to be answered as part of the assignment are provided in Section 2. There are some practice questions in Section 3 which need not be submitted as part of the assignment.

2 Assignment Questions (Solutions need to be submitted)

1. (a) **[5 marks]** Recall that a hyperplane $H = (w, b)$ for some $w \neq \mathbf{0} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ is defined as $H = \{x \in \mathbb{R}^d : \langle w, x \rangle = b\}$. Show that for every hyperplane H and for every $\beta > 0$, there exists another hyperplane $\tilde{H} = (\tilde{w}, \tilde{b})$ such that $\|\tilde{w}\|_2 = \beta$. Illustrate the relationship between (w, b) and $(\tilde{w}, \tilde{b})$. (Recall that for $u \in \mathbb{R}^d$, $\|u\|_2 = \sqrt{\langle u, u \rangle} = \sqrt{\sum_{i=1}^d |u_i|^2}$ is the ℓ_2 norm of u).

   (b) **[5 marks]** Consider a data set $D = \{(x^1, y^1), \ldots, (x^n, y^n)\}$, where $x^j \in \mathbb{R}^d$, $y^j \in \{+1, -1\}, \forall j = 1, 2, \ldots, n$. Let $\max_j \|x^j\|_2 \leq R$. Recall that D is linearly separable if there exist $w^* \in \mathbb{R}^d$ and $\gamma > 0$ such that $y^j \langle w^*, x^j \rangle \geq \gamma, \; \forall j = 1, \ldots, n$. Show that if D is linearly separable, the mistake bound proved in class

$$M \leq \frac{R^2 \|w^*\|_2^2}{\gamma^2}$$

   can be rewritten simply as $M \leq \frac{R^2}{\eta^2}$, where $\eta > 0$ (which might be same as γ or different from γ). (**Hint: Use the previous result about hyperplane.**)

2. Consider the perceptron learning algorithm with a starting point $w^0 = \begin{bmatrix} \theta & \theta & \ldots & \theta \end{bmatrix}^\top$ where $\theta \in [0, 1]$, used to train on a linearly separable data set.

(a) **[7 marks]** Find a suitable upper bound on the number of mistakes for the choice of starting point w^0 given above.

(b) **[3 marks]** Compare and contrast the bound you obtained in part (a) with the bound discussed in class. Explain the changes observed in the bound, and explain the dependence of the bound you obtained in part (a) on the choice of w^0.

(c) **[2 marks]** Justify if your bound is tight.

3. For this exercise, you can use the code you would have written for Lecture 2 Homework. Consider a linearly separable data set D described as follows: generate 100 two dimensional points from a 2 dimensional Gaussian distribution with mean $(-2, -2)$ and variance $\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$ and label these points with $+1$; generate 100 two dimensional points from a 2 dimensional Gaussian distribution with mean $(1, 1)$ and variance $\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$ and label these points with -1. Make sure that the data set is linearly separable. For reproducibility of results, use a seed for random number generator used in your code. Denote the size of data set as $|D|$.

(a) Train a perceptron (until convergence) with a starting point with each coordinate sampled from a Gaussian distribution with zero mean and unit variance. Plot the number of mistakes made vs epochs. Plot the separator line obtained after convergence along with the data. [**5 marks**]

(b) Recall that in perceptron the prediction function is given by:

$$P(x; w) = \text{sign}(\langle w, x \rangle)$$

where $\text{sign}(a) = +1$ if $a \geq 0$ and -1 if $a < 0$. Consider a variant of perceptron called a **T-perceptron** where the prediction function is given by:

$$TP(x; w) = \text{sign}(\sigma(\langle w, x \rangle))$$

where $\sigma(a) = \frac{e^{2a}-1}{e^{2a}+1}$. Let $\sigma'(a)$ denote the derivative of the function $\sigma(a)$. We will use the following algorithm to train a T-perceptron.

Initialize: starting point w^0, $t = 0$, $epochs = 0$, $maxepochs = 100$
while *epochs*<*maxepochs* **do**
  Initialize NUM_MISTAKES = 0
  for $k = 1, 2, \ldots, |D|$ **do**
    Receive (x^k, y^k) from data set D.
    $x^k = (x^k, 1)$ (Append a constant feature to the data features)
    $\hat{y} = TP(x^k; w^t)$.
    if $\hat{y} \neq y^k$ **then**
      NUM_MISTAKES = NUM_MISTAKES+1
    end
    $w^{t+1} = w^t - \eta(y^k - \sigma(\langle w^t, x^k \rangle))\sigma'(\langle w^t, x^k \rangle)x^k$
    $t = t + 1$
  end
  epochs=epochs+1
  if *NUM_MISTAKES equals 0* **then**
    break from the While loop
  end
end

Algorithm 1: Training algorithm for T-Perceptron

i. Implement Algorithm 1 using the same starting point w^0 as that used in your perceptron training. Choose $\eta \in \{0.01, 0.001, 0.0001\}$. [**5 marks**]

ii. In a single graph, plot the number of mistakes vs epochs for different η values along with the corresponding plot obtained for perceptron (use different colors for different η values and use a different color for perceptron). Explain your observations obtained by comparing the plots obtained for T-perceptron with plot obtained from perceptron. [**4 marks**]

iii. For each η, consider a 2D grid of points, and for each point $z = (z_1, z_2)$ in the 2D grid, use a color to represent the point when $\text{sign}(\sigma(\langle w^\eta, z\rangle))$ is $+1$ and use a different color to represent the point when $\text{sign}(\sigma(\langle w^\eta, z\rangle))$ is -1. Note that w^η denotes the final weights obtained for corresponding η values. On each graph, plot the data points and plot the the separating line obtained from perceptron. Explain your observations obtained by comparing the separating surfaces obtained for T-perceptron with the separating line obtained from perceptron. [**6 marks**]

(c) Consider a linearly non-separable data D_1 by making appropriate modifications to the variance matrices used to generate D; in particular, use the variance matrix as $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$. Check if D_1 is indeed linearly non-separable. Verify if Perceptron training does not end even after 100 epochs when trained on D_1. Perform T-Perceptron training on D_1 using Algorithm 1. For each $\eta \in \{0.01, 0.001, 0.0001\}$ check if Algorithm 1 converges within 100 epochs or not. Generate the plots similar to that obtained for data set D and perform a comparative analysis. Explain your observations. Based on your observations, explain if T-perceptron can be useful for linearly non-separable data. [**8 marks**]

3 Practice Questions (Not for submission)

- The inner product (or dot product or scalar product) between two vectors $u, v \in \mathbb{R}^d$ is defined as $\langle u, v\rangle = \sum_{i=1}^{d} u_i v_i$. The following properties are related to the inner product:
  1. Prove $\langle u, v\rangle = \langle v, u\rangle, \forall u, v \in \mathbb{R}^d$.
  2. Prove $\langle u, v + w\rangle = \langle u, v\rangle + \langle u, w\rangle, \forall u, v, w \in \mathbb{R}^d$.
  3. Prove $\langle \alpha u, v\rangle = \alpha\langle u, v\rangle, \forall u, v \in \mathbb{R}^d, \forall \alpha \in \mathbb{R}$.
- A vector norm defined on $\mathbb{R}^d$ is a function $\|\cdot\| : \mathbb{R}^d \longrightarrow [0, +\infty]$, which satisfies the following properties:
  - **[Non-negativity]** $\|u\| \geq 0, \forall u \in \mathbb{R}^d$ and $\|u\| = 0$ if and only if $u = \mathbf{0}$ is the zero vector.
  - **[Absolute scaling]** $\|\alpha u\| = |\alpha|\|u\|, \forall u \in \mathbb{R}^d, \forall \alpha \in \mathbb{R}$.
  - **[Triangle inequality]** $\|u + v\| \leq \|u\| + \|v\|, \forall u, v \in \mathbb{R}^d$.

  The following questions are related to vector norms.
  1. Consider $\|u\|_2 = \sqrt{\langle u, u\rangle} = \sqrt{\sum_{i=1}^{d} u_i^2}$. Show that $\|\cdot\|_2$ is a vector norm. You must verify all the three properties described above. (Note that this is the popular Euclidean norm and is induced by the inner product definition given in the previous question.)
  2. Consider $\|u\|_1 = \sum_{i=1}^{d} |u_i|$. Show that $\|\cdot\|_1$ is a vector norm.
  3. Consider $\|u\|_p = \left[\sum_{i=1}^{d} (u_i)^p\right]^{\frac{1}{p}}$, where $p > 2$. Show that $\|\cdot\|_p$ is a vector norm.
  4. Consider $\|u\|_\infty = \max\{|u_1|, |u_2|, \ldots, |u_d|\}$. Show that $\|\cdot\|_\infty$ is a vector norm.
- Prove the Cauchy-Schwarz inequality $|\langle u, v\rangle| \leq \|u\|_2\|v\|_2, \ \forall u, v \in \mathbb{R}^d$.