Machine Learning Course 2024 Spring: Homework 1

March 1, 2024

1 Problem 1

Given a data set $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. A L_2-regularized least squares linear regression model (ridge regression) is employed to best fit this data set. It can be formulated as the following optimization problem:

$$\min_{\boldsymbol{w},b} \ell(\boldsymbol{w}, b) = \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b - y_i)^2 + \frac{\lambda}{2}||\boldsymbol{w}||_2^2, \tag{1.1}$$

where $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are the weight and bias terms respectively, and λ is the regularization parameter. Try to answer the following questions:

1. Rewrite the optimization problem into matrix form. Please clearly demonstrate the definition and shape of the matrix represented by each letter you use.

2. Is the optimal parameter $(\boldsymbol{w}^*, b^*)$ unique for any $\lambda > 0$? Please prove your conclusion.

3. The data set D with 6 instances is shown in Table 1, where each sample has 3 dimensions. Please calculate the optimal parameter $(\boldsymbol{w}^*, b^*)$ for $\lambda = 0.1$.

Table 1: Training set for ridge regression.

ID	x_1	x_2	x_3	y	**ID**	x_1	x_2	x_3	y
1	2	1	3	0	4	3	5	2	-3
2	5	3	6	0	5	1	7	2	-3
3	4	2	5	0	6	6	1	4	3

4. Consider a random noise $\varepsilon \sim N(0, \sigma^2)$ is added to the simple linear regression model, that is,

$$y_i = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}_i + \varepsilon_i. \tag{1.2}$$

Assume a Gaussian prior over each element of $\boldsymbol{\theta}$ with mean 0 and standard deviation τ, i.e. $\theta_j \sim N(0, \tau^2)$. Show that the estimate of $\boldsymbol{\theta}^*$ by maximizing the conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, where $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$, is equivalent to solving the optimization problem Eq.(1.1) with $b = 0$.

2 Problem 2

In a binary classification problem, each instance $\boldsymbol{x}_i \in \mathbb{R}^d$ in a data set $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ has a label $y_i \in \{0, 1\}$. A powerful tool to handle this kind of problem is the logistic regression model with the definition of the sigmoid function Eq.(2.1).

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad \text{, such that} \quad z = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b \tag{2.1}$$

To simplify this problem, we assume that $\boldsymbol{\beta} = (\boldsymbol{w}; b), \hat{\boldsymbol{x}} = (\boldsymbol{x}; 1)$. Since its negative log-likelihood function Eq.(2.2) is convex, we can optimize it efficiently with Gradient Descent method, Newton's Method, and so on.

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} (-y_i \boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i + \ln(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i})) \tag{2.2}$$

1. Prove the sigmoid function Eq.(2.1) is non-convex, and Eq.(2.2) is convex for parameter $\boldsymbol{\beta}$.

2. Suppose we are facing a K-class classification problem instead of a binary classification problem, where $y_i \in \{1, 2, \ldots, K\}$. Please expand the logistic regression model Eq.(2.1) to a multi-class version and write down the log-likelihood function of this multi-class logistic regression model.

3 Problem 3

In a binary classification problem, given the true label of the instance and the predicted values of the two classifiers C_1, C_2, calculate the relevant performance measures.

1. Calculate AUC (for C_1 and C_2 respectively).

2. Confusion Matrix (threshold=0.3 and 0.5 for C_1 and C_2 respectively).

3. $F1$-Score (threshold=0.3 and 0.5 for C_1 and C_2 respectively).

Table 2: True label and predicted values of two classifiers.

ID	y	y_{C_1}	y_{C_2}	**ID**	y	y_{C_1}	y_{C_2}
1	0	0.38	0.19	6	1	0.43	0.49
2	0	0.28	0.89	7	0	0.88	0.23
3	1	0.67	0.47	8	1	0.54	0.66
4	1	0.38	0.89	9	1	0.29	0.15
5	0	0.11	0.95	10	0	0.75	0.66