Homework4

Name: Chenqing Ji
Student ID: 11911303
Statistical Learning for Data Science

April 29, 2023

1 Conceptual Questions

1.1

Please explain the meaning of the bias-variance tradeoff and how to strike a balance between model complexity and prediction error.

Solution:

The bias-variance tradeoff means the tradeoff between a model's ability to fit the training data and its ability to generalize to new, unseen data. A model with high bias is too simplistic and unable to capture the complexity of the underlying relationships in the data, resulting in underfitting. However, a model with high variance will become too complex and overfits the training data, resulting in poor generalization performance.

In order to strike a balance between the model complexity and prediction error, it is essential to evaluate the model's performance on the testing set to assess its generalization performance. By comparing the performance on the training set and the testing set, we can easily assess whether the model is overfitting or underfitting. One main approach to strike a balance between bias and variance is to use regularization techniques such as L1 and L2 regularization to constrain the model's complexity. Additionally, cross-validation can be used to assess the model's performance on different subsets of the data and to tune the model's hyperparameters to improve its generalization performance.

1.2

Please briefly describe the differences between L1 regularization and L2 regularization and their roles in model optimization.

Solution:

Firstly, L1 and L2 regularization are techniques used to prevent overfitting in machine learning models by adding a penalty term to the loss function.

L1 regularization always adds the sum of the absolute values of the model's coefficients multiplied by a hyperparameter lambda to the loss function. This penalty term encourages the model to reduce the number of features it uses and to set some coefficients to zero. Thus, L1 regularization can be useful for feature selection and creating sparse models.

On the other hand, L2 regularization always adds the sum of the squared values of the model's coefficients multiplied by a hyperparameter lambda to the loss function. This penalty term encourages the model to shrink the coefficients towards zero without setting them exactly to zero. Thus, L2 regularization can be useful for reducing the impact of collinearity between features and preventing overfitting in high-dimensional datasets.

Generally, L1 regularization tends to create sparser models with fewer features, while L2 regularization tends to produce smoother models with more stable and balanced coefficients. Therefore, the choice between L1 and L2 regularization, or a combination of both, depends on the specific problem and the characteristics of the dataset.

1.3

Please explain the basic principle of the K-Nearest Neighbors (KNN) algorithm and how to choose an appropriate K value.

Solution:

The basic principle of the KNN algorithm is to classify new data points based on the majority vote of their k nearest neighbors in the training dataset. In other words, the KNN algorithm assumes that similar data points tend to be in the same class or have similar target values. To apply the KNN algorithm, we first need to select a value for k, which determines the number of nearest neighbors to consider. Choosing an appropriate k value is critical for the performance of the algorithm. A smaller value of k leads to a more complex decision boundary and may result in overfitting, while a larger value of k may lead to underfitting and oversimplification of the model. Therefore, it is important to choose a k value that balances bias and variance.

One way to choose an appropriate k value is to use cross-validation techniques, such as k-fold cross-validation, to evaluate the performance of the model for different k values. Another approach is to use a heuristic rule of thumb, such as the square root of the number of samples in the training dataset, which tends to work well in practice. However, the optimal k value may vary depending on the specific dataset and task and we usually use different k values in experiment to find the best one.

1.4

Please explain the basic concepts of Support Vector Machines (SVM), including support vectors, margin, and kernel function.

Solution:

The basic idea of SVM is to find the hyperplane that best separates the different classes in the dataset. The key concepts of SVM is shown below:

(1) Support Vectors: In SVM, support vectors are the data points that are closest to the decision boundary, also known as the hyperplane. These points determine the position and orientation of the hyperplane and play a crucial role in the training process. The SVM algorithm seeks to maximize the margin between the support vectors and the decision boundary, which helps in improving the model's accuracy.

(2) Margin: Margin refers to the distance between the decision boundary and the closest support vectors. A larger margin implies that the model is more confident about the classification of the data points. In SVM, the goal is to maximize the margin, as this helps in reducing the generalization error of the model.

(3) Kernel Function: Kernel functions are used in SVM to map the input data into a higher-dimensional feature space where the data can be better separated. The kernel function calculates the dot product between the input data points in the higher-dimensional space, without actually having to compute the coordinates of the data points in that space. The choice of kernel function plays an important role in the performance of the SVM model. There are various types of kernel functions available, including linear, polynomial and radial basis function (RBF). The RBF kernel is one of the most commonly used kernels in SVM, as it is effective in handling non-linearly separable data.

2 Calculation Questions

2.1

Given the following dataset: $\boldsymbol{X}$ **:** $[\mathbf{2, 4, 6, 8, 10, 12}]$ $\boldsymbol{Y}$ **:** $[\mathbf{3, 5, 7, 9, 11, 13}]$. Please calculate the regression coefficient w and the intercept b using simple linear regression (either manually or by writing code).

Solution:

To calculate the regression coefficient w and the intercept b using simple linear regression, we need to find the equation of a straight line that best fits the given data. This equation is given by:

$$\hat{\boldsymbol{y}} = \boldsymbol{wx} + \boldsymbol{b} \tag{1}$$

where $\boldsymbol{y}$ is the dependent variable, $\boldsymbol{x}$ is the independent variable, $\boldsymbol{w}$ is the regression coefficient and $\boldsymbol{b}$ is the intercept. Then, for the least squares coefficient estimate in simple linear regression, we can know that:

$$\boldsymbol{w} = \frac{\sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{y_i} - \overline{\boldsymbol{y}})}{\sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^{\mathbf{2}}} \tag{2}$$

$$\boldsymbol{b} = \overline{\boldsymbol{y}} - \boldsymbol{w}\overline{\boldsymbol{x}} \tag{3}$$

According to the $\boldsymbol{XY}$ dataset given in this problem, we can easily find that:

$$\overline{\boldsymbol{x}} = \mathbf{7}, \overline{\boldsymbol{y}} = \mathbf{8} \tag{4}$$

Using the above equation(2) and (3), we can find that:

$$\boldsymbol{w} = \mathbf{1}, \boldsymbol{b} = \mathbf{1} \tag{5}$$

Therefore, the regressing Y on X can be displayed as:

$$\hat{\boldsymbol{y}} = \boldsymbol{x} + \mathbf{1} \tag{6}$$

2.2

Assume a binary classification problem with the following confusion matrix:

Actual\Predicted	Positive	Negative
Positive	80	20
Negative	10	90

Figure 1: Confusion matrix for Problem 2.2

Please calculate the following evaluation metrics: Accuracy, Precision, Recall, and F1-score.

Solution:

According to the confusion matrix, we can find the : $\boldsymbol{TP = 80, TN = 90, FP = 10, FN = 20}$.

Therefore, based on the equation in the ppt and textbook, the value of the evaluation metrics: Accuracy, Precision, Recall and F1-score are:

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{80 + 90}{200} = 85\%} \tag{7}$$

$$\boldsymbol{Precision = \frac{TP}{TP + FP} = \frac{80}{80 + 10} \approx 88.9\%} \tag{8}$$

$$\boldsymbol{Recall = \frac{TP}{TP + FN} = \frac{80}{80 + 20} = 80\%} \tag{9}$$

$$\boldsymbol{F1 = 2 * \frac{Recall * Precision}{Recall + Precision} = \frac{2 * 8/10 * 8/9}{8/9 + 8/10} = 84.2\%} \tag{10}$$

In a conclusion, the Accuracy is $\boldsymbol{85\%}$, the Precision is $\boldsymbol{88.9\%}$, the Recall is $\boldsymbol{80\%}$ and the F1-score is $\boldsymbol{84.2\%}$.

2.3

Suppose we have a regression problem with the following true and predicted values: True values: $\boldsymbol{[10, 15, 20, 25, 30]}$ Predicted values: $\boldsymbol{[12, 18, 19, 24, 32]}$ Please calculate the Mean Squared Error (MSE) and $\boldsymbol{R^2}$ score.

Solution:

Firstly, the mean of the true value is:

$$\boldsymbol{\overline{y} = \frac{10 + 15 + 20 + 25 + 30}{5} = 20} \tag{11}$$

Then, based on the equation in the textbook and ppt, the value of the Mean Squared Error (MSE) and $\boldsymbol{R^2}$ score are:

$$\boldsymbol{MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} = \frac{4+9+1+1+4}{5} = \frac{19}{5} = 3.8} \tag{12}$$

$$\boldsymbol{R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{19}{250} = \frac{231}{250} = 0.924} \tag{13}$$

Therefore, the Mean Squared Error (MSE) is **3.8** and the $\boldsymbol{R^2}$ score is **0.924**.

3 Inference Questions

3.1

In a binary classification problem, you used a logistic regression model for prediction. The model performed well during training but had poor predictive performance on new data. Please analyze the possible reasons and provide corresponding solutions.

Solution:

Here are some possible reasons and corresponding solutions about this problem:

(1) Overfitting: The model may have overfit the training data, meaning it has learned the noise in the data and is not able to generalize well to new data. One solution to this is to use regularization techniques such as L1 or L2 regularization, dropout, or early stopping to prevent overfitting.

(2) Incorrect Features: The features used for training the model may not be relevant or informative for predicting the target variable on new data. One solution to this is to carefully analyze and select the features to use, or to use feature selection techniques such as PCA (Principal Component Analysis) or Random Forest feature importance to identify the most relevant features.

(3) Imbalanced Data: The class distribution in the training data may be different from the class distribution in the new data, leading to poor predictive performance on new data. One solution to this is to balance the classes by either oversampling the minority class or undersampling the majority class.

3.2

When dealing with a regression problem with multiple features, we may encounter the issue of multicollinearity. Please explain what multicollinearity is and how to detect and address it.

Solution:

Multicollinearity is a phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other. It can lead to unstable and unreliable regression coefficients, making it difficult to interpret the impact of each individual variable on the dependent variable.

To detect the multicollinearity, we can examine the correlation matrix between all of the independent variables. High correlation coefficients (close to 1 or -1) indicate a

potential multicollinearity issue. Moreover, we can also compute the Variance Inflation Factor (VIF) for each independent variable. VIF measures the degree to which the variance of an estimated regression coefficient is increased because of collinearity in the model. Generally, a VIF value greater than 10 indicates that multicollinearity may be present.

To address the multicollinearity, we can firstly remove one of the highly correlated variables, which means: If two or more variables are highly correlated, we can consider removing one of them from the model. Then, we can combine the collinear variables together into a single predictor to create a new predictor. Moreover, we can also use some regularization methods like ridge regression to select important variables and penalize the unimportant ones, thus reducing the impact of multicollinearity.