Module 4 – Center and Spread, Other Loss Functions

DSC 40A, Summer 2023

Announcements

- Homework 1 is due **July 11 at 11:59pm**.
  - LaTeX template available if you want to type your answers.
  - Make sure to explain your answers! Don't just write a number; show how you got it.
- Discussion section is on Friday.

Agenda

- Recap of empirical risk minimization.
- Center and spread.
- A new loss function.

Recap of empirical risk minimization

Empirical risk minimization

- **Goal**: Given a dataset $y_1, y_2, ..., y_n$, determine the best prediction h^*.

- Strategy:
  1. Choose a **loss function**, $L(h, y)$, that measures how far any particular prediction h is from the "right answer" y.

     cost

  2. Minimize **empirical risk** (also known as average loss) over the entire dataset. The value(s) of h that minimize empirical risk are the resulting "best predictions".

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

Absolute loss and squared loss

- General form of empirical risk:

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

- **Absolute loss**: $L_{\text{abs}}(h, y) = |y - h|$.
  - Empirical risk: $R_{abs}(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|$. Also called "**mean absolute error**".
  - Minimized by $h^* =$ **Median**$(y_1, y_2, ..., y_n)$.
- **Squared loss**: $L_{\text{sq}}(h, y) = (y - h)^2$.
  - Empirical risk: $R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n} (y_i - h)^2$. Also called "**mean squared error**".
  - Minimized by $h^* =$ **Mean**$(y_1, y_2, ..., y_n)$.

Discussion Question

Consider a dataset $y_1, y_2, ..., y_n$.
Recall,

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

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

Is it true that, for any h, $[R_{abs}(h)]^2 = R_{sq}(h)$?

a) True

b) False

Center and spread

What does it mean?

- General form of empirical risk:

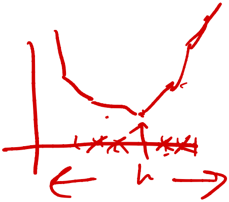
$$R(h) = \frac{1}{n} \sum_{i=1}^{n} L(h, y_i)$$

Single value

- The input h^* that minimizes $R(h)$ is some measure of the **center** of the data set.
  - e.g. median, mean, mode.

- 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

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

- $R_{abs}(h)$ is minimized at $h^* = \text{Median}(y_1, y_2, \dots, y_n)$.

- Therefore, the minimum value of $R_{abs}(h)$ is

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

Mean absolute deviation from the median

- The minimum value of $R_{abs}(h)$ is the **mean absolute deviation from the median**.

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

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

Median = 3

$\frac{1+0+0+1}{4}$

$\frac{2}{4}$

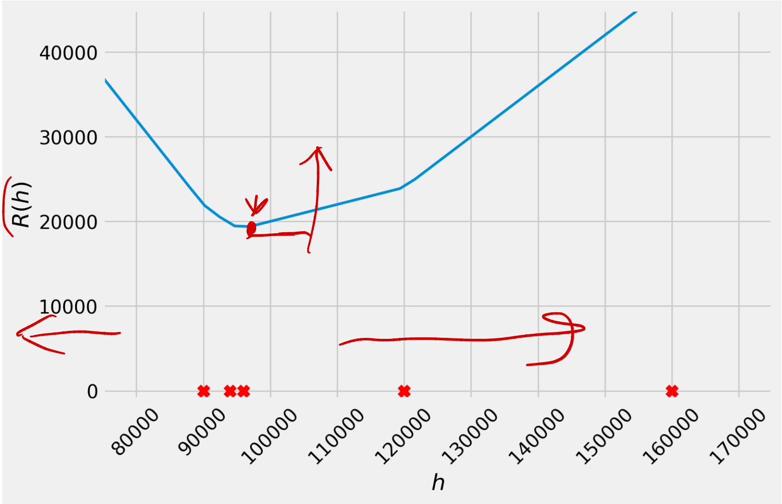
$\frac{1}{2}$

Discussion 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



Squared loss

- The empirical risk for the squared loss is

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

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

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

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

Variance

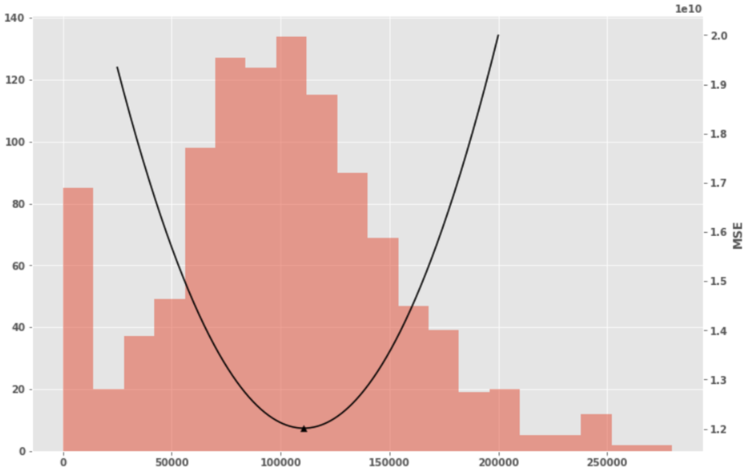
- The minimium value of $R_{\text{sq}}(h)$ is the mean squared deviation from the mean, more commonly known as the **variance**.

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

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

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

Variance



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

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 the proportion (between 0 and 1) of data points not equal to h.

- $R_{0,1}(h)$ is minimized at $h^* = \text{Mode}(y_1, y_2, \dots, y_n)$.

- Therefore, $R_{0,1}(h^*)$ is the proportion of data points not equal to the mode.

A poor way to measure spread

- The minimium value of $R_{0,1}(h)$ is the proportion of data points not equal to the mode.

- A 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.

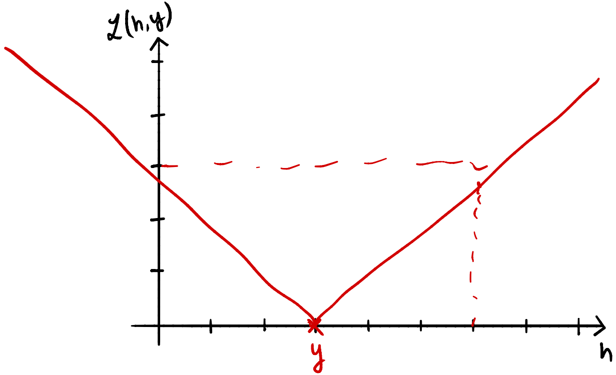
Summary of center and spread

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

A new loss function

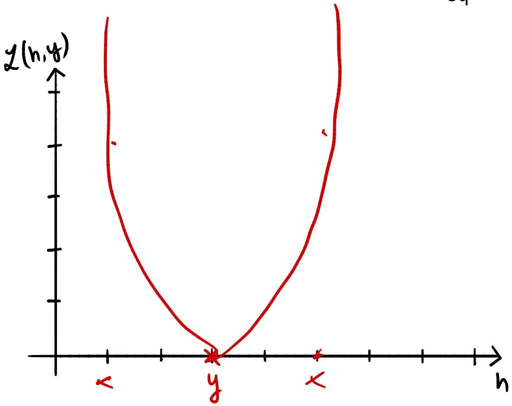
Plotting a loss function

- The plot of a loss function tells us how it treats outliers.
- Consider y to be some fixed value. Plot $L_{\text{abs}}(h, y) = |y - h|$:



Plotting a loss function

- The plot of a loss function tells us how it treats outliers.
- Consider y to be some fixed value. Plot $L_{sq}(h, y) = (y - h)^2$:



Discussion Question

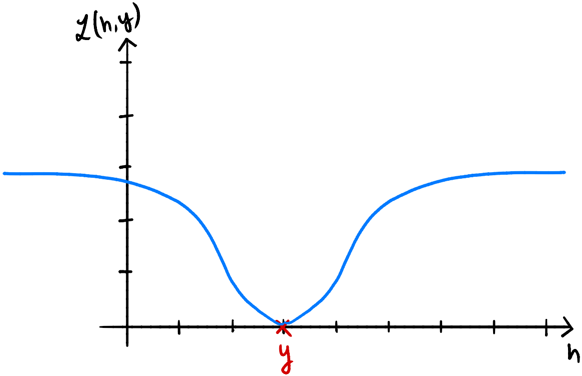
Suppose L considers all outliers to be equally bad. What would it look like far away from y?

a) flat

b) rapidly decreasing

c) rapidly increasing

A very insensitive loss



- We'll call this loss L_{ucsd} because we made it up at UCSD.

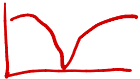
Discussion Question

Which of these could be $L_{ucsd}(h, y)$?

a) $\cancel{e^{-(y-h)^2}}$

b) $1 - e^{-(y-h)^2}$

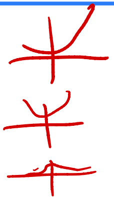
c) $\cancel{1 - (y - h)^2}$

d) $\cancel{1 - e^{-|y-h|}}$ $\leftarrow$ 

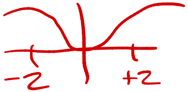
$e^{x} \rightarrow$

e^{x^2}

e^{-x^2}



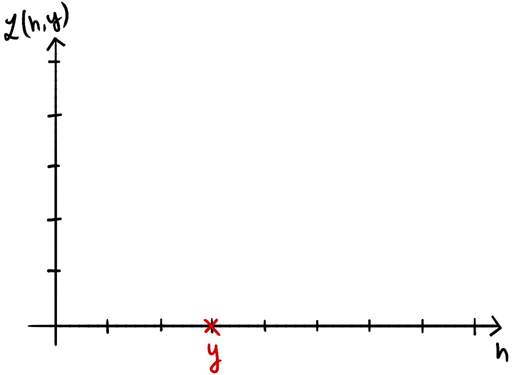
Adding a scale parameter



- Problem: L_{ucsd} has a fixed scale. This won't work for all datasets.
  - If we're predicting temperature, and we're off by 100 degrees, that's bad.
  - If we're predicting salaries, and we're off by 100 dollars, that's pretty good.
  - What we consider to be an outlier depends on the scale of the data.
- Fix: add a **scale parameter**, σ:

$$L_{ucsd}(h, y) = 1 - e^{-(y-h)^2/\sigma^2}$$

Scale parameter controls width of bowl



Empirical risk minimization

- We have salaries $y_1, y_2, ..., y_n$.
- To find optimal prediction, ERM says to minimize the average loss:

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

$$= \frac{1}{n} \sum_{i=1}^{n} \left[1 - e^{-(y_i - h)^2 / \sigma^2} \right]$$

Let's plot R_{ucsd}

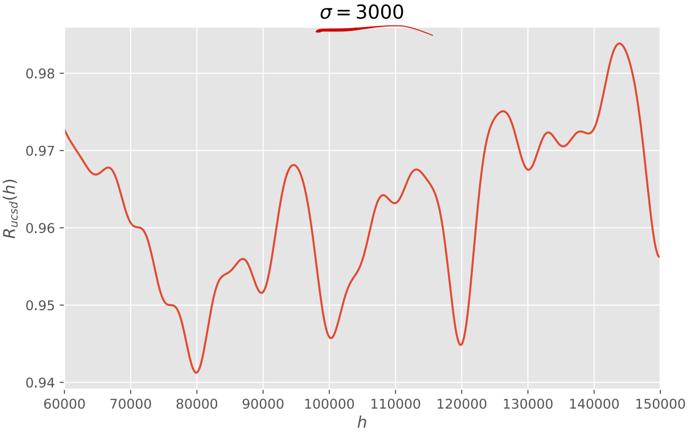
- Recall:

$$R_{ucsd}(h) = \frac{1}{n} \sum_{i=1}^{n} \left[1 - e^{-(y_i - h)^2 / \sigma^2} \right]$$

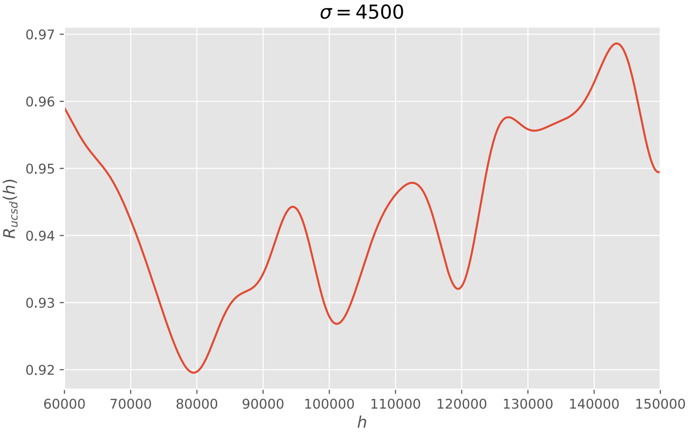
- Once we have data $y_1, y_2, ..., y_n$ and a scale σ, we can plot $R_{ucsd}(h)$.

- Let's try several scales, σ, for the data scientist salary data.

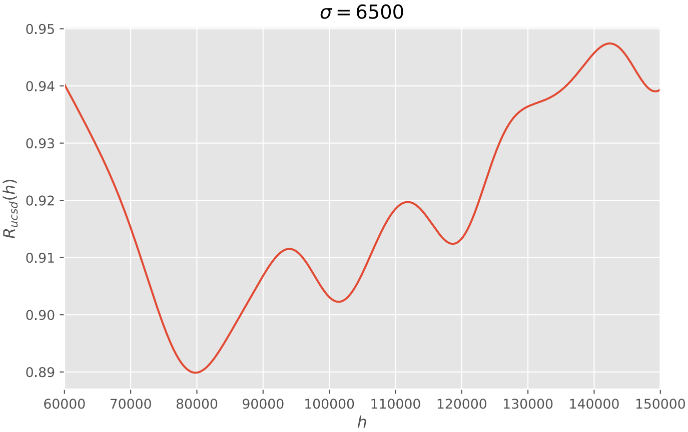
Plot of $R_{ucsd}(h)$



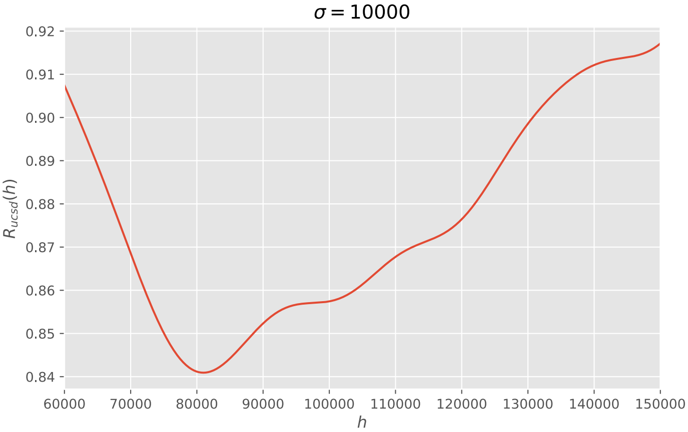
Plot of $R_{ucsd}(h)$



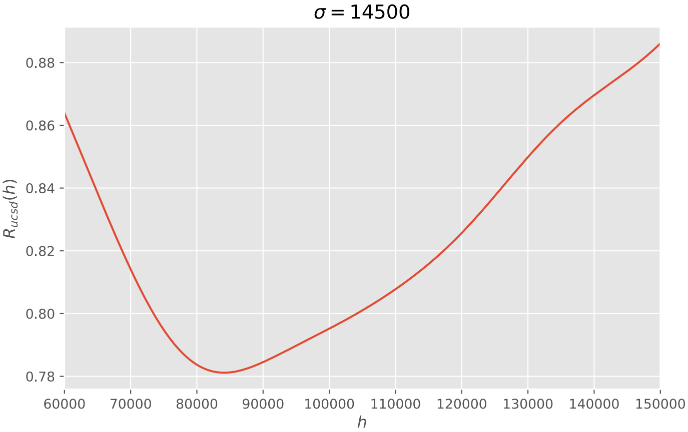
Plot of $R_{ucsd}(h)$



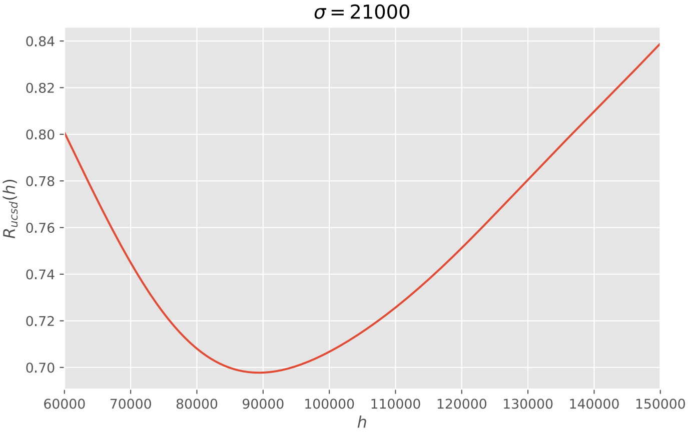
Plot of $R_{ucsd}(h)$



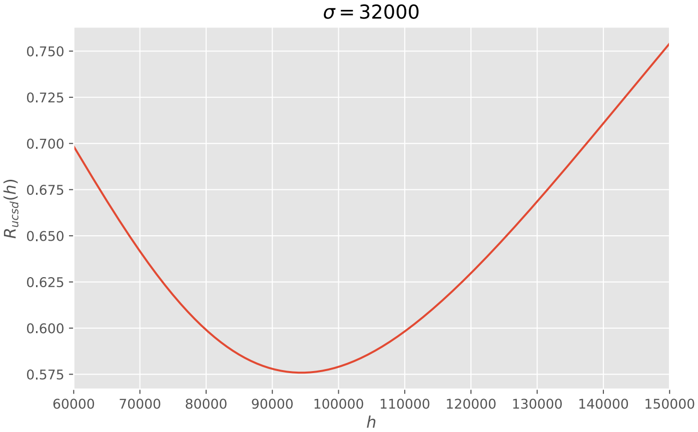
Plot of $R_{ucsd}(h)$



Plot of $R_{ucsd}(h)$



Plot of $R_{ucsd}(h)$



Minimizing R_{ucsd}

- To find the best prediction, we find h^* minimizing $R_{ucsd}(h)$.
- $R_{ucsd}(h)$ is **differentiable**.
- To minimize: take derivative, set to zero, solve.

Step 1: Taking the derivative

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

Step 2: Setting to zero and solving

- We found:

$$\frac{d}{dh} R_{ucsd}(h) = \frac{2}{n\sigma^2} \sum_{i=1}^{n} (h - y_i) \cdot e^{-(h-y_i)^2/\sigma^2}$$

- Now we just set to zero and solve for h:

$$0 = \frac{2}{n\sigma^2} \sum_{i=1}^{n} (h - y_i) \cdot e^{-(h-y_i)^2/\sigma^2}$$

- We **can** calculate derivative, but we **can't** solve for h; we're stuck again.

Summary

- Different loss functions lead to empirical risk functions that are minimized at various measures of **center**.
- The minimum values of these empirical risk functions are various measures of **spread**.
- We came up with a more complicated loss function, L_{ucsd}, that treats all outliers equally.
  - We weren't able to minimize its empirical risk R_{ucsd} by hand.
- **Next Time:** We'll learn a computational tool to approximate the minimizer of R_{ucsd}.