DSC 40A

Theoretical Foundations of Data Science I

In This Video

The optimal prediction h^* that minimizes $R(h)$ is a measure of center. What is the meaning of the value of $R(h^*)$?

Recommended Reading

Course Notes: Supplement 1

General Approach

- We start with a loss function $L(h, y)$.
- Then we minimize the empirical risk:

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} L(h, y_i).$$

- The input h^* that minimizes $R(h)$ is some measure of the **center** of the data set.
- The minimum output $R(h^*)$ represents some measure of the **spread**, or variation, in the data set.

Absolute Loss

- The empirical risk for the absolute loss is

$$R(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|.$$

- $R(h)$ is minimized at $h^* = \text{median}(y_1, y_2, \ldots, y_n)$.

Absolute Loss

- The empirical risk for the absolute loss is

$$R(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|.$$

- $R(h)$ is minimized at $h^* = \text{median}(y_1, y_2, \ldots, y_n)$.

- Therefore, the minimum value of $R(h)$ is

$$\begin{aligned} R(h^*) &= R(\text{median}(y_1, y_2, \ldots, y_n)) \\ &= \frac{1}{n}\sum_{i=1}^{n} |y_i - \text{median}(y_1, y_2, \ldots, y_n)|. \end{aligned}$$

Mean Absolute Deviation from the Median

- The minimum value of $R(h)$ is the **mean absolute deviation from the median**.

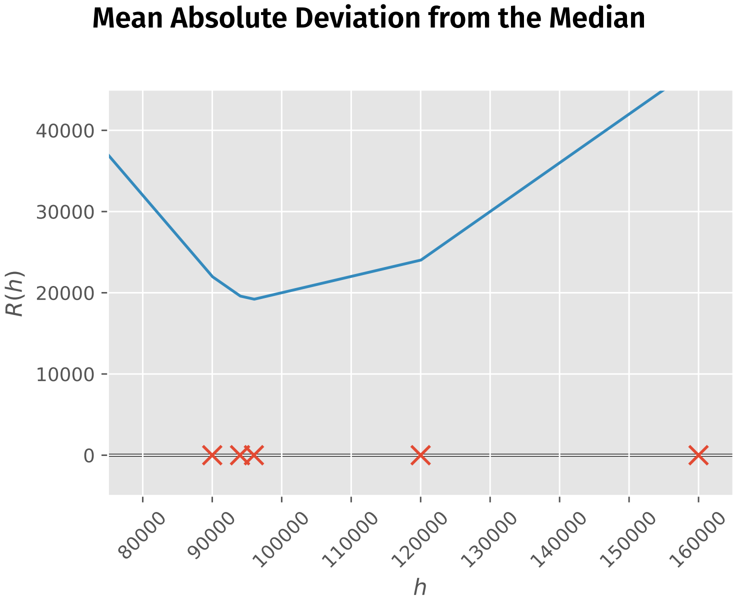
$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \text{median}(y_1, y_2, \ldots, y_n)|$$

- It measures how far each data point is from the median, on average.

Question

For the data set $2, 3, 3, 4$, what is the mean absolute deviation from the median?

a) 0 b) $\frac{1}{2}$ c) 1 d) 2



Mean Absolute Deviation from the Median
R(h)
h
40000
30000
20000
10000
0
80000
90000
100000
110000
120000
130000
140000
150000
160000

Square Loss

- The empirical risk for the square loss is

$$R_{\mathrm{sq}}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h)^2.$$

- $R_{\mathrm{sq}}(h)$ is minimized at $h^* = \mathrm{mean}(y_1, y_2, \dots, y_n)$.

Square Loss

- The empirical risk for the square loss is

$$R_{\text{sq}}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2.$$

- $R_{\text{sq}}(h)$ is minimized at $h^* = \text{mean}(y_1, y_2, \ldots, y_n)$.

- Therefore, the minimum value of $R_{\text{sq}}(h)$ is

$$\begin{aligned} R_{\text{sq}}(h^*) &= R_{\text{sq}}(\text{mean}(y_1, y_2, \ldots, y_n)) \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \text{mean}(y_1, y_2, \ldots, y_n))^2. \end{aligned}$$

Variance

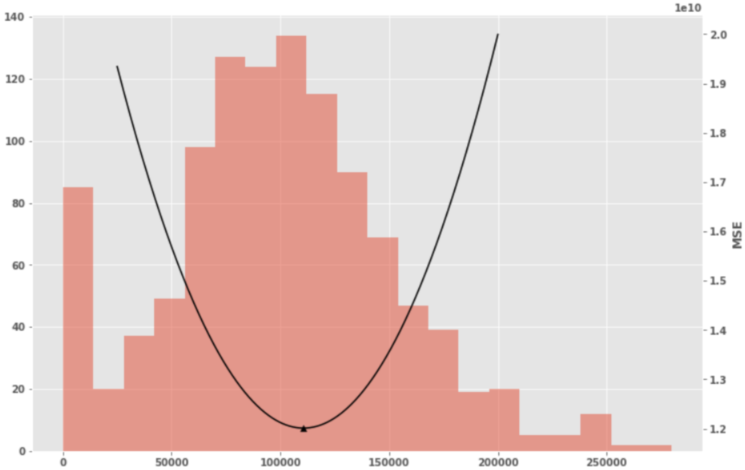
- The minimum value of $R_{\text{sq}}(h)$ is the mean squared deviation from the mean, more commonly known as the **variance**.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \text{mean}(y_1, y_2, \ldots, y_n))^2$$

- It measures the squared distance of each data point from the mean, on average.

- Its square root is called the **standard deviation.**

Variance



1e10
MSE
140
120
100
80
60
40
20
0
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
0
50000
100000
150000
200000
250000

0-1 Loss

- The empirical risk for the 0-1 loss is

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 0, & \text{if } h = y_i \\ 1, & \text{if } h \neq y_i \end{cases}$$

- This is simply a count of the number of data points not equal to h.
- $R_{0,1}(h)$ is minimized at $h^* = \text{mode}(y_1, y_2, \ldots, y_n)$.
- Therefore, $R_{0,1}(h^*)$ is a count of the number of data points not equal to the mode.

A Poor Way to Measure Spread

- The minimium value of $R_{0,1}(h)$ is the number of data points not equal to the mode.
- Higher value means less of the data is clustered at the mode.
- Just as the mode is a very simplistic way to measure the center of the data, this is a very crude way to measure spread.

A Poor Way to Measure Spread

- The minimum value of $R_{0,1}(h)$ is the number of data points not equal to the mode.
- Higher value means less of the data is clustered at the mode.
- Just as the mode is a very simplistic way to measure the center of the data, this is a very crude way to measure spread.

Question

For two different data sets, does it make sense say the data set with more data points not equal to the mode is more spread out?

Summary

- Different loss functions lead to empirical risk functions that are minimized at various measures of **center**.
- The minimum values of these risk runctions are various measures of **spread**.
- There are many different ways to measure both center and spread. These are sometimes called **descriptive statistics**.