Assignment Three Solutions
ECE 4200

- 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)
- Submission rules are the same as previous assignments.
- **Please write your net-id on top of every page. It helps with grading.**

Problem 1 (10 points). Consider a classification problem with d dimensional features. A linear classifier is specified by $(\vec{w}, t)$, and a decision rule $\overrightarrow{w} \cdot X \lesseqgtr t$ for a new feature X.

Recall the perceptron algorithm from the lectures. Suppose $d = 2$, $n = 4$, and there are four training examples, given by:

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 the 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.

Solution.

1. Start with the weight vector $(0, 0, 1)$ (the x-axis). Implement the perceptron algorithm **by hand**. Go over the data-points in order. Output a table of the following form, which shows the updates you make. We have filled in some entries in the first two rows. You need to add rows until no mistakes happen on any example.

starting weight	example	label	predicted label	new weight
$(0, 0, 1)$	$(-1, -0.6, 1)$	-1	$+1$	$(1, 0.6, 0)$
$(1, 0.6, 0)$	$(-1, -3, -4)$	-1	-1	$(1, 0.6, 0)$
$(1, 0.6, 0)$	$(-1, 3, -2)$	$+1$	$+1$	$(1, 0.6, 0)$
$(1, 0.6, 0)$	$(-1, 0.5, 1)$	$+1$	-1	$(0, 1.1, 1)$
$(0, 1.1, 1)$	$(-1, -0.6, 1)$	-1	$+1$	$(1, 1.7, 0)$
$(1, 1.7, 0)$	$(-1, -3, -4)$	-1	-1	$(1, 1.7, 0)$
$(1, 1.7, 0)$	$(-1, 3, -2)$	$+1$	$+1$	$(1, 1.7, 0)$
$(1, 1.7, 0)$	$(-1, 0.5, 1)$	$+1$	-1	$(0, 2.2, 1)$
$(0, 2.2, 1)$	$(-1, -0.6, 1)$	-1	-1	$(0, 2.2, 1)$
$(0, 2.2, 1)$	$(-1, -3, -4)$	-1	-1	$(0, 2.2, 1)$
$(0, 2.2, 1)$	$(-1, 3, -2)$	$+1$	$+1$	$(0, 2.2, 1)$
$(0, 2.2, 1)$	$(-1, 0.5, 1)$	$+1$	$+1$	$(0, 2.2, 1)$

2. 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.

The equation of the separating hyperplane is $2.2x + y = 0$.

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 γ. We showed in class that perceptron converges with at most $4/\gamma^2$ updates when initialized to the all all zero vectors, 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 jth update. Let $(\vec{w}_{opt}, t_{opt})$ be the optimal hyperplane. Then,

Similar to what we did in class, assume that the $d + 1$ dimensional vector $(\vec{w}_{opt}, t_{opt})$ satisfies

$$||(\vec{w}_{opt}, t_{opt})||^2 \leq 2.$$

1. Show that

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

Solution. Recall the perceptron updates:

$$\begin{aligned} \vec{w_j} &= \vec{w}_{j-1} + y_i \overrightarrow{X}_i, \\ t_j &= t_{j-1} - y_i. \end{aligned}$$

Suppose we made a mistake on X_i, y_i for the jth update. Then, similar to class (where we use that $y_i(\vec{w}_{j-1} \cdot X_i - t + j - 1 < 0$,

$$\begin{aligned} (\vec{w_j}, t_j) \cdot (\vec{w}_{opt}, t_{opt}) &= (\vec{w}_{j-1} + y_i X_i, t_{j-1} - y_i) \cdot (\vec{w}_{opt}, t_{opt}) \\ &= (\vec{w}_{j-1}, t_{j-1}) \cdot (\vec{w}_{opt}, t_{opt}) + y_i(\vec{w}_{opt} \cdot X_i - t_{opt}) \\ &\geq (\vec{w}_{j-1}, t_{j-1}) \cdot (\vec{w}_{opt}, t_{opt}) + \gamma. \end{aligned}$$

Inductively following these steps, we obtain,

$$(\vec{w}_j, t_j) \cdot (\vec{w}_{opt}, t_{opt}) \geq (\vec{w}_0, t_0) \cdot (\vec{w}_{opt}, t_{opt}) + j\gamma.$$

Now, by using $\cos(\theta) \leq 1$,

$$[(\vec{w}_0, t_0) \cdot (\vec{w}_{opt}, t_{opt})]^2 \leq \|(\vec{w}_0, t_0)\|_2^2 \cdot \|(\vec{w}_{opt}, t_{opt})\|_2^2 \leq 2 \cdot (2R^2) = 4R^2.$$

Therefore,

$$(\vec{w}_0, t_0) \cdot (\vec{w}_{opt}, t_{opt}) \geq -2R.$$

Plugging this above gives the bound.

2. Show that

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

Solution.

$$\begin{aligned}
(\vec{w}_j, t_j) \cdot (\vec{w}_j, t_j) =& (\vec{w}_{j-1} + y_i X_i, t_{j-1} - y_i) \cdot (\vec{w}_{j-1} + y_i X_i, t_{j-1} - y_i) \\
=& (\vec{w}_{j-1}, t_{j-1}) \cdot (\vec{w}_{j-1}, t_{j-1}) + 2y_i(\vec{w}_{j-1} \cdot X_i - t_{j-1}) + y_i^2 + \|X_i\| \\
\leq& (\vec{w}_{j-1}, t_{j-1}) \cdot (\vec{w}_{j-1}, t_{j-1}) + 2y_i(\vec{w}_{j-1} \cdot X_i - t_{j-1}) + 2 \\
\leq& (\vec{w}_{j-1}, t_{j-1}) \cdot (\vec{w}_{j-1}, t_{j-1}) + 2
\end{aligned}$$

where we used that we made a mistake in the $j-1$th update and that $y_i^2 = 1$. The final step uses the bound on the norm of $(\vec{w}_0, t_0)$.

3. Using these conclude that the number of updates before perceptron converges is at most

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

updates.

Solution. Finally note that

$$\begin{aligned}
[(\vec{w}_j, t_j) \cdot (\vec{w}_{opt}, t_{opt})]^2 &\leq \|(\vec{w}_{opt}, t_{opt})\|_2^2 \cdot \|(\vec{w}_j, t_j)\|_2^2 \\
[j\gamma - 2R]^2 &\leq 2(2j + 2R^2),
\end{aligned}$$

which holds for $j \leq \frac{4+4R\gamma}{\gamma^2}$.

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.)

   Solution. You can do it in many ways, including showing that the double derivative of e^x is always positive, and then using convexity. One simple way is the following.

$$\exp(a) + \exp(b) - 2\exp\left(\frac{a+b}{2}\right) = \left(\exp\left(\frac{a}{2}\right) - \exp\left(\frac{b}{2}\right)\right)^2 \geq 0.$$

2. Extending this, show that for any vectors $\vec{w}_1, \vec{w}_2, x \in \mathbb{R}^d$,

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

   Solution. Let $a = \vec{w}_1 \cdot x$, and $b = \vec{w}_2 \cdot x$, then

$$\frac{\vec{w}_1 + \vec{w}_2}{2} \cdot x = \frac{\vec{w}_1 \cdot x}{2} + \frac{\vec{w}_2 \cdot x}{2} = \frac{a+b}{2}.$$

   Plugging these values in the first part gives the solution.

3. Show that $J(\vec{w}, t)$ is concave. You can show it any way you want. One way is to first show that for any $\vec{w}_1, \overrightarrow{w}_2 \in \mathbb{R}^d$, and $t_1, t_2 \in \mathbb{R}$,

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

   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 $\vec{w} \cdot X_i - t = \vec{w}^* \cdot X_i^*$.

Solution. Recall from the slides and class,

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

Since sum of concave functions are concave it will suffice to shat that each of the n terms in the summation above are concave in $\overrightarrow{w}^*$. In other words, we need to show that

$$\frac{1+y_i}{2}(\vec{w}^* \cdot X_i^*) - \log\left(1 + \exp(\vec{w}^* \cdot X_i^*)\right)$$

is concave in $\vec{w}^*$. Now the first term is linear, and is therefore concave (and convex). Because of the $-$ sign, we need to show that

$$\log\left(1 + \exp(\vec{w}^* \cdot X_i^*)\right)$$

is convex. In particular we will show that

$$\log\left(1+\exp\left(\vec{w}_1^*\cdot X_i^*\right)\right)+\log\left(1+\exp\left(\vec{w}_2^*\cdot X_i^*\right)\right)\geq 2\log\left(1+\exp\left(\frac{\vec{w}_1^*+\vec{w}_2^*}{2}\cdot X_i^*\right)\right).$$

Since log is a monotonically increasing function, it will suffice to show that

$$\left(1+\exp\left(\vec{w}_1^*\cdot X_i^*\right)\right)\cdot\left(1+\exp\left(\vec{w}_2^*\cdot X_i^*\right)\right)\geq\left(1+\exp\left(\frac{\vec{w}_1^*+\vec{w}_2^*}{2}\cdot X_i^*\right)\right)^2.$$

Expanding and canceling the terms, this reduces to proving

$$\exp\left(\vec{w}_1^*\cdot X_i^*\right)+\exp\left(\vec{w}_2^*\cdot X_i^*\right)\geq 2\exp\left(\frac{\vec{w}_1^*+\vec{w}_2^*}{2}\cdot X_i^*\right),$$

which is the previous part.

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 (or $t=0$). Instead of logistic regression, suppose the class probabilities are given by the following function, where $X\in\mathbb{R}^d$ are the features:

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

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

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

1. Show that the log-likelihood function is

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

2. Compute the gradient and write one step of gradient ascent. Namely fill in the blank:

$$\vec{w}_{j+1}=\vec{w}_j+\eta\cdot\_\_\_\_\_$$

Solution:

1. Note that,

$$\begin{aligned}P(y=-1|X,\vec{w})=&1-P(y=+1|X,\vec{w})\\=&1-\frac{1}{2}\left(1+\frac{\vec{w}\cdot X}{\sqrt{1+(\vec{w}\cdot X)^2}}\right)\\=&\frac{1}{2}\left(1-\frac{\vec{w}\cdot X}{\sqrt{1+(\vec{w}\cdot X)^2}}\right).\end{aligned} \tag{3}$$

Therefore, for both $y = +1$, and $y = -1$,

$$P(y|X, \vec{w}) = \frac{1}{2}\left(1 + \frac{y(\vec{w} \cdot X)}{\sqrt{1 + y(\vec{w} \cdot X)^2}}\right), \tag{4}$$

$$\begin{aligned} J(\vec{w}) &= \log \prod_{i=1}^{n} P(y_i|X_i, \vec{w}) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{2}\left(1 + \frac{y_i(\vec{w} \cdot X_i)}{\sqrt{1 + (\vec{w} \cdot X_i)^2}}\right)\right) \\ &= -n \log 2 + \sum_{i=1}^{n} \log\left(1 + \frac{y_i(\vec{w} \cdot X_i)}{\sqrt{1 + (\vec{w} \cdot X_i)^2}}\right). \end{aligned} \tag{5}$$

2. Compute the gradient and write one step of gradient ascent. Namely fill in the blank:

$$\begin{aligned} \nabla(J(\vec{w})) &= \sum_{i=1}^{n} \left(1 + \frac{y_i(\vec{w} \cdot X_i)}{\sqrt{1 + (\vec{w} \cdot X_i)^2}}\right)^{-1} \cdot \frac{y_i \cdot X_i \sqrt{1 + (\vec{w} \cdot X_i)^2} - \frac{y_i X_i \cdot (\vec{w} \cdot X_i)^2}{\sqrt{1 + (\vec{w} \cdot X_i)^2}}}{1 + (\vec{w} \cdot X_i)^2} \\ &= \sum_{i=1}^{n} \left(1 + \frac{y_i(\vec{w} \cdot X_i)}{\sqrt{1 + (\vec{w} \cdot X_i)^2}}\right)^{-1} \cdot \frac{y_i \cdot X_i\big(1 + (\vec{w} \cdot X_i)^2\big) - y_i X_i \cdot (\vec{w} \cdot X_i)^2}{(1 + (\vec{w} \cdot X_i)^2)^{\frac{3}{2}}} \\ &= \sum_{i=1}^{n} \left(1 + \frac{y_i(\vec{w} \cdot X_i)}{\sqrt{1 + (\vec{w} \cdot X_i)^2}}\right)^{-1} \cdot \frac{y_i \cdot X_i}{(1 + (\vec{w} \cdot X_i)^2)^{\frac{3}{2}}} \end{aligned} \tag{6}$$

Therefore,

$$\vec{w}_{j+1} = \vec{w}_j + \eta \cdot \left(\sum_{i=1}^{n} \left(1 + \frac{y_i(\vec{w_j} \cdot X_i)}{\sqrt{1 + (\vec{w_j} \cdot X_i)^2}}\right)^{-1} \cdot \frac{y_i \cdot X_i}{(1 + (\vec{w_j} \cdot X_i)^2)^{\frac{3}{2}}}\right)$$