STATS 790 - Homework 1

Yicen Li

January 19, 2023

1 Question 1

I agree with Breiman's opinion that the goal when dealing with practical problems is not interpretability, but accurate information. The most inspiring idea in this article is that he argued to develop a new definition such as prediction accuracy to replace the original goodness of fit methods. It is forethoughtful that Breiman pointed out the limitations of traditional statistics (data models) and advocated machine learning (algorithmic models) that are able to overcome difficulties to obtain more accurate and valid conclusions in 2001 [1].

(Sorry exceeds 2-3 sentences)[Due to the trade-off between interpretability and model simplicity, linear regressions are simple and interpretable but far less accurate than neural networks on complex datasets when making predictions. Especially nowadays people conduct research on text, images and various types of complicated data which can not be easily handled by prior distribution assumptions [1]. A recent comment claims that "Breiman was a provocateur in the best possible terms. Without people like him, statistics would not be where it is today." [2]. It is also cheerful that statistics has done more since then such as lasso and more model selection methods.

2 Question 2

I replicate the ESL 2.1 figure. Codes partially refer to the open source.

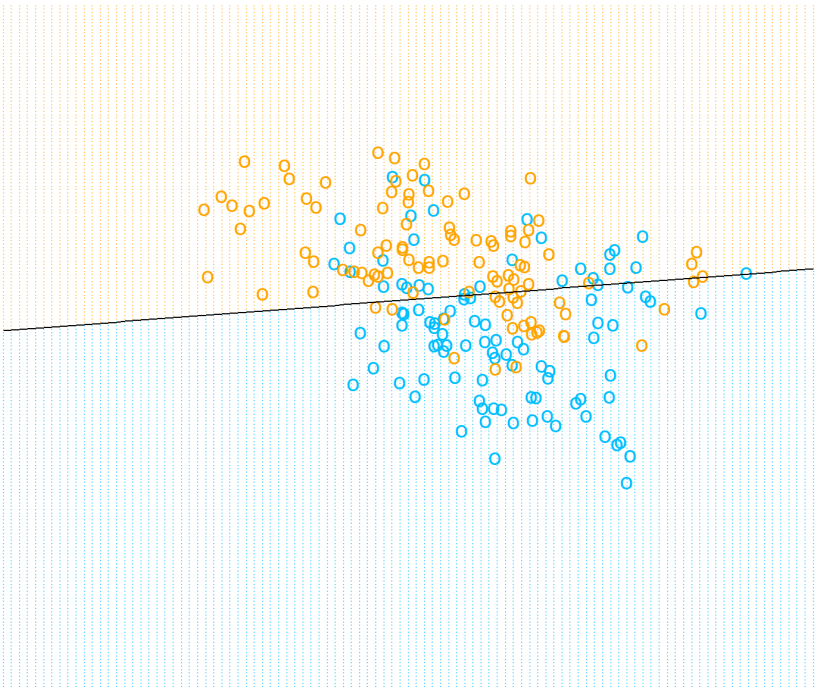

Figure 1: ESL2.1. Codes available on Github

3 Question 3

Suppose $\mathbf{Y}$ is a continuous random variable with pdf $f(y)$.

$$\begin{aligned} MAE(m) &= \mathbb{E}[|\mathbf{Y} - m|] \\ &= \int_{-\infty}^{\infty} |y - m| f(y) dy \\ &= \int_{-\infty}^{m} (m - y) f(y) dy + \int_{m}^{\infty} (y - m) f(y) dy \\ &= \int_{-\infty}^{m} m f(y) dy - \int_{-\infty}^{m} y f(y) dy + \int_{m}^{\infty} y f(y) dy - \int_{m}^{\infty} m f(y) dy \end{aligned}$$

To maximize MAE, we have:

$$\begin{aligned} \frac{\partial MAE}{\partial m} &= \frac{\partial \int_{-\infty}^{m} m f(y) dy}{\partial m} - \frac{\partial \int_{-\infty}^{m} y f(y) dy}{\partial m} + \frac{\partial \int_{m}^{\infty} y f(y) dy}{\partial m} - \frac{\partial \int_{m}^{\infty} m f(y) dy}{\partial m} \\ &= \int_{-\infty}^{m} f(y) dy + m f(m) - m f(m) - m f(m) - \int_{m}^{\infty} f(y) dy + m f(m) \\ &= \int_{-\infty}^{m} f(y) dy - \int_{m}^{\infty} f(y) dy \\ &= 0 \end{aligned}$$

Hence, we obtain:

$$\int_{-\infty}^{m} f(y) dy - \int_{m}^{\infty} f(y) dy = 0$$

And we have:

$$\int_{-\infty}^{m} f(y) dy + \int_{m}^{\infty} f(y) dy = 1$$

Thus, it is clear that:

$$\int_{-\infty}^{m} f(y) dy = \int_{m}^{\infty} f(y) dy = \frac{1}{2}$$

which is exactly the definition of the median:

$$\mathbf{P}(y \leq m) = \mathbf{P}(y \geq m) = \frac{1}{2}$$

We conclude that the median of $\mathbf{Y}$ minimizes the MAE. In general, MAE works better if the training data is polluted by outliers (eg, there are a lot of wrong negative and positive labels in the training data, but not in the test set) When dealing with outliers, the MAE loss function is more stable, but its derivative is discontinuous or even does not exist, thus there will be troubles for converges to update parameters. The MSE loss function is more sensitive to outliers since we squared the error. But on the other side, MSE has a better mathematical expression when differentiating. By setting its derivative to 0, a more stable closed solution can be obtained. Therefore, it more depends on our training set and task to decide which loss function should be applied.

4 Question 4

If we take the global mean as a linear smoother, we have:

$$\hat{\mu}(x) = \sum_{i=1}^{n} \frac{y_1 + ... + y_n}{n}$$

we can rewrite it in matrix form as:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ ... \\ \mu_n \end{pmatrix} = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & ... & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & ... & \frac{1}{n} \\ ... & ... & ... & ... \\ \frac{1}{n} & \frac{1}{n} & ... & \frac{1}{n} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}$$

or

$$\hat{\mu} = \mathbf{w}\mathbf{y}$$

where $\mathbf{w}$ is a $n \times n$ matrix and it is the hat matrix in this case. According to Eq. 1.70 from the ADA book [3],

$$\begin{aligned} df\hat{\mu} &= \mathbf{tr}\ \mathbf{w} \\ &= \mathbf{tr} \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ \dots & \dots & \dots & \dots \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix} \\ &= \sum_{1}^{n} \frac{1}{n} \\ &= 1 \end{aligned}$$

Hence we show that it has one degree of freedom.

5 Question 5

If we take the k-nearest-neighbours regression as a linear smoother, we have:

$$\hat{\mu}(x) = \sum_i y_i \hat{w}(x_i, x)$$

where

$$\hat{w}(x_i, x) = \begin{cases} \frac{1}{k} & x_i \ one\ of\ the\ k\ nearest\ neighbors\ of\ x \\ 0 & otherwise \end{cases}$$

Similarly, we can rewrite it as:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \hat{w}(x_1, x) & \hat{w}(x_2, x) & \dots & \hat{w}(x_n, x) \\ \hat{w}(x_1, x) & \hat{w}(x_2, x) & \dots & \hat{w}(x_n, x) \\ \dots & \dots & \dots & \dots \\ \hat{w}(x_1, x) & \hat{w}(x_2, x) & \dots & \hat{w}(x_n, x) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$$

or

$$\hat{\mu} = \mathbf{w}\mathbf{y}$$

where $\mathbf{w}$ is a $n \times n$ matrix and it is the hat matrix in this case. According to Eq. 1.70 from the ADA book [3],

$$\begin{aligned} df\hat{\mu} &= \mathbf{tr}\ \mathbf{w} \\ &= \mathbf{tr} \begin{pmatrix} \hat{w}(x_1, x) & \hat{w}(x_2, x) & \dots & \hat{w}(x_n, x) \\ \hat{w}(x_1, x) & \hat{w}(x_2, x) & \dots & \hat{w}(x_n, x) \\ \dots & \dots & \dots & \dots \\ \hat{w}(x_1, x) & \hat{w}(x_2, x) & \dots & \hat{w}(x_n, x) \end{pmatrix} \\ &= \sum_{i=1}^{n} \hat{w}(x_i, x) \end{aligned}$$

(The question states the expression should be in terms of k and n. However, I think that $\sum_{i=1}^{n} \hat{w}(x_i, x)$ should always equal 1 no matter which k is since it goes through all the x_i. Maybe I got this question wrong.)

6 Question 6

We set the threshold as 2.5 for linear regression to classify. The Zip dataset is provided by ESL [4].

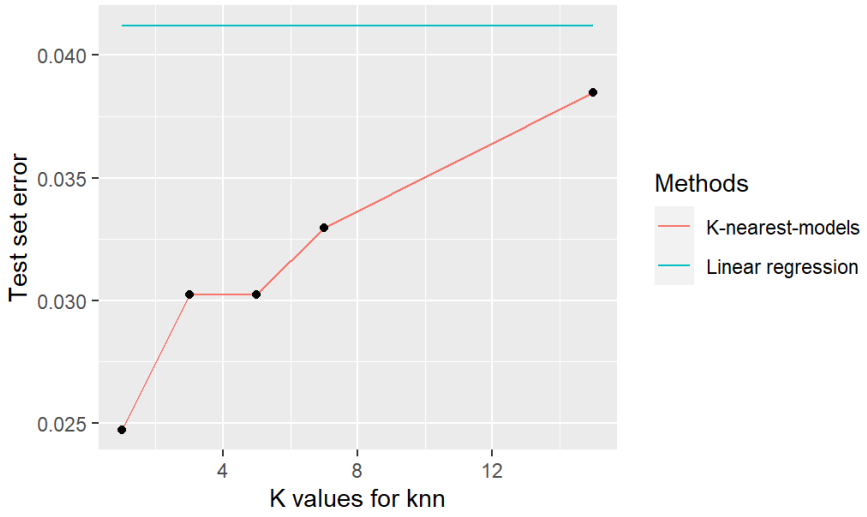

Figure 2: The test set error of linear regression and k– nearest neighbour classification on the zipcode data. Codes available on Github

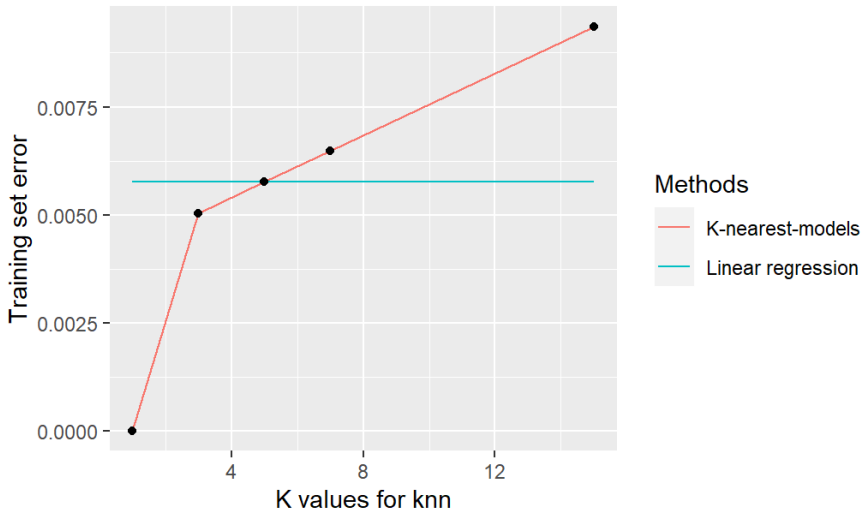

Figure 3: The training set error of linear regression and k– nearest neighbour classification on the zipcode data. Codes available on Github

We can observe that as the increase of k, the performance of knn reduced on both training and test set. In general, the k-nearest neighbour models outperform linear regression on the test set.

References

[1] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[2] Jelena Bradic and Yinchu Zhu. Comments on leo breiman's paper: "statistical modeling: The two cultures" (statistical science, 2001, 16(3), 199-231). *Observational studies*, 7(1):21–31, 2021.

[3] Cosma Shalizi. Advanced data analysis from an elementary point of view. 2013.

[4] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.