Homework II

Due date: Feb. 22nd 11:59pm, 2023

Submission Guidelines: 1. Put all the documents into one folder, name that folder as firstnamelastname_UIN and compress it into a .zip file.

2. For coding problems, your submission should include a code file (either .py or .ipynb), and also a pdf file (in one file together with other non-coding questions) to report the results that are required in the question.

Problem 1: Least Absolute Deviation (15 points)

In class, we have assumed the following data generative model

$$y = f(\mathbf{x}) + \epsilon$$

where ϵ follows a standard gaussian distribution, i.e., $\epsilon \sim \mathcal{N}(\epsilon|0,1)$. Assume a linear model for $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. We now modify the data generative model by assuming that ϵ follows a Laplacian distribution whose probability density function is

$$p(\epsilon) = \frac{\lambda}{2}\exp(-\lambda|\epsilon|)$$

where λ is a positive constant. For more about Laplacian distribution please check the following wiki page `http://en.wikipedia.org/wiki/Laplace_distribution`.

Based on the above noise model about ϵ, derive the log-liklihood for the observed training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and the objective function for computing the solution $\mathbf{w}$. Does the problem have a closed form solution like Least Square Regression?

Problem 2: Regression with Ambiguous Data (30 points)

In the regression model we talked about in class, we assume that for each training data point $\mathbf{x}_i$, its output value y_i is observed. However in some situations that we can not measure the exact value of y_i. Instead we only have information about if y_i is larger or less than some value z_i. More specifically, the training data is given as a triplet $(\mathbf{x}_i, z_i, b_i)$, where

- $\mathbf{x}_i$ is represented by a vector $\phi(\mathbf{x}_i) = (\phi_0(\mathbf{x}_i), \ldots, \phi_{M-1}(\mathbf{x}_i))^\top$
- $z_i \in \mathbb{R}$ is a scalar, $b_i \in \{0, 1\}$ is a binary variable indicating that if the true output y_i is larger than z_i ($b_i = 1$) or not $b_i = 0$

Develop a regression model for the ambiguous training data $(\mathbf{x}_i, z_i, b_i), i = 1, \ldots, n$.

Hint: Define a Gaussian noise model for y and derive a log-likelihood for the observed data. You can derive the objective function using the error function given below (note that there is no closed-form solution). The error function is defined as

$$\mathrm{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt$$

It is known that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right], \text{ and } \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-t^2/2} dt = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$$

Problem 3: Regularization Penalizes Large Magnitudes of Parameters (15 points)

In class, we have learned that when increasing the regularization parameter λ in the regularized least square problem

$$\min_{\mathbf{w}} \frac{1}{2}\|\Phi\mathbf{w} - \mathbf{y}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

where $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n, \Phi^\top = (\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)) \in \mathbb{R}^{M \times n}$, the magnitude of the optimal solution will decrease. Let the optimal solution $\mathbf{w}_*$ be

$$\mathbf{w}_* = (\lambda I + \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$$

You are asked to show that the Euclidean norm of the optimal solution $\|\mathbf{w}_*\|_2$ will decrease as λ increases.

Hint: (1) use the result from the Problem 2 in homework 1. (2) for any vector $\mathbf{u} \in \mathbb{R}^d$ if $V^\top V = I$ where $V \in \mathbb{R}^{d \times d}$ then $\|V\mathbf{u}\|_2 = \|\mathbf{u}\|_2$

Problem 4: Ridge Regression and Lasso (40 points)

In this problem, you are asked to learn regression models using Ridge regression and Lasso. The data set that we are going to use is the E2006-tfidf[1].

The first column is the target output y, and the remaining columns are features in the form of (feature_index:feature_value). You can load the data by sklearn[2]. If we let $\mathbf{x} \in \mathbb{R}^d$ denote the feature vector, the prediction is given by $\mathbf{w}^\top \mathbf{x} + w_0$, where $\mathbf{w} \in \mathbb{R}^d$ contains the coefficients for all features and w_0 is a intercept term. Denoting $\mathbf{X}^\top = (\mathbf{x}_1, ..., \mathbf{x}_n)$, the problem becomes

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{X}\mathbf{w} + w_0 - \mathbf{y}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2,$$

which is the Lasso regression problem when the regularization term $\|\mathbf{w}\|^2 = \|\mathbf{w}\|_1^2$ and is the Ridge regression problem when the regularization term $\|\mathbf{w}\|^2 = \|\mathbf{w}\|_2^2$.

You can use the Python sklearn library for Lasso[3] and Ridge regression[4].

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html

[2] https://scikit-learn.org/stable/datasets/loading_other_datasets.html

[3] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

(1) Solution of Ridge Regression and Lasso: Set the value of the regularization parameter $\lambda = 0.1$, compute the optimal solution for Ridge regression and Lasso. Report the number of nonzero coefficient in the solution $\mathbf{w}$ for both Ridge regression and Lasso. You may observe that the solutions of Ridge regression and Lasso contain very different numbers of nonzero elements. What is the cause of that? What can this imply? Justify your observation and hypothesis. (Note: If you use the sklearn Lasso, the value of alpha should be set to λ/n, where n is the number of training examples, and in the sklearn Ridge, set alpha to be λ. Same for following questions.)

(2) Training and testing error with different values of λ: (i) For each value of λ in [0, 1e-5, 1e-3, 1e-2 , 0.1, 1, 10, 100, 1e3, 1e4, 1e5, 1e6] run the Ridge regression and Lasso on training data to obtain a model $\mathbf{w}$ and then compute the root mean square error (RMSE[5]) on both the training and the testing data of the obtained model. (ii) Plot the error curves for root mean square error on both the training data and the testing data vs different values of λ. You need to show the curves, and discuss your observations of the error curves, and report the best value of λ and the corresponding testing error. (iii) Plot the curve of number of nonzero elements in the solution $\mathbf{w}$ vs different values of λ. Discuss your observations. (iv) Plot the curve of $\|\mathbf{w}\|_2^2$ vs different values of λ. Discuss your observations.

(3) Cross-validation: Use the given training data and follow the 5-fold cross-validation procedure to select the best value of λ for both Ridge regression and Lasso. Then train the model on the whole training data using the selected λ and compute the root mean square error on the testing data. Report the best λ and the testing error for both Ridge regression and Lasso.

[5] For a set of examples $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, the root mean square error of a prediction function $f(\cdot)$ is computed by RMSE $= \sqrt{\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2/n}$.