Machine Learning, 2024 Fall
Homework 3

Notice

Due 23:59 (CST), Dec 12, 2024
Plagiarizer will get 0 points.
LaTeXis highly recommended. Otherwise you should write as legibly as possible.

1 Regularization in Overfitting [20pts]

This question explores how incorporating L2 regularization can reduce the risk of overfitting in ordinary least squares regression. The objective function of L2-regularized regression (ridge regression):

$$\min_{\tilde{w}} \sum_{i=1}^{n} (\tilde{w}^\top \tilde{x}_i - y_i)^2 + \lambda \|\tilde{w}\|_2^2$$

can be reformulated as:

$$\min_{\tilde{w}} \sum_{i=1}^{n} (\tilde{w}^\top \tilde{x}_i - y_i)^2 \quad \text{subject to} \quad \|\tilde{w}\|_2^2 \leq B^2$$

We are going to focus on the second expression for simplicity. In addition, we are going to assume the following:

1. Each data point $(\tilde{x}_i, y_i)$ is drawn identically and independently from the distribution P, namely, the dataset $D \sim P^n$.
2. For any $(\tilde{x}, y)$ sampled from P, we have $\|\tilde{x}\|_2^2 = 1$.

With the above assumptions, solve the following questions:

(1) [10pts] Notice that $\tilde{w}(D)$ is a function of D, and since D is random, so is $\tilde{w}(D)$. Define $\bar{w} = \mathbb{E}_D[\tilde{w}(D)]$. Show that

$$\|\tilde{w}(D) - \bar{w}\|_2^2 \leq 4B^2$$

Hint: using the triangular inequality:

$$\|\mathbf{a} - \mathbf{b}\|_2 \leq \|\mathbf{a}\|_2 + \|\mathbf{b}\|_2$$

(2) [10pts] Define the model $h_D(\tilde{x}) = \tilde{w}(D)^\top \tilde{x}$ and $\bar{h}(\tilde{x}) = \mathbb{E}_D[h_D(\tilde{x})]$. Show that the variance of the model

$$\mathbb{E}_{\tilde{x},D}\left[\left(h_D(\tilde{x}) - \bar{h}(\tilde{x})\right)^2\right] \leq 4B^2$$

Hint: first showing that

$$h_D(\tilde{x}) - \bar{h}(\tilde{x}) = (\tilde{w}(D) - \bar{w})^\top \tilde{x}$$

and then using the Cauchy-Schwarz inequality:

$$(\mathbf{a}^\top \mathbf{b})^2 \leq (\mathbf{a}^\top \mathbf{a})(\mathbf{b}^\top \mathbf{b})$$

to conclude the result.

Solution: (a) Using the triangular inequality, we have:

$$\|\tilde{w}(D) - \bar{w}\|_2 \leq \|\tilde{w}(D)\|_2 + \|\bar{w}\|_2$$

Taking the square of each side:

$$\|\tilde{w}(D) - \bar{w}\|_2^2 \leq \|\tilde{w}(D)\|_2^2 + \|\bar{w}\|_2^2 + 2\|\tilde{w}(D)\|_2\|\bar{w}\|_2$$

Since $\bar{w} = \mathbb{E}_D[\tilde{w}(D)]$, we have:

$$\|\bar{w}\|_2^2 \leq B^2$$

Thus:

$$\|\tilde{w}(D) - \bar{w}\|_2^2 \leq \|\tilde{w}(D)\|_2^2 + \|\bar{w}\|_2^2 + 2\|\tilde{w}(D)\|_2\|\bar{w}\|_2 \leq B^2 + B^2 + 2B^2 = 4B^2$$

(b) Since $h_D(\tilde{x}) = \tilde{w}(D)^T\tilde{x}$, then:

$$h_D(\tilde{x}) - \bar{h}(\tilde{x}) = \tilde{w}(D)^T\tilde{x} - \mathbb{E}_D[\tilde{w}(D)^T\tilde{x}]$$

Because $\bar{w} = \mathbb{E}_D[\tilde{w}(D)]$ and the expectation of $\tilde{w}(D)$ does not depend on $\tilde{x}$, we have:

$$h_D(\tilde{x}) - \bar{h}(\tilde{x}) = \tilde{w}(D)^T\tilde{x} - \bar{w}^T\tilde{x}$$

Using the Cauchy-Schwarz inequality:

$$(h_D(\tilde{x}) - \bar{h}(\tilde{x}))^2 \leq ((\tilde{w}(D) - \bar{w})^T(\tilde{w}(D) - \bar{w}))(\tilde{x}^T\tilde{x})$$

This can be written as:

$$(h_D(\tilde{x}) - \bar{h}(\tilde{x}))^2 \leq \|\tilde{w}(D) - \bar{w}\|_2^2 \cdot \|\tilde{x}\|_2^2$$

Because $\|\tilde{x}\|_2^2 = 1$, we have:

$$(h_D(\tilde{x}) - \bar{h}(\tilde{x}))^2 \leq \|\tilde{w}(D) - \bar{w}\|_2^2$$

Using our result from part (a), we get:

$$(h_D(\tilde{x}) - \bar{h}(\tilde{x}))^2 \leq 4B^2$$

Finally, taking the expectation, we obtain:

$$\mathbb{E}_{\tilde{x},D}\big((h_D(\tilde{x}) - \bar{h}(\tilde{x}))^2\big) \leq 4B^2$$

Takeaway: By adding regularization, we essentially bound the variance of the model, which reduces overfitting.

2 Decision Tree [20pts]

In this problem you are asked to induce a decision tree from the training data set and evalate your decision tree on the test set. The data is about the canteen selection data of ShanghaiTech University.

Id	Student Type	Budget	Study Pressure	Weather	Canteen Choice
1	Undergraduate	High	Low	Sunny	Canteen 1
2	Graduate	Low	High	Overcast	Canteen 3
3	Undergraduate	Low	Medium	Rainy	Canteen 2
4	Graduate	High	Medium	Sunny	Canteen 1
5	Undergraduate	High	High	Overcast	Canteen 3

Table 1: **Training Set**

Id	Student Type	Budget	Study Pressure	Weather	Canteen Choice
1	Undergraduate	Low	High	Sunny	Canteen 1
2	Graduate	High	Low	Overcast	Canteen 3

Table 2: **Test Set**

(1) Based on the information gain, determine which feature is the root of the decision tree and explain your reason. [10pts]

(2) According to your result above, whats your accuracy on test set if we only have the root node and leaf with majority label. [5pts]

(3) Complete your decision tree and whats your accuracy on test set now. [5pts]

Solution:
(1)
Student Type:

1. Undergraduate <Canteen1, Canteen2, Canteen3> = <1, 1, 1>
2. Graduate <Canteen1, Canteen2, Canteen3> = <1, 1, 1>

So the entropy of Student Type is high.

Budget:

1. High <Canteen1, Canteen2, Canteen3> = <2, 0, 1>
2. Low <Canteen1, Canteen2, Canteen3> = <0, 2, 1>

The entropy of Budget is $\log 3 - \frac{2}{3}$.

Study Pressure:

1. High <Canteen1, Canteen2, Canteen3> = <0, 0, 2>
2. Medisum <Canteen1, Canteen2, Canteen3> = <1, 2, 0>
3. Low <Canteen1, Canteen2, Canteen3> = <1, 0, 0>

The entropy of Study Pressure is $\frac{1}{2}\log 3 - \frac{1}{3}$.

Weather:

1. Sunny <Canteen1, Canteen2, Canteen3> = <2, 0, 0>

2. Overcast <Canteen1, Canteen2, Canteen3> = <0, 1, 2>
3. Rainy <Canteen1, Canteen2, Canteen3> = <0, 1, 0>

The entropy of Study Pressure is $\frac{1}{2}\log 3 - \frac{1}{3}$.

Therefore, both "Study Pressure" and "Weather" can be the root.

(2)
If we choose "Study Pressure", the accuracy on test is 0%.
If we choose "Weather". the accuracy on test is 100%.

(3)
All reasonable decision tree can be accepted.

3 Breast Cancer Classification Using K-Nearest Neighbors [30pts]

You will work with a dataset containing measurements of tumor cells extracted from medical images. The goal is to predict whether a tumor is benign (Class 0) or malignant (Class 1) based on these measurements. The dataset includes the following features:

- Id: Sample identifier
- Cl.thickness: Clump Thickness
- Cell.size: Uniformity of Cell Size
- Cell.shape: Uniformity of Cell Shape
- Marg.adhesion: Marginal Adhesion
- Epith.c.size: Single Epithelial Cell Size
- Bare.nuclei: Bare Nuclei
- Bl.cromatin: Bland Chromatin
- Normal.nucleoli: Normal Nucleoli
- Mitoses: Mitoses
- Class: Breast cancer class (0 for benign, 1 for malignant)

Given a cancer dataset containing measurements of tumor cells, your task is to implement the K-Nearest Neighbors (KNN) algorithm to perform classification predictions on the preprocessed data. Split the dataset into training and testing sets (e.g., 70% training, 30% testing), and use a fixed random seed to ensure reproducibility of results.

(1) **Implement the KNN algorithm and provide the result.** The results should contain the model's performance metrics, including Accuracy, Precision, Recall, and F1-score.

Your answer should include: Embedded code, comment on your code and visual screenshot. Hint: Implement the knn algorithm by hand (Dont use the sklearn implementation) [10pts].

(2) **Analyze the Impact of Training Set Size on KNN Performance.** Using the same random seed, split the dataset into different training set proportions (e.g., 50%, 70%, 90%). Run a KNN classifier with $k = 1$ for each proportion and record the test accuracy. Discuss how training set size affects accuracy (overfitting and underfitting) [10pts].

(3) Make predictions on the test set using different values of k (1, 2, ..., 50). Analyze how changes in k affect accuracy, and discuss the underlying reasons for these variations. Plot the test accuracy for different k values to visualize how k affects model performance [10pts].

Solution:
(1) Refer to lecture+13-knn, the algorithm consists of the following steps:
1. Calculate Distance 2. Select K Nearest Neighbors 3. Vote for Majority Class 4. Predict Output

(2) Provide results on different training set sizes. All reasonable discussions can be accepted.

(3) Plot the test accuracy on different k values. All reasonable anlaysis can be accepted.

4 Linear Regression for Wine Quality Dataset [30 pts]

A dataset of wine quality is provided, consisting of red wine (1599 samples) and white wine (4898 samples). The dataset includes 11 features related to wine (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) and the final quality score of the wine. We treat the wine quality score as a regression problem. We set the first 4000 rows of white wine data are used as the training set, and the rest as the test set.

Hint: Implement the linear regression algorithm by hand.

(1) Use the original feature data to solve this problem with a linear regression algorithm. Provide the linear regression expression, the linear regression weights, embedded core code, and performance (MSE) on the test set. [10pts]

(2) Analyze which features have the greatest impact on white wine quality scores. Are these features positively or negatively correlated with quality? [5pts]

(3) Implement L2 regularization, set the regularization coefficient to 1, and provide the linear regression with L2 regularization expression, the linear regression weights, embedded code, and performance (MSE) on the test set. [10pts]

(4) Based on question 3, implement K-fold cross-validation (K=5) to determine the optimal L2 regularization coefficient from the values $\{0.01, 0.1, 1, 10, 100\}$ based on performance on the validation set and provide the performance (MSE) on the test set. [5pts]

Solution:
(1)
Linear regression expression:

$$Xw = Y$$
$$w = (X^T X)^{-1} X^T Y$$

(2)
Pay attention to normalizing the data to obtain meaningful weight values.

(3)
Linear regression expression with L2 regularization:

$$Xw + \lambda w^T w = Y$$
$$w = (X^T X + \lambda I)^{-1} X^T Y$$

(4)
Refer to lecture+11-validation.