Assignment Four
ECE 4950

- 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 3/01/2019, 23.59.59 ET**.
- 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) 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 $(\overrightarrow{w}, 0)$, 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 $\overrightarrow{X} \in \mathbb{R}^d$ are the features:

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

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

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

1. Show that the log-likelihood function is

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

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

$$\overrightarrow{w}_{j+1} = \overrightarrow{w}_j + \eta \cdot \underline{\hspace{3cm}}$$

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 $\overrightarrow{w} = (w^1, \ldots, w^d)$, and $\overrightarrow{X} = (\overrightarrow{X}^1, \ldots, \overrightarrow{X}^d)$, the regression we want is:

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

We considered the following regularized least squares objective, which is called as **Ridge Regression**. For n examples $(\overrightarrow{X}_i, y_i)$,

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

Problem 2 (10 points) Gradient Descent for regression.

1. Instead of using the closed form expression we derived in the class, suppose we want to perform gradient descent to find the optimal solution for $J(\overrightarrow{w})$. 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 $\overrightarrow{X}_{n+1}$, what will the predicted y_{n+1} be when $\lambda \to \infty$?

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 $\overrightarrow{w}_1$, and $\overrightarrow{w}_2$ be the minimizers of $J(\overrightarrow{w}, \lambda_1)$, and $J(\overrightarrow{w}, \lambda_2)$ respectively.

1. Show that $\|\overrightarrow{w}_1\|_2^2 \geq \|\overrightarrow{w}_2\|_2^2$. Therefore more regularization implies smaller norm of solution!

   Hint: Observe that $J(\overrightarrow{w}_1, \lambda_1) \leq J(\overrightarrow{w}_2, \lambda_1)$, and $J(\overrightarrow{w}_2, \lambda_2) \leq J(\overrightarrow{w}_1, \lambda_2)$ (why?).

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

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

   Hint: Use the first part of the problem.

Problem 4 (25 points) Linear and Quadratic Regression. Please refer to the Jupyter Notebook in the assignment, and complete the coding part in it! You can use sklearn regression package: `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html`