Midterm Review Session

DSC 40A, Fall 2021 @ UC San Diego

Suraj Rampure, with help from **many others**

Agenda

- Bird's eye view of the course.
- Select past exam problems.
- Review of select problems from Homeworks 1, 2, and 3.

Bird’s eye view of the course

What is this course about?

- So far, this course has really been about one thing: **learning from data**.
- The recipe:
    1. Choose a prediction rule.
    2. Choose a loss function.
    3. Minimize empirical risk, i.e. average loss, on your dataset to find the best predictions/parameters.
- Let's look at all of this a little more deeply.

Choosing a prediction rule

In lecture, we've studied two prediction rules in depth:

- The **constant hypothesis**, h (Lectures 1-5).
  - We didn't call it a "prediction rule" at the time, but it is one.
  - Equivalent to saying $H(x) = h$, i.e. we predict the same output for everyone.
- The **simple linear** prediction rule, $H(x) = w_0 + w_1 x$ (Lectures 6-7).
  - Now, predictions vary depending on x (x is called a **feature**).
- You also looked at $H(x) = w_1 x$ in Homework 3, 1f.

Some questions...

- Suppose I've chosen to use a constant prediction rule h. Which h do I use?
- Suppose I've chosen to use a simple linear prediction rule $H(x) = w_0 + w_1 x$. What should w_0 and w_1 be?
- **Answer:** Loss functions can help us.

Loss functions

- A **loss function** $L(h, y)$ measures how a prediction h is from the truth y.
- We've seen several loss functions so far:
  - Absolute loss: $L(h, y) = |y - h|$.
  - Squared loss: $L(h, y) = (y - h)^2$.
  - UCSD loss: $L(h, y) = 1 - e^{-(y-h)^2/\sigma^2}$.
  - 0-1 loss: $L(h, y) = 0$ if h = y, otherwise 1.
- Different loss functions have different properties, the key ones being their ease of minimization and their robustness to outliers.

Empirical risk

- Loss functions are great — but they only measure the quality of a single prediction for a single true value.

- In order to get a sense of the quality of a prediction on our entire dataset, we must take the average of our chosen loss function over our entire dataset.

- The result is called empirical risk:

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

  - If using absolute loss, R is called **mean absolute error**.

  - If using squared loss, R is called **mean squared error**.

The constant hypothesis, h

To find the best h (denoted as h^*) to make constant predictions with, we need to choose a loss function.

- If we choose absolute loss, the resulting empirical risk (i.e. mean absolute error) is

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

- If we choose squared loss, the resulting empirical risk (i.e. mean squared error) is

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

- We also looked at the resulting empirical risk when we choose 0-1 loss and UCSD loss.

The simple linear prediction rule, $H(x) = w_0 + w_1 x$

w_0 and w_1 are called **parameters**. To find the **optimal parameters** (denoted as w_0^* and w_1^*), we again need a loss function

- We could choose absolute loss — see Homework 3, Q3.
  - The resulting problem is called "least absolute deviations regression."

- The more common choice, though, is squared loss:

$$R_{sq}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

  - This problem is called **least squares regression**.

Minimizing empirical risk

After choosing a prediction rule and loss function, and writing out the corresponding empirical risk, we need to **minimize** the empirical risk to find the best predictions/parameters.
Some ways we've minimized empirical risk:

- Calculus.
- Other algebraic arguments.
- Gradient descent.

Minimizing empirical risk with calculus

- Strategy: take derivative(s), set it to 0, and solve.

- Constant hypothesis, squared loss:

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

- Simple linear prediction rule, squared loss:

$$R_{sq}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

$$\implies w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r\frac{\sigma_y}{\sigma_x} \qquad w_0^* = \bar{y} - w_1^*\bar{x}$$

- Several homework problems.
  - HW 2 Q3, HW 3 Q1f.

Minimizing empirical risk with algebraic arguments

- Since absolute loss is not differentiable, the resulting empirical risk (mean absolute error) also isn't. We couldn't use calculus.
- For the constant hypothesis, $R_{abs}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$.
- We instead minimized R_{abs} by finding a formula for the slope of R at any h (that isn't one of the y_i):

$$\text{slope of R at h} = \frac{1}{n}\left(\#(y_i < h) - \#(y_i > h)\right)$$

- The **median** is where the slope of R goes from - to +; it minimizes $R_{abs}(h)$.

Minimizing empirical risk using gradient descent

- Sometimes, even when our empirical risk is differentiable, there is no **closed-form solution** for the minimizing input.
  - Example: Empirical risk for L_{ucsd}.

- Solution: **gradient descent**.

- Gradient descent tries to minimize a function $R(h)$ through an iterative process.
  - **Key idea**: Move opposite the direction of the slope.

  - Given an initial guess, h_0, for the minimizer and a step-size/learning rate α, gradient descent updates are made with the update equation

$$h_i = h_{i-1} - \alpha \cdot \frac{d}{dh} R(h_{i-1})$$

Gradient descent

- **Key theorem:** Gradient descent is guaranteed to find the global minimum of a function if that function is **convex** and **differentiable**, given an appropriate step size.

- A function f is **convex** if it is true that given any two inputs a, b, the line segment joining $(a, f(a))$ and $(b, f(b))$ does not go below the graph of f.
  - Convex functions are "bowl" shaped.

  - Second derivative test.

Summary of key results

Other concepts — spread

- Different loss functions lead to empirical risk functions that, for the constant prediction rule, are minimized at various measures of **center**.
  - Absolute loss: **median**.
  - Squared loss: **mean**.
  - 0-1 loss: **mode**.

- The minimum value of these empirical risks (i.e. the lowest height on the graph of R) is a measure of the **spread** of the data.
  - Absolute loss: **median absolute deviation from the median**.
  - Squared loss: **variance**.
  - 0-1 loss: **proportion of values not equal to the mode**.

Other concepts — correlation

- The correlation coefficient, r, is a measure of the **linear association between two variables**.

- It ranges between -1 and 1.
  - $r = 1$ indicates a perfect positive linear association (x and y lie exactly on a straight line that is sloped upwards).
  - $r = -1$ indicates a perfect negative linear association between x and y.
  - The closer r is to 0, the weaker the linear association between x and y is.

- w_1^* can be written in terms of r:

$$w_1^* = r\frac{\sigma_y}{\sigma_x}$$