Homework 3

4190.408 001, Artificial Intelligence

October 30, 2024

1 Instruction

In this homework, you are encouraged to study basic tools of linear algebra, probability, and optimization. Check the following concepts and address their meaning, algorithms, applications, and so on. You may search on the internet, but you should not copy and paste whole sentences. Please make your answers as concise as possible, with only essential parts (each answer should be within a few lines).

- In this homework you are to solve 4 questions and submit a report via Etl. Please name your report, {studentID}_{first name}_{last name}.pdf, e.g.2023-26403_Taeksoo_Kim. Please write your name in English. **Only the PDF generated with LaTeX is accepted.**
- Collaboration is allowed. Discussions are encouraged but you should think about the problems on your own.
- Even when you collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.
- Make sure you cite your work/collaborators at the end of the homework.
- **Honor Code**: This document is exclusively for Fall 2024, 4190.408 students with Professor Hanbyul Joo at Seoul National University. Any student who references this document outside of that course during that semester (including any student who retakes the course in a different semester), or who shares this document with another student who is not in that course during that semester, or who in any way makes copies of this document (digital or physical) without consent of Professor Hanbyul Joo is guilty of cheating, and therefore subject to penalty according to the Seoul National University Honor Code.

2 Linear Regression

Let's say we have N training data

$$\{ (\mathbf{x_1}, t_1), (\mathbf{x_2}, t_2), \cdots, (\mathbf{x_N}, t_N) \},$$

where $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N} \in \mathbb{R}^M$ are input vectors and $t_1, t_2, \ldots, t_N \in \mathbb{R}$ are target output values. Consider the linear model defined as below

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_M x_M = \mathbf{w^T} \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}.$$

Let $\bar{\mathbf{x}} = \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}$, $\mathbf{w} = \begin{bmatrix} w_0 & w_1 & \cdots & w_M \end{bmatrix}^T$ and we can find the best $\mathbf{w}$ that minimizes the Euclidean Loss by solving the optimization problem below

$$\min_{\mathbf{w}} \frac{1}{2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^T \bar{\mathbf{x}}_n \right)^2.$$

(a) Finding the analytic solution of the above optimization problem leads to solving the linear system

$$\mathbf{A}\mathbf{w} = \mathbf{b}.$$

Find the expressions for $\mathbf{A}, \mathbf{b}$ in terms of $\bar{\mathbf{x}}_n, t_n$.
Clarification. You may have to differentiate the above loss and set it to 0. Then clearing up the equation will lead to the form of $\mathbf{A}\mathbf{w} = \mathbf{b}$.

(b) Assume we have two training data $(\bar{\mathbf{x}}_1, t_1) = (\begin{bmatrix} 1 \\ 0 \end{bmatrix}, 1), (\bar{\mathbf{x}}_2, t_2) = (\begin{bmatrix} 1 \\ \epsilon \end{bmatrix}, 1)$, where $\epsilon \neq 0$.
Find $\mathbf{w}$ by solving the linear system in (a).

(c) Assume we have two training data $(\bar{\mathbf{x}}_1, t_1) = (\begin{bmatrix} 1 \\ 0 \end{bmatrix}, 1 + \epsilon), (\bar{\mathbf{x}}_2, t_2) = (\begin{bmatrix} 1 \\ \epsilon \end{bmatrix}, 1)$, where $\epsilon \neq 0$.
Find $\mathbf{w}$ by solving the linear system in (a).

(d) Let $\epsilon = 0.1$ and calculate the difference of $\mathbf{w}$ from (b) and (c).

3 Linear Regression with Regularization

Consider the linear regression setup in **problem 2**. To regularize our weight $\mathbf{w}$ close to 0, we can add a regularization term $\frac{\lambda}{2}\mathbf{w^T w}$ to the objective, where $\lambda > 0$. Then, the linear system defined in 2-(a) becomes

$$(\mathbf{A} + \lambda\mathbf{I})\mathbf{w} = \mathbf{b}.$$

(a) The **spectral radius** of the $N \times N$ square matrix $\mathbf{X}$ is defined as

$$\rho(\mathbf{X}) = \max\{|\lambda_1|, |\lambda_2|, \cdots, |\lambda_N|\},$$

where $\lambda_1, \lambda_2, \cdots, \lambda_N$ are the eigenvalues of $\mathbf{X}$. Show that $\rho((\mathbf{A} + \lambda\mathbf{I})^{-1}) \leq \frac{1}{\lambda}$.
(*Hint: Show* $\mathbf{A}$ *is a positive semidefinite matrix.*)

(b) Let $\epsilon = 0.1, \lambda = 0.05$. Find $\mathbf{w}$ with the two training data in **Problem 2-(b)** and **Problem 2-(c)** respectively. Then, calculate the difference of $\mathbf{w}$ from these two setups.

(c) Compare the difference of $\mathbf{w}$ in **Problem 2-(d)** and **Problem 3-(b)**.
With this observation, what benefit can we expect from adding the regularization term?
(*Hint: We can consider* ϵ *as small noise in training data.*)

4 LR with Regularization: A Probabilistic Perspective

In the class, we learned the probabilistic perspective on linear regression. Assuming we have the training data from the linear function perturbed with Gaussian noise, finding the best $\mathbf{w}$ by **maximum likelihood** is equivalent to the **linear regression**. To clarify,

$$\begin{aligned} t &= \mathbf{w}^T\bar{\mathbf{x}} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \\ \mathbf{w}^* &= \underset{\mathbf{w}}{\text{argmax}} \Pr(\mathbf{t}|\bar{\mathbf{X}}, \mathbf{w})) \\ &= \underset{\mathbf{w}}{\text{argmax}} \prod_n \exp(-\frac{(t_n - \mathbf{w}^T\bar{\mathbf{x}}_n)^2}{2\sigma^2}) \\ &= \cdots \\ &= \underset{\mathbf{w}}{\text{argmin}} \sum_n (t_n - \mathbf{w}^T\bar{\mathbf{x}}_n)^2 \end{aligned} \tag{1}$$

Instead of using likelihood, we can exploit **the posterior probability** with some prior on our weight $\mathbf{w}$. From the Bayes' rule, the posterior

$$\Pr(\mathbf{w}|\mathbf{X}, \mathbf{y}) \propto \Pr(\mathbf{w})\Pr(\mathbf{y}|\mathbf{X}, \mathbf{w}).$$

Consider the Gaussian prior is given for $\mathbf{w}$, $\Pr(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \frac{1}{\lambda}\mathbf{I})$. Show that finding the best $\mathbf{w}$ by maximizing the posterior is equivalent to **linear regression with L2 regularization**.
(Remark: The Gaussian prior can be considered as enforcing $\mathbf{w}$ *to* 0 *in a probabilistic way.)*

5 Logistic Regression

As you learned in our class, logistic regression is a form of **classification** in spite of the word 'regression' in its name. It can be viewed as regression because classification is done by estimating the class probability of given data $\Pr(c_k|x)$, which is a continuous function of input x. For binary classification, for example, we use a sigmoid function $\sigma(\cdot)$ for $\Pr(c_k|x)$ as below,

$$\Pr(c_0|\mathbf{x}) = \sigma(\mathbf{w}^T\bar{\mathbf{x}})$$

$$\Pr(c_1|\mathbf{x}) = 1 - \sigma(\mathbf{w}^T\bar{\mathbf{x}}).$$

Then, the loss function to find the best $\mathbf{w}$ can be expressed as below

$$L(\mathbf{w}) = -\sum_n \{t_n \ln \sigma(\mathbf{w}^T\bar{\mathbf{x}}_n) + (1-t_n)\ln(1-\sigma(\mathbf{w}^T\bar{\mathbf{x}}_n)\}.$$

We may differentiate the objective function and set the derivative to 0 to find the closed form solution as we did for **linear regression**. By doing this, explain why we cannot get the closed form solution of $\mathbf{w}$. (Hint: $\frac{\partial\sigma(x)}{\partial x} = \sigma(x)(1-\sigma(x))$)