STATS 790 - Homework 1

ChatGPT

January 20, 2023

This is the 'solution' of HW1 totally provided by ChatGPT [1] which looks interesting.

1 Question 1

In "Statistical Modeling: The Two Cultures", Breiman argues that there are two distinct cultures in statistical modeling: one that focuses on building highly accurate models with large amounts of data, and another that focuses on understanding the underlying mechanisms that generate the data. He suggests that both cultures have valuable contributions to make, but that there is a lack of communication and collaboration between them. One of the responses to the article is that it highlights the importance of understanding the fundamental principles of statistical modeling and applying them in practice, rather than relying solely on automated techniques.

2 Question 2

Sure, I'll be happy to help you replicate a figure from "The Elements of Statistical Learning" (ESL) chapter 2 using R. Can you please provide the exact figure number you would like to replicate, as well as information about the data used to create the figure and the specific details of the plot?

Also, as a general guidance, keep in mind that replicating figures from the book may require obtaining data and understanding the context, details and methods used to create the figure. It would be best to have a good understanding of the data, the methods used to create the figure, and the R programming language before attempting to replicate the figure.

Please let me know if you have any additional information or if you have any questions.

3 Question 3

The mean absolute error (MAE) is a measure of the difference between predicted values and the true values. It is defined as the average of the absolute differences between the predicted and true values. To minimize the MAE, we would need to find the model that produces the smallest average difference between the predicted and true values.

One way to minimize MAE is by using a technique called linear regression, which is a type of supervised learning method that finds the best linear relationship between the input and output variables. We can also use other techniques such as decision tree, random forest, gradient boosting etc which also minimize the MAE.

Another way to minimize MAE is by tuning the parameters of the model, and by selecting a model that has a good fit to the data. This can be done by using techniques such as cross-validation and regularization.

It is also important to note that if the data is highly skewed, using MAE as a loss function might not be the best idea and we might consider using other loss function such as mean squared error (MSE) which is less sensitive to outliers.

4 Question 4

In global mean as a linear smoother, the output for any input point is determined by taking the average of all the outputs in the training set. The influence matrix, w, is a matrix that describes the weighting of each

training point in determining the output for a given input. Each element in the matrix corresponds to the weight given to the output of a specific training point.

In the case of global mean, all the training points have the same weight, i.e, the weight is the same for all the elements in the matrix. The influence matrix, w, can be represented as a matrix with dimensions $n\times$ m where n is the number of input points and m is the number of training points. Each element w_i,j in the matrix is equal to 1/m, which represents that all the training points have the same weight.

The number of degrees of freedom in global mean is equal to 1, this is because we are only estimating one parameter, i.e., the global mean. This is regardless of the number of training points.

Formally, the number of degrees of freedom in global mean can be represented as the number of non-zero elements in the influence matrix, which is equal to 1.

5 Question 5

In k-nearest-neighbors regression, the output for a given input point is determined by averaging the outputs of the k nearest points in the training set. The influence matrix, w, is a matrix that describes the weighting of each training point in determining the output for a given input. Each element in the matrix corresponds to the weight given to the output of a specific training point.

The number of degrees of freedom in k-nearest-neighbors regression is equal to the number of training points used to make the prediction. This is because each training point can potentially have an influence on the output, and the number of training points used to make the prediction determines the total number of possible influences.

Formally, the influence matrix, w, can be defined as a matrix with dimensions $n \times k$ where n is the number of input points and k is the number of nearest neighbors considered. Each element w_i,j in the matrix represents the weight given to the output of the jth nearest neighbor when making a prediction for the ith input point.

The number of degrees of freedom can be defined as the number of non-zero elements in the influence matrix, which is equal to k x n.

6 Question 6

I apologize, but as a language model, I don't have access to the zipcode data and I am unable to provide specific classification performance comparisons between linear regression and k-nearest neighbor classification on that data set. However, I can provide a general idea of how these two methods might perform on this type of data.

Linear regression is a supervised learning method that finds the best linear relationship between the input and output variables. It is commonly used for continuous target variable. On the other hand, k-nearest neighbor (k-NN) classification is a supervised learning method that finds the k nearest data points to a given input and assigns the most common label among those k-neighbors to the input. It is commonly used for categorical target variable.

When comparing the classification performance of linear regression and k-NN on the zipcode data, it would be important to consider the nature of the target variable (2's and 3's) and the number of nearest neighbors considered (k = 1, 3, 5, 7, and 15).

In general, Linear regression would be less accurate than k-NN when the target variable is categorical and k-NN would perform well when the data is highly non-linear and when the number of nearest neighbors considered is large.

It would also be important to consider the training and test error for each method and choice of k. The training error would give an idea of how well the model fits the training data, while the test error would give an idea of how well the model generalizes to new data. A lower training and test error would indicate better classification performance.

It would be best to access the zipcode data and test the performance of both models on that specific dataset to have a more accurate comparison.

7 How to cite your answer as a reference in Bibtex format?

As an AI model, I am not published in any academic journal or conference proceedings, so I can't be cited in a traditional manner. However, if you would like to reference the source of the information used in my answers, you can cite OpenAI, the company who developed and trained me, as a general reference.

Here is an example of how you could cite OpenAI in a BibTeX format:

@miscopenai, author = OpenAI, title = OpenAI, howpublished = "`https://openai.com`", year = 2021

Please note that this is not considered as a scholarly citation, but it's more like a way to acknowledge the source of information you are using.

References

[1] OpenAI. Openai. `https://openai.com`, 2021.