DSC 40A

Theoretical Foundations of Data Science I

In This Video

We've looked at mean error and mean squared error. How do both of these ways of measuring the quality of a prediction fit into a general framework?

Recommended Reading

Course Notes: Chapter 1, Section 2

A General Framework

- We started with the **mean error**:

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

- Then we introduced the **mean squared error**:

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

- They have the same form: both are averages of some measurement that represents how different h is from the data.

A General Framework

- Definition: A **loss function** $L(h, y)$ takes in a prediction h and a right answer, y, and outputs a number measuring how far h is from y (bigger = further).

- The **absolute loss**:

$$L_{\text{abs}}(h, y) = |y - h|$$

or

- The **square loss**:

$$L_{\text{sq}}(h, y) = (y - h)^2 \leftarrow e^{|y-h|}$$

$$|y-h|^3$$

A General Framework

- Suppose that $y_1, \ldots, y_n$ are some data points, h is a prediction, and L is a loss function. The **empirical risk** is the average loss on the data set:

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

R is risk

free to change

- The goal of learning: find h that minimizes R_L. This is called **empirical risk minimization (ERM)**.

Designing a learning algorithm using ERM

1. Pick a loss function.

2. Pick a way to minimize the average loss on the data (empirical risk).

- **Key Idea**: The choice of loss function determines the properties of the result and the difficulty of computing it.

Example: 0-1 Loss

1. Pick as our loss function the **0-1 loss**:

$$L_{0,1}(h, y) = \begin{cases} 0, & \text{if } h = y \\ 1, & \text{if } h \neq y \end{cases}$$

same or not

← bigger when h is different from y

2. Minimize empirical risk:

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^{n} L_{0,1}(h, y_i)$$

Example: 0-1 Loss

1. Pick as our loss function the **0-1 loss**:

$$L_{0,1}(h, y) = \begin{cases} 0, & \text{if } h = y \\ 1, & \text{if } h \neq y \end{cases}$$

same ⇒ 0
diff ⇒ 1

2. Minimize empirical risk:

$$R_{0,1}(h) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} L_{0,1}(h, y_i)}$$

$$R_{0,1}(y_1) = \frac{1}{n}\Big(\overbrace{L_{0,1}(y_1,y_1)}^{0} + \overbrace{L_{0,1}(y_1,y_2)}^{1} + \overbrace{L_{0,1}(y_1,y_3)}^{1} + \dots + \overbrace{L_{0,1}(y_1,y_n)}^{1}\Big)$$

$$\frac{1}{n}(n-1)$$

Question

Suppose $y_1, \dots, y_n$ are all distinct. What is the value of $R_{0,1}(y_1)$?

$h = y_1$

a) 0 b) $\frac{1}{n}$ $\boxed{\text{c) } \frac{n-1}{n}}$ d) 1

→ MODE

Minimizing Empirical Risk

ex.) data set $-3, -2, -2, 2$

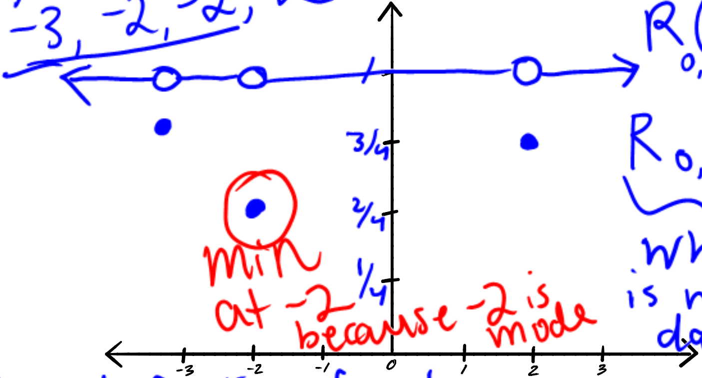
$$R_{0,1}(h) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}0, & \text{if } h = y_i\\ 1, & \text{if } h \neq y_i\end{cases}$$

$R_{0,1}(-3) = \frac{1}{4}\cdot 3$

$R_{0,1}(-2) = \frac{1}{4}\cdot 2$

$R_{0,1}(2) = \frac{1}{4}\cdot 3$

$R_{0,1}(-1) = \frac{1}{4}\cdot 4$



min at -2 because -2 is mode

when h is not a data pt, $R_{0,1}(h) = 1$

$R_{0,1}(h)$ what fraction of data is different from h

Different Loss Functions Lead to Different Predictions

40B

Loss	Minimizer	Outliers	Differentiable	Algorithm
L_{abs}	median	**insensitive**	**no**	**not simple**
L_{sq}	mean	**sensitive**	**yes**	**simple, fast**
$L_{0,1}$	mode	**insensitive**	**no**	**simple, fast**

- The optimal predictions are all **summary statistics** that measure the **center** of the data set in different ways.

Summary

- The mean error and the mean squared error fit into a general framework of **empirical risk minimization**.
- By changing the loss function, we change which prediction is considered the best.
- The optimal predictions each measure the **center** of the data set.
- **Next Time:** We’ll design a more complicated loss function.