Assignment Four
ECE 4200/5420

September 30, 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.
- Questions marked with an asterisk **is optional for students taking the class as ECE 4200**.
- You can look up definitions/basics online (e.g., wikipedia, stack-exchange, etc).
- **The due date is 10/08/2021, 23.59.59 ET**.
- Submission rules are the same as previous assignments.

In **Problems 1, 2, and 3** we will study linear regression. We will assume in these problems that $w^0 = 0$. (This can be done by centering the features and labels to have mean 0, but we will not worry about it). For $\bar{w} = (w^1, \ldots, w^d)$, and $X = (X^1, \ldots, X^d)$, the regression we want is:

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

We considered the following regularized least squares objective, which is called as **Ridge Regression**. For the dataset $S = \{(X_1, y_1), \ldots, (X_n, y_n)\}$,

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

We then find a $\bar{w}$ to minimize $J(\bar{w}, \lambda)$, namely

$$\arg\min_{\bar{w}} J(\bar{w}, \lambda). \tag{3}$$

Problem 1 (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 X_{n+1}, what will the predicted y_{n+1} be when $\lambda \to \infty$?

Problem 2 (15 points) Regularization increases training error. Note the two terms in $J(\bar{w}, \lambda)$. The first is the training error, and the second is the regularization. When $\lambda = 0$ then we minimize just the training error. As λ increases, the second term will become more and more dominant, and this means two things: (1) We are giving more weight to the regularization term and therefore, the training error should perhaps decrease, and (2) and as λ increases we should obtain a $\bar{w}$ with smaller norm. We will formalize both these things rigorously below.

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

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

   This shows that as we regularize more, the training error grows. **Hint:** Use the first part of the problem.

Problem 3* (15 points) Ridge regression $\equiv$ MAP estimation with Gaussian prior. In class we provided a Maximum Likelihood (ML) interpretation of least square regression without regularization (i.e., $\lambda = 0$) under the Gaussian noise model.

The Gaussian noise model is $y_i = \bar{w} \cdot \bar{X}_i + N(0, \sigma^2)$, where $N(0, \sigma^2)$ is a Gaussian distribution with mean 0 and variance σ^2. In other words,

$$p\big(y_i - \bar{w} \cdot \bar{X}_i = \nu | \bar{w}, \bar{X}\big) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\nu^2}{2\sigma^2}\right). \tag{4}$$

We proved in class that Equation (2) with $\lambda = 0$ is equivalent to Maximum Likelihood estimation under the Gaussian noise model.

The goal of this problem is to show that least squares regression with regularization is equivalent to Maximum Aposteriori (MAP) estimation under a Gaussian prior over $\bar{w}$.

Suppose the noise model is still Gaussian (Equation (4)). Furthermore, suppose that $\bar{w}$ has a **prior distribution** that is Gaussian with mean 0 and variance σ^2/λ, namely

$$p(\bar{w}) = \frac{1}{(2\pi\sigma^2/\lambda)^{d/2}} \exp\left(-\frac{\lambda\|\bar{w}\|_2^2}{2\sigma^2}\right).$$

1. Show that the MAP estimator under Gaussian noise model and Gaussian prior:

$$\arg\max_{\bar{w}} p(w|y_1, \ldots, y_n, X_1, \ldots, X_n) \tag{5}$$

   is equivalent to solving (3).

   Hint: Start by noting that

$$\arg\max_{\bar{w}} p(\bar{w}|y_1, \ldots, y_n, X_1, \ldots, X_n) = \arg\max_{\bar{w}} p(y_1, \ldots, y_n|X_1, \ldots, X_n, \bar{w}) \cdot p(\bar{w}).$$

   Now the first term is what we had in class for ML interpretation, and the second term is what will introduce the regularization.

2. What is the MAP estimator $\bar{w}$ as $\lambda \to \infty$, namely when the prior distribution of $\bar{w}$ is Gaussian with mean 0 and variance close to 0?

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`