Jonathan Nusantara jan265 Assignment 4

Problem 1 (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 $(\bar{w}, 0)$, 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 $\bar{X} \in \mathbb{R}^d$ are the features:

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

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

Suppose we obtain n examples $(\bar{X}_i, y_i)$ for $i = 1, \ldots, n$.

1. Show that the log-likelihood function is

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

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

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

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

$$P(y = -1 | \bar{x}, \bar{w}) = \frac{1}{2}\left(1 + \frac{-1(\bar{w} \cdot \bar{x})}{\sqrt{1 + (\bar{w} \cdot \bar{x})^2}}\right)$$

① For $i = 1, \ldots, n$:

$$P(y_1 \ldots y_n | \bar{x}_1 \ldots \bar{x}_n, \bar{w}) = \prod_{i=1}^{n} P(y_i | \bar{x}_i, \bar{w})$$

$$= \prod_{i=1}^{n} \frac{1}{2}\left(1 + \frac{y_i(\bar{w} \cdot \bar{x})}{\sqrt{1 + (\bar{w} \cdot \bar{x})^2}}\right)$$

log ⟶

$$= \sum_{i=1}^{n} \log\left(\frac{1}{2}\left(1 + \frac{y_i(\bar{w} \cdot \bar{x})}{\sqrt{1 + (\bar{w} \cdot \bar{x})^2}}\right)\right)$$

$$= \sum_{i=1}^{n} \log \frac{1}{2} + \sum_{i=1}^{n} \log\left(1 + \frac{y_i(\bar{w} \cdot \bar{x})}{\sqrt{1 + (\bar{w} \cdot \bar{x})^2}}\right)$$

$$= \sum_{i=1}^{n} \log(2^{-1}) + \sum_{i=1}^{n} \log\left(1 + \frac{y_i(\bar{w} \cdot \bar{x})}{\sqrt{1 + (\bar{w} \cdot \bar{x})^2}}\right)$$

$$J(\bar{w}) = -n \log 2 + \sum_{i=1}^{n} \log\left(1 + \frac{y_i(\bar{w} \cdot \bar{x})}{\sqrt{1 + (\bar{w} \cdot \bar{x})^2}}\right) \quad \text{(proven)}$$

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

Find $\nabla J(\bar{w})$

$$\nabla J(\bar{w}) = \frac{d}{d\bar{w}} \sum_{i=1}^{n} \log\left(1 + \frac{y_i(\bar{w}\cdot\bar{x}_i)}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}\right) + \frac{d}{d\bar{w}}(-n\log 2) \quad (\frac{d}{d\bar{w}}(-n\log 2) \to 0)$$

$$= \sum_{i=1}^{n} \frac{1}{1 + \frac{y_i(\bar{w}\cdot\bar{x}_i)}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}} \cdot \frac{\left(\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}\right)\cdot y_i(\bar{x}_i) - \frac{y_i(\bar{w}\cdot\bar{x}_i)(\bar{x}_i(\bar{w}\cdot\bar{x}_i))}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}}{(1+(\bar{w}\cdot\bar{x}_i)^2)}$$

$$= \sum_{i=1}^{n} \frac{1}{1 + \frac{y_i(\bar{w}\cdot\bar{x}_i)}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}} \cdot \frac{y_i(\bar{x}_i)\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2} - \frac{y_i(\bar{x}_i)(\bar{w}\cdot\bar{x}_i)^2}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}}{(1+(\bar{w}\cdot\bar{x}_i)^2)}$$

$$= \sum_{i=1}^{n} \frac{1}{1 + \frac{y_i(\bar{w}\cdot\bar{x}_i)}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}} \cdot \frac{y_i(\bar{x}_i)\left(\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2} - \frac{(\bar{w}\cdot\bar{x}_i)^2}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}\right)}{(1+(\bar{w}\cdot\bar{x}_i)^2)}$$

$$= \sum_{i=1}^{n} \frac{1}{1 + \frac{y_i(\bar{w}\cdot\bar{x}_i)}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}} \cdot \frac{y_i(\bar{x}_i)\left(\frac{1+(\bar{w}\cdot x_i)^2 - (\bar{w}\cdot x_i)^2}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}\right)}{(1+(\bar{w}\cdot\bar{x}_i)^2)}$$

$$= \sum_{i=1}^{n} \frac{1}{1 + \frac{y_i(\bar{w}\cdot\bar{x}_i)}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}} \cdot \frac{y_i(\bar{x}_i)}{(1+(\bar{w}\cdot\bar{x}_i)^2)\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}$$

$$= \sum_{i=1}^{n} \frac{1}{1 + \frac{y_i(\bar{w}\cdot\bar{x}_i)}{\sqrt{1+(\bar{w}\cdot\bar{x}_i)^2}}} \cdot \frac{y_i(\bar{x}_i)}{(1+(\bar{w}\cdot\bar{x}_i)^2)^{3/2}}$$

$$= \sum_{i=1}^{n} \frac{y_i(x_i)}{(1+(\bar{w}\cdot\bar{x}_i)^2)^{\frac{3}{2}} + y_i(\bar{w}\cdot\bar{x}_i)(1+(\bar{w}\cdot\bar{x}_i)^2)}$$

$$\bar{w}_{j+1} = \bar{w}_j + \eta * \sum_{i=1}^{n} \frac{y_i(\bar{x}_i)}{(1+(\bar{w}\cdot\bar{x}_i)^2)^{3/2} + y_i(\bar{w}\cdot\bar{x}_i)(1+(\bar{w}\cdot\bar{x}_i)^2)}$$

In **Problem 2**, and **Problem 3**, we will study linear regression. We will assume in both the problems that $w^0 = 0$. (This can be done by translating the features and labels to have mean zero,

but we will not worry about it). For $\bar{w} = (w^1, \ldots, w^d)$, and $\bar{X} = (\bar{X}^1, \ldots, \bar{X}^d)$, the regression we want is:

$$y = w^1 \bar{X}^1 + \ldots + w^d \bar{X}^d = \bar{w} \cdot \bar{X}. \tag{3}$$

We considered the following regularized least squares objective, which is called as **Ridge Regression**. For the dataset $S = \{(\bar{X}_1, y_1), \ldots, (\bar{X}_n, y_n)\}$,

$$J(\bar{w}, \lambda) = \sum_{i=1}^{n} \left(y_i - \bar{w} \cdot \bar{X}_i\right)^2 + \lambda \cdot \|\bar{w}\|_2^2.$$

Problem 2 (10 points) Gradient Descent for regression.

1. Instead of using the closed form expression we mentioned in class, suppose we want to perform gradient descent to find the optimal solution for $J(\bar{w})$. Please compute the gradient of J, and write one step of the gradient descent with step size η.
2. Suppose we get a new point $\bar{X}_{n+1}$, what will the predicted y_{n+1} be when $\lambda \to \infty$?

① $\nabla_w J(\bar{w}, \lambda) = \frac{d}{d\bar{w}} \sum_{i=1}^{n} (y_i - \bar{w} \cdot \bar{x}_i)^2 + \frac{d}{dw} \lambda \cdot \|\bar{w}\|_2^2$

$= \sum_{i=1}^{n} \left(2(y_i - \bar{w} \cdot \bar{x}_i)(-\bar{x}_i)\right) + 2\lambda \bar{w}$

$= -2 \sum_{i=1}^{n} \bar{x}_i (y_i - \bar{w} \cdot \bar{x}_i) + 2\lambda \bar{w}$

$w_{j+1} \longleftarrow w_j - \eta * \left(2 \sum_{i=1}^{n} \bar{x}_i (y_i - \bar{w} \cdot \bar{x}_i) + 2\lambda \bar{w}\right)$

or write in matrix

$\nabla_w J(\bar{w}, \lambda) = \nabla \left[(y - \bar{x} \cdot \bar{w})^T (y - \bar{x}\bar{w})\right] + \nabla \left[\lambda \bar{w}^T \bar{w}\right]$

$= -2\bar{x}^T \left[y - \bar{x}\bar{w}\right] + 2\lambda \bar{w}$

$w_{j+1} \longleftarrow w_j - \eta * \left(2\bar{x}^T \left[y - \bar{x}\bar{w}\right] + 2\lambda \bar{w}\right)$

(2) So we have $$J(\bar{w}, \lambda) = \sum_{i=1}^{n} (y_i - \bar{w} \cdot \bar{X}_i)^2 + \lambda \cdot \|\bar{w}\|_2^2.$$

and our goal is to minimize $J(\bar{w}, \lambda)$ and find $\bar{w}$.

If $\lambda \to \infty$, than our w can only be 0, so that we can minimize $J(\bar{w}, \lambda)$.

Our prediction $y = \underbrace{w_0}_{=0} + \bar{w}_1 \bar{x}_1 + \bar{w}_2 \bar{x}_2 + \dots + \bar{w}_n \bar{x}_n$

$$= \bar{w} \cdot \bar{x}$$

If $\bar{w} = 0$

then $y = 0 \cdot \bar{x}$

$$= \underline{\underline{0}}$$

Our predicted $y = 0$, which is wrong and this is because our model is very underfitting and has too much smoothing (λ too big)

Problem 3 (15 points) Regularization increases training error. In the class we said that when we regularize, we expect to get weight vectors with smaller, but never proved it. We also displayed a plot showing that the training error increases as we regularize more (larger λ). In this assignment, we will formalize the intuitions rigorously.

Let $0 < \lambda_1 < \lambda_2$ be two regularizer values. Let $\bar{w}_1$, and $\bar{w}_2$ be the minimizers of $J(\bar{w}, \lambda_1)$, and $J(\bar{w}, \lambda_2)$ respectively.

1. Show that $\|\bar{w}_1\|_2^2 \geq \|\bar{w}_2\|_2^2$. Therefore more regularization implies smaller norm of solution!
   Hint: Observe that $J(\bar{w}_1, \lambda_1) \leq J(\bar{w}_2, \lambda_1)$, and $J(\bar{w}_2, \lambda_2) \leq J(\bar{w}_1, \lambda_2)$ (why?).

2. Show that the training error for $\bar{w}_1$ is less than that of $\bar{w}_2$. In other words, show that

$$\sum_{i=1}^{n} \left(y_i - \bar{w}_1 \cdot \bar{X}_i\right)^2 \leq \sum_{i=1}^{n} \left(y_i - \bar{w}_2 \cdot \bar{X}_i\right)^2.$$

$$J(\bar{w}, \lambda) = \sum_{i=1}^{n} \left(y_i - \bar{w} \cdot \bar{X}_i\right)^2 + \lambda \cdot \|\bar{w}\|_2^2.$$

   Hint: Use the first part of the problem.

① Using the hint:

Proof that $J(\bar{w}_1, \lambda_1) \leq J(\bar{w}_2, \lambda_1)$

Since $\bar{w}_1$ is the minimizer of $J(\bar{w}, \lambda_1)$, if we use $\bar{w}_2$ in $J(w, \lambda_1)$, then $J(\bar{w}_1, \lambda_1) \leq J(\bar{w}_2, \lambda_1)$.

Any $\bar{w}_n$ used beside the minimizer, will be $J(\bar{w}_1, \lambda_1) \leq J(\bar{w}_n, \lambda_1)$

The same goes to $\bar{w}_2$ and $J(\bar{w}_2, \lambda_2)$.

So, we proved $J(\bar{w}_1, \lambda_1) \leq J(\bar{w}_2, \lambda_1)$ and $J(\bar{w}_2, \lambda_2) \leq J(\bar{w}_1, \lambda_2)$

$$J(\bar{w}_1, \lambda_1) \leq J(\bar{w}_2, \lambda_1)$$

$$\sum_{i=1}^{n} (y_i - w_1 \cdot \bar{x}_i)^2 + \lambda_1 \cdot \|\bar{w}_1\|_2^2 \leq \sum_{i=1}^{n} (y_i - \bar{w}_2 \cdot \bar{x}_i)^2 + \lambda_1 \cdot \|\bar{w}_2\|_2^2$$

$$J(\bar{w}_2, \lambda_2) \leq J(\bar{w}_1, \lambda_2)$$

$$\sum_{i=1}^{n} (y_i - w_2 \cdot \bar{x}_i)^2 + \lambda_2 \cdot \|\bar{w}_2\|_2^2 \leq \sum_{i=1}^{n} (y_i - w_1 \cdot \bar{x}_i)^2 + \lambda_2 \cdot \|\bar{w}_1\|_2^2$$

$$J(\bar{w}_1, \lambda_1) + J(\bar{w}_2, \lambda_2) \leq J(\bar{w}_2, \lambda_1) + J(\bar{w}_1, \lambda_2)$$

$$\sum_{i=1}^{n} (y_i - w_1 \cdot \bar{x}_i)^2 + \lambda_1 \cdot \|\bar{w}_1\|_2^2 + \sum_{i=1}^{n} (y_i - w_2 \cdot \bar{x}_i)^2 + \lambda_2 \cdot \|\bar{w}_2\|_2^2$$

$$\leq \sum_{i=1}^{n} (y_i - w_2 \cdot \bar{x}_i)^2 + \lambda_1 \cdot \|\bar{w}_2\|_2^2 + \sum_{i=1}^{n} (y_i - w_1 \cdot \bar{x}_i)^2 + \lambda_2 \cdot \|\bar{w}_1\|_2^2$$

$$\lambda_1 \cdot \|\bar{w}_1\|_2^2 + \lambda_2 \cdot \|\bar{w}_2\|_2^2 \leq \lambda_1 \cdot \|\bar{w}_2\|_2^2 + \lambda_2 \cdot \|\bar{w}_1\|_2^2$$

$$\lambda_1 \cdot \|\bar{w}_1\|_2^2 - \lambda_2 \cdot \|\bar{w}_1\|_2^2 \leq \lambda_1 \cdot \|\bar{w}_2\|_2^2 - \lambda_2 \cdot \|\bar{w}_2\|_2^2$$

$$(\lambda_1 - \lambda_2) \cdot \|\bar{w}_1\|_2^2 \leq (\lambda_1 - \lambda_2) \cdot \|\bar{w}_2\|_2^2$$

Since $\lambda_2 > \lambda_1$ ->

$$-\|\bar{w}_1\|_2^2 \leq -\|\bar{w}_2\|_2^2$$

← flip sign

$$\boxed{\|\bar{w}_1\|_2^2 \geq \|\bar{w}_2\|_2^2} \quad \text{(proven)}$$

② Show that $\sum_{i=1}^{n} (y_i - \bar{w}_1 \cdot \bar{X}_i)^2 \leq \sum_{i=1}^{n} (y_i - \bar{w}_2 \cdot \bar{X}_i)^2$.

$$J(\bar{w}_1, \lambda_1) \leq J(\bar{w}_2, \lambda_1)$$

$$\sum_{i=1}^{n} (y_i - w_1 \cdot \bar{x}_i)^2 + \lambda_1 \cdot \|\bar{w}_1\|_2^2 \leq \sum_{i=1}^{n} (y_i - \bar{w}_2 \cdot \bar{x}_i)^2 + \lambda_1 \cdot \|\bar{w}_2\|_2^2$$

From ①, we know that $\boxed{\|\bar{w}_1\|_2^2 \geq \|\bar{w}_2\|_2^2}$

So, $\lambda_1 \cdot \|\bar{w}_1\|_2^2 \geq \lambda_1 \|\bar{w}_2\|_2^2$

Thus, $\leq \lambda_1 \cdot \|\bar{w}_1\|_2^2$, so we can make the substitution.

$$\sum_{i=1}^{n} (y_i - w_1 \cdot \bar{x}_i)^2 + \overbrace{\lambda_1 \cdot \|\bar{w}_2\|_2^2} \leq \sum_{i=1}^{n} (y_i - \bar{w}_2 \cdot \bar{x}_i)^2 + \lambda_1 \cdot \|\bar{w}_2\|_2^2$$

$$\boxed{\sum_{i=1}^{n} (y_i - \bar{w}_1 \cdot \bar{x}_i)^2 \leq \sum_{i=1}^{n} (y_i - \bar{w}_2 \cdot \bar{x}_i)^2} \quad \text{(proven)}$$