

Module 3 – Mean Squared Error and Empirical Risk Minimization



DSC 40A, Summer 2023

Agenda

- ▶ Recap from Module 2 – minimizing mean absolute error and formulating mean squared error.
- ▶ Minimizing mean squared error.
- ▶ Comparing different minimizers.
- ▶ Empirical risk minimization.

Recap from Module 2

The median minimizes mean absolute error

- ▶ Our problem was: find h^* which minimizes the mean

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

- ▶ **Regardless of if n is odd or even**, the answer is $h^* = \text{Median}(y_1, \dots, y_n)$. The **best prediction**, in terms of mean absolute error, is the **median**.
 - ▶ When n is odd, this answer is unique.
 - ▶ When n is even, any number between the middle two data points also minimizes mean absolute error.
 - ▶ We define the median of an even number of data points to be the mean of the middle two data points.

The mean absolute error is **not differentiable**

- ▶ We can't compute $\frac{d}{dh} |y_i - h|$.
- ▶ Remember: $|y_i - h|$ measures how far h is from y_i .
- ▶ Is there something besides $|y_i - h|$ which:
 1. Measures how far h is from y_i , and
 2. is **differentiable**?

The mean absolute error is **not differentiable**

- ▶ We can't compute $\frac{d}{dh} |y_i - h|$.
- ▶ Remember: $|y_i - h|$ measures how far h is from y_i .
- ▶ Is there something besides $|y_i - h|$ which:
 1. Measures how far h is from y_i , and
 2. is **differentiable**?

Discussion Question

Which of these would work?

a) $e^{|y_i - h|}$

b) $|y_i - h|^2$

c) $|y_i - h|^3$

d) $\cos(y_i - h)$

The squared error

- ▶ Let h be a prediction and y be the true value (i.e. the “right answer”). The **squared error** is:

$$|y - h|^2 = (y - h)^2$$

- ▶ Like absolute error, squared error measures how far h is from y .
- ▶ But unlike absolute error, the squared error is **differentiable**:

$$\frac{d}{dh}(y - h)^2 =$$

The new idea

- ▶ Find h^* by minimizing the **mean squared error**:

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

- ▶ Strategy: Take the derivative, set it equal to zero, and solve for the minimizer.

Minimizing mean squared error

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

Discussion Question

Which of these is dR_{sq}/dh ?

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

b) 0

c) $\sum_{i=1}^n y_i$

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

Solution

$$\frac{dR_{sq}}{dh} = \frac{d}{dh} \left[\frac{1}{n} \sum_{i=1}^n (y_i - h)^2 \right]$$

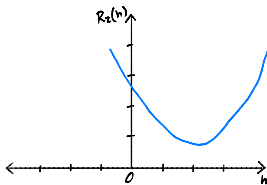
Set to zero and solve for minimizer

The mean minimizes mean squared error

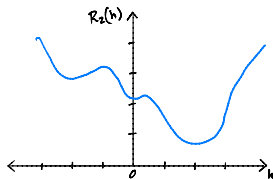
- ▶ Our new problem was: find h^* which minimizes the mean squared error, $R_{sq}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$.
 - ▶ The answer is: $\text{Mean}(y_1, \dots, y_n)$.
 - ▶ The **best prediction**, in terms of mean squared error, is the **mean**.
 - ▶ This answer is always unique!

Discussion Question

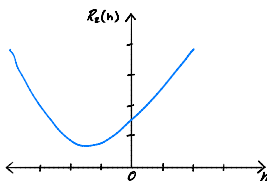
Suppose y_1, \dots, y_n are salaries. Which plot could be $R_{sq}(h)$?



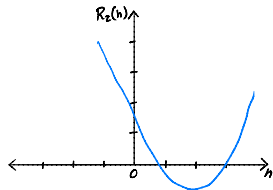
(a)



(b)



(c)



(d)

Comparing the median and mean

Outliers

- ▶ Consider our original dataset of 5 salaries.

90,000 94,000 96,000 120,000 160,000

- ▶ As it stands, the **median is 96,000** and the **mean is 112,000**.

- ▶ What if we add 300,000 to the largest salary?

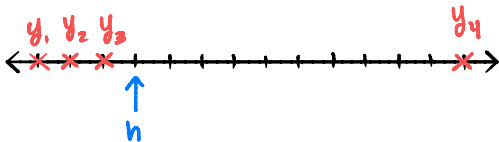
90,000 94,000 96,000 120,000 460,000

- ▶ Now, the **median is still 96,000** but the **mean is 172,000!**

- ▶ **Key Idea:** The mean is quite **sensitive** to outliers.

Outliers

- ▶ The mean is quite **sensitive** to outliers.

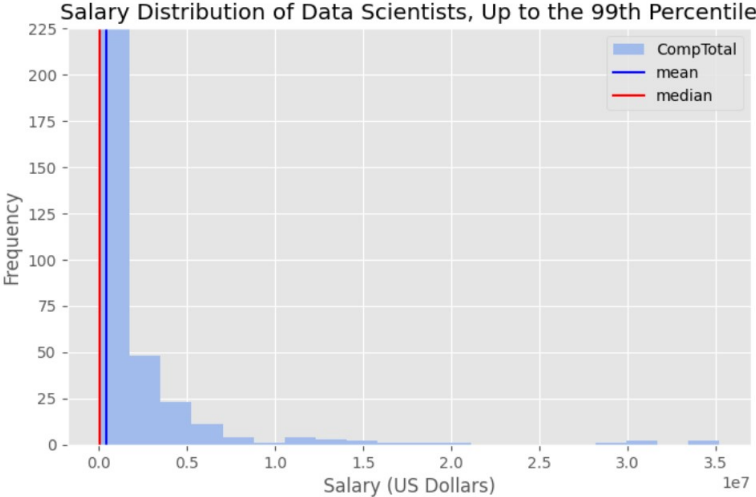


- ▶ $|y_4 - h|$ is 10 times as big as $|y_3 - h|$.
- ▶ But $(y_4 - h)^2$ is 100 times as big as $(y_3 - h)^2$.
 - ▶ This “pulls” h^* towards y_4 .
- ▶ Squared error can be dominated by outliers.

Example: Data Scientist Salaries

- ▶ Dataset of 2,016 self-reported data science salaries in the United States from the 2022 StackOverflow survey.
- ▶ Median = \$86,700.
- ▶ Mean = \$501,425,531.
- ▶ Min = \$20.
- ▶ Max = \$1,000,000,000,000.
- ▶ 90th Percentile: \$700,000.

Example: Data Scientist Salaries



Example: Income Inequality

Average vs median income

Median and mean income between 2012 and 2014 in selected OECD countries, in USD; weighted by the currencies' respective [purchasing power](#) (PPP).

■ Average income in USD ■ Median income

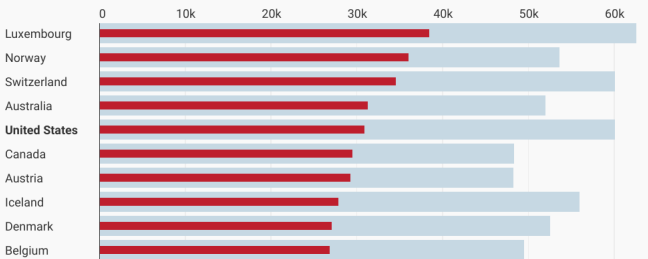
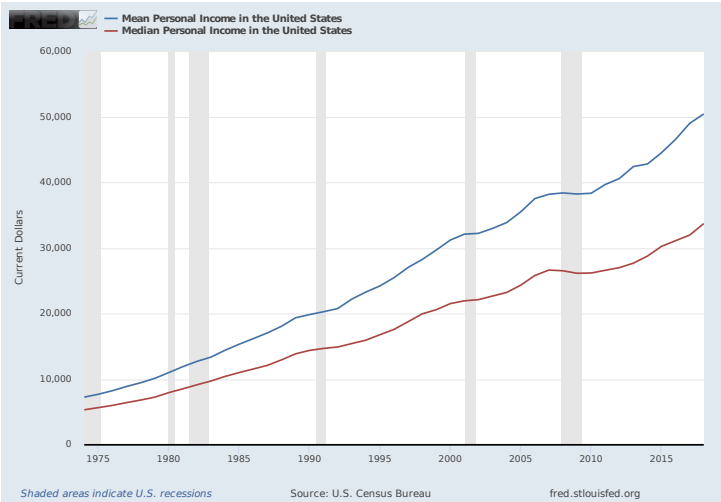


Chart: Lisa Charlotte Rost, Datawrapper

Example: Income Inequality



Empirical risk minimization

A general framework

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

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

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

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

- ▶ 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 true value (i.e. a “right answer”), y , and outputs a number measuring how far h is from y (bigger = further).
- ▶ The **absolute loss**:

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

- ▶ The **squared loss**:

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

A general framework

- ▶ Suppose that y_1, \dots, 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:

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

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

The learning recipe

1. Pick a loss function.
2. Pick a way to minimize the average loss (i.e. empirical risk) on the data.
 - ▶ **Key Idea:** The choice of loss function determines the properties of the result. **Different loss function = different minimizer = different prediction!**
 - ▶ Absolute loss yields the median.
 - ▶ Squared loss yields the mean.
 - ▶ The mean is easier to calculate but is more sensitive to outliers.

Example: 0-1 Loss

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

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

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**:

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

2. Minimize empirical risk:

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

Discussion Question

Suppose y_1, \dots, y_n are all distinct. Find $R_{0,1}(y_1)$.

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

Minimizing empirical risk

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

Different loss functions lead to different predictions

Loss	Minimizer	Outliers	Differentiable
L_{abs}	median	insensitive	no
L_{sq}	mean	sensitive	yes
$L_{0,1}$	mode	insensitive	no

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

Summary

Summary

- ▶ $h^* = \text{Mean}(y_1, \dots, y_n)$ minimizes $R_{sq}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$, i.e. the mean minimizes mean squared error.
- ▶ The mean absolute error and the mean squared error fit into a general framework called **empirical risk minimization**.
 - ▶ Pick a loss function. We've seen absolute loss, $|y - h|$, squared loss, $(y - h)^2$, and 0-1 loss.
 - ▶ Pick a way to minimize the average loss (i.e. empirical risk) on the data.
- ▶ By changing the loss function, we change which prediction is considered the best.

Next time

- ▶ **Spread** – what is the meaning of the value of $R_{abs}(h^*)$?
 $R_{sq}(h^*)$?
- ▶ Creating a new loss function and trying to minimize the corresponding empirical risk.
 - ▶ We'll get stuck and have to look for a new way to minimize.