

# Assignment Three

## ECE 4200/5420, Fall 2021

- Provide credit to **any sources** other than the course staff that helped you solve the problems. This includes **all students** you talked to regarding the problems.
- You can look up definitions/basics online (e.g., wikipedia, stack-exchange, etc)
- **The due date is 10/01/2021, 23.59.59 Eastern time.**
- Submission rules are the same as previous assignments.

**Problem 1 (10 points).** Recall that a linear classifier is specified by  $(\vec{w}, t)$ , where  $\vec{w} \in \mathbb{R}^d$ , and  $t \in \mathbb{R}$ . The decision rule for a feature vector  $X \in \mathbb{R}^d$  is given by  $\vec{w} \cdot X \leq t$ .

Recall the perceptron algorithm from the lectures. Suppose  $d = 2$ ,  $n = 4$ , and a dataset with four examples:

$i$	feature $X_i$	label $y_i$
1	$(-0.6, 1)$	$-1$
2	$(-3, -4)$	$-1$
3	$(3, -2)$	$+1$
4	$(0.5, 1)$	$+1$

1. In class we started with the the initial  $(\vec{w}, t) = (\vec{0}, 0)$ , and derived the convergence of perceptron. In this problem:

- Start with the initialization  $(\vec{w}, t) = ((0, 1), 0)$  (the  $x$ -axis).
- Implement the perceptron algorithm **by hand**. Go over the data-points **in order**. Output a table of the form below where each row works with one example in some iteration. We have filled in some entries in the first two rows. You need to add rows until no mistakes happen on any example.

starting $\vec{w}$	starting $t$	features $X_i$	label $y_i$	predicted label	new $\vec{w}$	new $t$
$(0, 1)$	0	$(-0.6, 1)$	$-1$	...	...	...
...	...	$(-3, -4)$	$-1$	...	...	...
...	...	...	...	...	...	...

- Draw a 2-D grid. On this grid, mark the four examples (like we do on the board). Draw the line you obtain as the final result.

**Problem 2. (15 points).** Consider the same set-up as in perceptron, where the features all satisfy  $\|X_i\| \leq 1$ , and the training examples are separable with margin  $\gamma$ . We showed in class that perceptron converges with at most  $4/\gamma^2$  updates when initialized to the all zero vector, namely to  $(\vec{0}, 0)$ . Suppose instead the initial start with some initial  $(\vec{w}_0, t_0)$  with  $\|\vec{w}_0\| \leq R$ , and  $|t_0| \leq R$ . We run the perceptron algorithm with this initialization. Suppose,  $(\vec{w}_j, t_j)$  is the hyperplane after  $j$ th update.

Let  $(\vec{w}_{opt}, t_{opt})$  be the optimal hyperplane with margin  $\gamma$  with  $\|\vec{w}_{opt}\|^2 = 1$

1. Similar to the class, show that  $|t_{opt}| \leq 1$ , and therefore conclude that

$$\|\vec{w}_{opt}\|^2 + |t_{opt}|^2 \leq 2.$$

2. Show that

$$(\vec{w}_j, t_j) \cdot (\vec{w}_{opt}, t_{opt}) \geq j\gamma - 2R.$$

3. Show that

$$(\vec{w}_j, t_j) \cdot (\vec{w}_j, t_j) \leq 2j + 2R^2.$$

4. Using these conclude that the **total number of updates** before perceptron converges is at most

$$\frac{4 + 4R\gamma}{\gamma^2}.$$

**Problem 3 (10 points).** Recall the log-likelihood function for logistic regression:

$$J(\vec{w}, t) = \sum_{i=1}^n \log \Pr(y_i | X_i, \vec{w}, t),$$

where  $\Pr(y_i | X_i, \vec{w}, t)$  is the same as defined in the class for logistic regression, and  $y_i \in \{-1, +1\}$ . We will show that this function is concave as a function of  $\vec{w}, t$ .

1. Show that for two real numbers  $a, b$ ,  $\exp(a) + \exp(b) \geq 2 \exp((a+b)/2)$ . (Basically this says that exponential function is convex.)
2. Extending this, show that for any vectors  $\vec{w}_1, \vec{w}_2, \vec{x} \in \mathbb{R}^d$ ,

$$\exp(\vec{w}_1 \cdot \vec{x}) + \exp(\vec{w}_2 \cdot \vec{x}) \geq 2 \exp\left(\frac{(\vec{w}_1 + \vec{w}_2)}{2} \cdot \vec{x}\right).$$

3. Show that  $J(\vec{w}, t)$  is concave (you can only show the concavity holds for  $\lambda = 1/2$ ). You can show it any way you want. One way is to first show that for any  $\vec{w}_1, \vec{w}_2 \in \mathbb{R}^d$ , and  $t_1, t_2 \in \mathbb{R}$ ,

$$\frac{1}{2}J(\vec{w}_1, t_1) + \frac{1}{2}J(\vec{w}_2, t_2) \leq J\left(\frac{\vec{w}_1 + \vec{w}_2}{2}, \frac{t_1 + t_2}{2}\right).$$

You can use that sum of concave functions are concave. A linear function is both concave and convex.

In this problem you can also work with the vector  $\vec{w}^*$ , which is the  $d + 1$  dimensional vector  $(\vec{w}, t)$ , and the  $d + 1$  dimensional feature vectors  $X_i^* = (X_i, -1)$ , which are the features appended with a  $-1$ . Just like in class, this can simplify some of the computations by using the fact that  $\vec{w} \cdot X_i - t = \vec{w}^* \cdot X_i^*$ .

**Problem 4 (10 points) Different class conditional probabilities.** Consider a classification problem with features in  $\mathbb{R}^d$  and labels in  $\{-1, +1\}$ . Consider the class of linear classifiers of the form  $(\vec{w}, 0)$ , namely all the classifiers (hyper planes) that pass through the origin (i.e.,  $t = 0$ ). Instead of logistic regression, suppose the class probabilities are given by the following function:

$$P(y = +1|X, \vec{w}) = \frac{1}{2} \left( 1 + \frac{\vec{w} \cdot X}{\sqrt{1 + (\vec{w} \cdot X)^2}} \right), \quad (1)$$

where  $\vec{w} \cdot X$  is the dot product between  $\vec{w}$  and  $X$ .

Suppose we obtain  $n$  independent examples  $(X_i, y_i)$  for  $i = 1, \dots, n$ .

1. Show that the log-likelihood function is

$$J(\vec{w}) = -n \log 2 + \sum_{i=1}^n \log \left( 1 + \frac{y_i(\vec{w} \cdot X_i)}{\sqrt{1 + (\vec{w} \cdot X_i)^2}} \right). \quad (2)$$

2. Compute the gradient of  $J(\vec{w})$  and write one step of gradient ascent. Namely fill in the blank:

$$\vec{w}_{j+1} = \vec{w}_j + \eta \cdot \underline{\hspace{2cm}}$$

**hint:** use the chain rule and  $\nabla_{\vec{w}} \vec{w} \cdot X = X$ .

**Problem 5 (25 points).** See the attached notebook for details.