Assignment 1: Linear Models Regression, Classification and Regularization

Due Oct 9 at 11:59pm

This assignment is to be done individually.

Important Note: The university policy on academic dishonesty (cheating) will be taken very seriously in this course. You may not provide or use any solution, in whole or in part, to or by another student.

You are encouraged to discuss the concepts involved in the questions with other students. If you are in doubt as to what constitutes acceptable discussion, please ask! Further, please take advantage of office hours offered by the instructor and the TA if you are having difficulties with this assignment.

DO NOT:

- Give/receive code or proofs to/from other students
- Use Google to find solutions for assignment

DO:

- Meet with other students to discuss assignment (it is best not to take any notes during such meetings, and to re-work assignment on your own)
- Use online resources (e.g. Wikipedia) to understand the concepts needed to solve the assignment.

Submitting Your Assignment

The assignment must be submitted online at `https://canvas.sfu.ca/`. You must submit the following file:

1. An assignment report in **PDF format**, named `report.pdf`. This report should contain the solutions to questions 1-3.

1 Linear Regression

Optimal function (2 marks)

Given a joint probability distribution model P_{XY} for X and Y, find the optimal function, $f^* : X \to Y$ which minimizes the given Mean Squared Error:

$$MSE \; = \; \mathbb{E}[(f(X) - Y)^2] \tag{1}$$

Please note that a detailed step by step derivation is required while solving for f^*.

Gaussian Noise Regression Model (3 marks)

Consider a simple regression model with a noise variable ϵ as follows:

$t = y(x, w) + \epsilon$, where $y(x, w) = \boldsymbol{w}^T \boldsymbol{\phi}(x)$; ϕ is the basis function and w is the coefficients vector. Assume that ϵ has a Gaussian Distribution.

Now consider a data set of inputs X = $x_1, ..., x_N$ with corresponding target values $t_1, ..., t_N$. Unlike Gaussian noise with the same value of variance at every training data point, consider that there is a different value of noise variance at each training data point.

The probability density function with different variances is given as below:

$$\boldsymbol{p}(\mathbf{t}|\mathbf{X}, \mathbf{w},\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathbf{T}}\phi(\mathbf{x_n}), \beta_n^{-1}) \tag{2}$$

where β_n is the inverse variance.

In this scenario, solve the following:

(a) Derive the log likelihood of $\boldsymbol{p}(\mathbf{t}|\mathbf{X}, \mathbf{w},\beta)$.

(b) Comment on the relationship between the sum of squares error function and the log likelihood of $\boldsymbol{p}(\mathbf{t}|\mathbf{X}, \mathbf{w},\beta)$.

Weighted Linear Regression (2 marks)

Given training data of the form: $\mathcal{D} = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i)\}, i = 1, 2, \ldots, N$, where $\mathrm{x}_i \in \mathcal{R}^{1\times M}$, i.e. $\mathrm{x}_i = (x_{i,1}, \cdots, x_{i,M})^{\mathrm{T}}, y_i \in \mathcal{R}, \mathbf{X} \in \mathcal{R}^{N\times M}$, where row i of $\mathbf{X}$ is $\mathbf{x}_i^{\mathrm{T}}$, and $\mathrm{y} = (y_1, \cdots, y_N)^{\mathrm{T}}$, linear regression assumes a parametric model of the form: $y_i = \mathrm{x}_{\mathrm{i}}^{\mathrm{T}}\beta + \epsilon_i$ where ϵ_i are noise terms from a given distribution, and seeks to find the parameter vector β that provides the best of fit of the above regression model. One criteria to measure fitness, is to find β that minimizes a given loss function $\mathcal{J}(\beta)$. In class, we have shown that if we take the loss function to be the square-error, i.e.:

$$\mathcal{J}_1(\beta) = \sum_i \left(y_i - \mathrm{x}_{\mathrm{i}}^{\mathrm{T}}\beta\right)^2 = (\mathbf{X}\beta - \mathrm{y})^{\mathrm{T}}(\mathbf{X}\beta - \mathrm{y})$$

Then,

$$\beta^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{3}$$

Moreover, we have also shown that if we assume that $\epsilon_1, \ldots, \epsilon_N$ IID and sampled from the same zero mean Gaussian that is, $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ then the least square estimate is also the MLE estimate for $p(\mathrm{y} \mid \mathbf{X}; \beta)$.

(a) [**1 mark**] Now we assume that $\epsilon_1, \ldots, \epsilon_N$ are independent but each $\epsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right)$. Write down the formula for calculating the MLE of β.

(b) [**1 mark**] Calculate the MLE of β based on question (a). Please show detailed derivation.

2 Regularization

Based on the background in *Weighted Linear Regression*, we assume that the noise terms are IID distributed according to $\mathcal{N}\left(0, \sigma^2\right)$. Also we assume that the number of features M is much larger than the number of training instances N (i.e., $M \gg N$).

(a) [**1 mark**] Explain why in this situation, we can NOT compute β according to (3).

Hints: (1) If matrix A is $m \times n$ then rank$(A) \leq \min(n, m)$. (2) An $n \times n$ matrix A is invertible iff it is full-rank, i.e rank$(A) = n$

(b) [**2 marks**] Instead of minimizing $\mathcal{J}(\beta)$, we minimize the following loss function:

$$\mathcal{J}_R(\beta) = \sum_i \left(y_i - \mathrm{x}_\mathrm{i}^{\mathrm{T}}\beta\right)^2 + \lambda \sum_{j=1}^{M} \beta_j^2 = (\mathbf{X}\beta - \mathrm{y})^{\mathrm{T}}(\mathbf{X}\beta - \mathrm{y}) + \lambda\|\beta\|^2 \tag{4}$$

Derive the value of β^* that minimizes (4) in closed form and show that it is given by $\beta^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$. Please show you work in details to get full credit.

3 Classification

As part of testing a vaccine for a virus, a laboratory is trying to differentiate between sample cells which might be affected by staining them with a colored dye.

Logistic regression (2 marks)

The two kinds of cells that are being examined - (i) unaffected cells (Type A) (ii) virus affected cells (Type B). Let's assume that one of the features that be used to differentiate between the cells is the diameter of dye stain on the cells which can be represented as x_1.

The log of odds for the cells being unaffected is repesented as the below equation:

$$log(odds) \quad = \quad -10 + 2 * x_1 \tag{5}$$

If we use *Logistic Regression* as the classifier,

(a) What is the probability of a cell being unaffected if its diameter of dye stain is 6?

(b) What should be the minimum diameter of the dye stain on a cell to make sure that the cell is unaffected with a probability of 90%?

Softmax for Multi-Class Classification (3 marks)

A new cell type is added to the experiment which represents the cells affected with virus that were cured by the vaccine. Let's call them Type C. Now, along with the diameter of the stain, we will consider the depth of the stain color as a feature, which can be represented as x_2.

The *softmax function* is a multi-class generalization of the logistic sigmoid:

$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \tag{6}$$

Consider a case where the activation functions a_j are linear functions of the input features. The weights and bias details are provided in the table below.

Table 1: Weights and bias for input features

	Type A	**Type B**	**Type C**
w_1	2	5	5
w_2	5	10	2
$bias$	5	1.5	1

where
x_1 = diameter of the stain
x_2 = depth of the stain color

Answer the following questions.

(a) Write the equations for activation functions and class probabilities for Type A, Type B and Type C cells.

(b) Find the class probabilities for Type A, Type B and Type C cells for a sample cell with diameter=10 and depth=2. What will be the predicted type for the mentioned sample cell?

Note: *Use https://www.wolframalpha.com/ for calculations, if needed.*