Lecture 7 – Linear Prediction Rules

DSC 40A, Spring 2023

Announcements

- Homework 2 is due **tomorrow at 11:59pm**.
    - LaTeX template provided if you want to type your answers.
    - Please come to office hours!
- Review Homework 1 solutions on Campuswire.
- Discussion section is on Wednesday.

Agenda

- Recap of convexity.
- Prediction rules.
- Minimizing mean squared error, again.

Recap: convexity

Convexity: Definition

formal

- A function $f : \mathbb{R} \to \mathbb{R}$ is **convex** if for every choice of a, b and $t \in [0, 1]$:

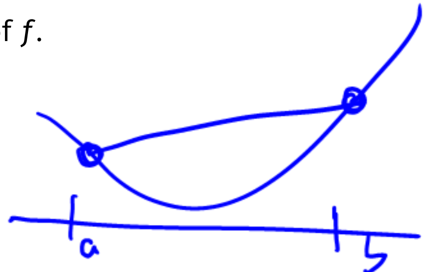
$$(1 - t)f(a) + tf(b) \geq f((1 - t)a + tb)$$

informal

- This means that for **every** a, b in the domain of f, the line segment between

$$(a, f(a)) \qquad \text{and} \qquad (b, f(b))$$

does not go below the plot of f.



Second derivative test for convexity

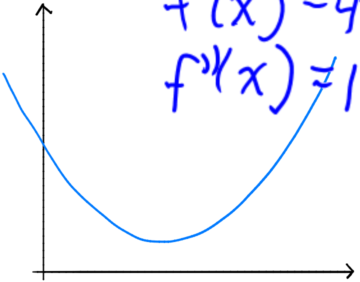
- If $f(x)$ is a function of a single variable and is twice differentiable, then:

- $f(x)$ is convex if and only if $\frac{d^2 f}{dx^2}(x) \geq 0$ for all x.

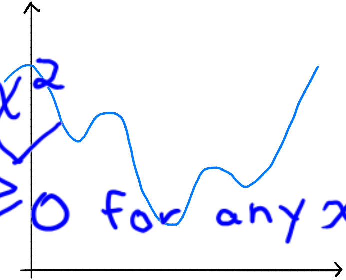
- Example: $f(x) = x^4$ is convex.

$f'(x) = 4x^3$

$f''(x) = 12x^2$ ✓ ≥ 0 for any x



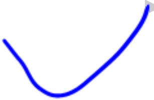
Convex



Non-convex

Convexity and gradient descent

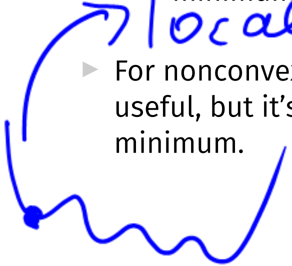
- **Theorem**: if $R(h)$ is convex and differentiable then gradient descent converges to a **global minimum** of R *provided* that the step size is small enough.
  - If a function is convex and has a local minimum, that local minimum must be a global minimum.



  - In other words, gradient descent won't get stuck/terminate in local minimums that aren't global minimums.

local min/not global

- For nonconvex functions, gradient descent can still be useful, but it's not guaranteed to converge to a global minimum.



Convexity of empirical risk

- If $L(h, y)$ is a convex function (when y is fixed) then

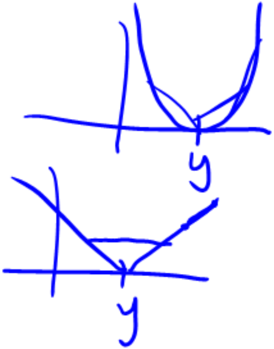
$$\underset{\text{convex}}{R(h)} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{L(h, y_i)}_{\text{convex}}$$

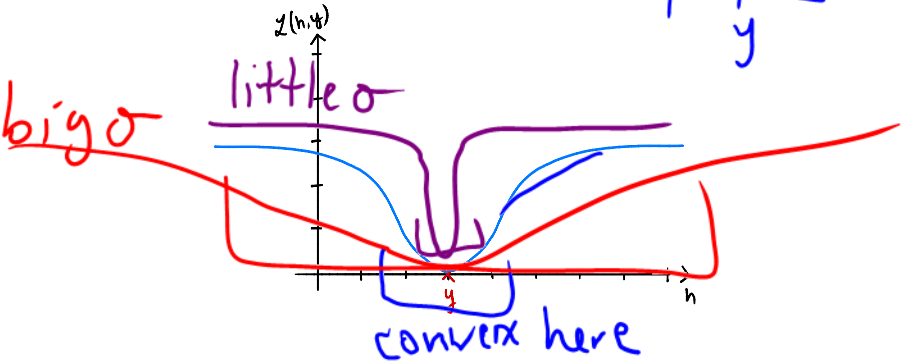
  is convex.
  - More generally, sums of convex functions are convex.

- What does this mean?
  - If a loss function is convex, then the corresponding empirical risk will also be convex.

Convexity of loss functions

- Is $L_{sq}(h, y) = (y - h)^2$ convex? **Yes** or **No**.
- Is $L_{abs}(h, y) = |y - h|$ convex? **Yes** or **No**.
- Is $L_{ucsd}(h, y)$ convex? **Yes** or **No**.

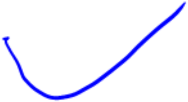




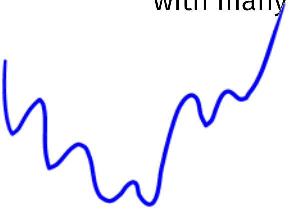
Convexity of R_{ucsd}

- A function can be convex in a region.

- If σ is large, $R_{ucsd}(h)$ is convex in a big region around data.
  - A large σ led to a very smooth, parabolic-looking empirical risk function with a single local minimum (which was a global minimum).



- If σ is small, $R_{ucsd}(h)$ is convex in only small regions.
  - A small σ led to a very bumpy empirical risk function with many local minimums.



bowl of line segments

Discussion Question

Recall the empirical risk for absolute loss,

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

Is $R_{abs}(h)$ **convex**? Is gradient descent **guaranteed** to find a global minimum, given an appropriate step size?

a) **YES** convex, **YES** guaranteed
b) **YES** convex, **NOT** guaranteed
c) ~~**NOT** convex, **YES** guaranteed~~
d) ~~**NOT** convex, **NOT** guaranteed~~

not differentiable

Prediction rules

How do we predict someone's salary?

After collecting salary data, we...

1. Choose a loss function.

2. Find the best prediction by minimizing the average loss across the entire data set (empirical risk).

- So far, we've been predicting future salaries without using any information about the individual (e.g. GPA, years of experience, number of LinkedIn connections).

- **New focus:** How do we incorporate this information into our prediction-making process?

Features

A **feature** is an attribute – a piece of information.

weird one

- **Numerical**: age, height, years of experience
- **Categorical**: college, city, education level
- **Boolean**: knows Python?, had internship?

Think of features as columns in a DataFrame or table.

not feature

	YearsExperience	Age	FormalEducation	Salary
0	6.37	28.39	Master's degree (MA, MS, M.Eng., MBA, etc.)	120000.0
1	0.35	25.78	Some college/university study without earning ...	120000.0
2	4.05	31.04	Bachelor's degree (BA, BS, B.Eng., etc.)	70000.0
3	18.48	38.78	Bachelor's degree (BA, BS, B.Eng., etc.)	185000.0
4	4.95	33.45	Master's degree (MA, MS, M.Eng., MBA, etc.)	125000.0

Variables

- The features, *x*, that we base our predictions on are called **predictor variables**.
  x's (could be many)
- The quantity, *y*, that we're trying to predict based on these features is called the **response variable**.
  only one y
- We'll start by predicting salary based on years of experience.
  y
  x

Prediction rules

- We believe that salary is a function of experience.
- In other words, we think that there is a function H such that:

$$\text{salary} \approx H(\text{years of experience})$$

- H is called a **hypothesis function** or **prediction rule**.
- **Our goal**: find a good prediction rule, H.

Possible prediction rules

$$H_1(\text{years of experience}) = \$50{,}000 + \$2{,}000 \times (\text{years of experience})$$

$$H_2(\text{years of experience}) = \$60{,}000 \times 1.05^{(\text{years of experience})}$$

$$H_3(\text{years of experience}) = \$100{,}000 - \$5{,}000 \times (\text{years of experience})$$

- These are all valid prediction rules.
- Some are better than others.

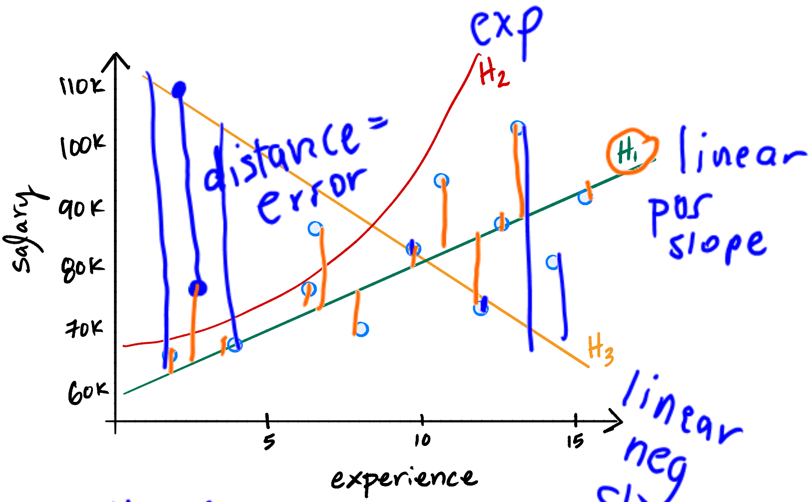
Comparing predictions

- How do we know which prediction rule is best: H_1, H_2, H_3?
- We gather data from n people. Let x_i be experience, y_i be salary:

$$\begin{matrix} (\text{Experience}_1, \text{Salary}_1) \\ (\text{Experience}_2, \text{Salary}_2) \\ \dots \\ (\text{Experience}_n, \text{Salary}_n) \end{matrix} \quad \rightarrow \quad \begin{matrix} (x_1, y_1) \\ (x_2, y_2) \\ \dots \\ (x_n, y_n) \end{matrix}$$

- See which rule works better on data.

Example



H_3 generally bad - many large errors

Quantifying the quality of a prediction rule H

x_i = exp.
y_i = salary
$H(x_i)$ = pred. salary

- Our prediction for person i's salary is $H(x_i)$.
- As before, we'll use a **loss function** to quantify the quality of our predictions.
  - Absolute loss: $|y_i - H(x_i)|$.

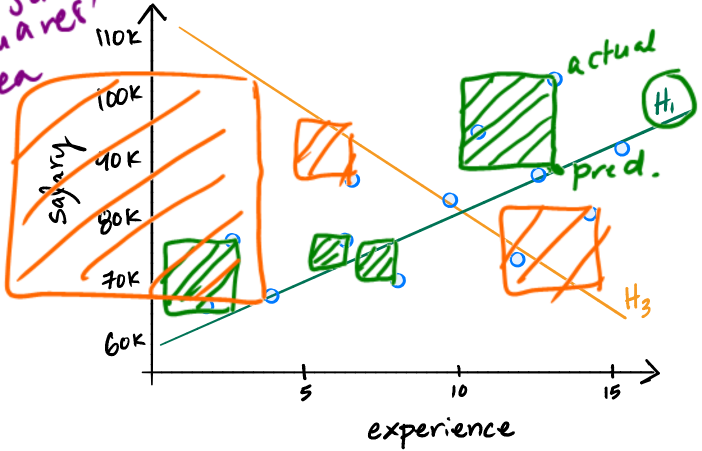
    actual pred.
  - Squared loss: $(y_i - H(x_i))^2$.
- We'll focus on squared loss, since it's differentiable.
- Using squared loss, the **empirical risk** (mean squared error) of the prediction rule H is:

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

MSE

Mean squared error

avg of all such squares' area



H_1 = generally small-area squares

H_3 = large squares

Finding the best prediction rule

- **Goal:** out of all functions $\mathbb{R} \to \mathbb{R}$, find the function H^* with the smallest mean squared error.

- That is, H^* should be the function that minimizes

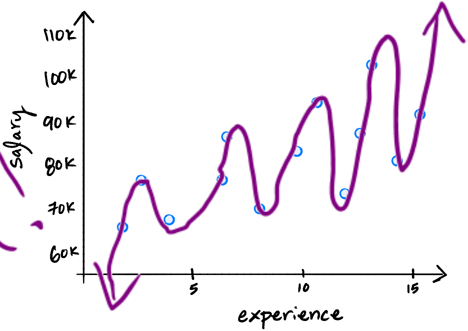
$$R_{sq}(H) = \frac{1}{n} \underbrace{\sum_{i=1}^{n} \left(y_i - H(x_i)\right)^2}_{\geq 0}$$

Discussion Question

Given the data below, is there a prediction rule H which has **zero** mean squared error?

a) Yes   b) No

it's stupid!



high degree poly.

Problem

- We can make mean squared error very small, even zero!
- But the function will be weird.
- This is called **overfitting**. doesn't generalize well
- Remember our real goal: make good predictions on data **we haven't seen**.

Solution

- Don't allow H to be just any function.
- Require that it has a certain form.

next

- Examples:
  - Linear: $H(x) = w_0 + w_1 x$.
  - Quadratic: $H(x) = w_0 + w_1 x + w_2 x^2$.
  - Exponential: $H(x) = w_0 e^{w_1 x}$.
  - Constant: $H(x) = w_0$.

simplest

family of functions

Finding the best linear prediction rule

- **Goal:** out of all **linear** functions $\mathbb{R} \to \mathbb{R}$, find the function H^* with the smallest mean squared error.
  - Linear functions are of the form $\boxed{H(x) = w_0 + w_1 x}$.

    $y = mx + b$
  - They are defined by a slope (w_1) and intercept (w_0).

    # for each year of experience

    Starting salary for new grad
- That is, H^* should be the linear function that minimizes

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

least squares

- This problem is called **linear regression**.
  - **Simple** linear regression refers to linear regression with a single predictor variable, x.

    vs. multiple

Minimizing mean squared error for the linear prediction rule

Minimizing the mean squared error

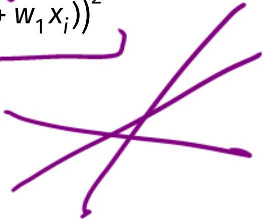
- The MSE is a function R_{sq} of a function H.

$$R_{\text{sq}}(H) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - H(x_i)\right)^2$$

- But since H is linear, we know $H(x_i) = w_0 + w_1 x_i$.

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

- Now R_{sq} is a function of w_0 and w_1.



- We call w_0 and w_1 **parameters**.
  - Parameters define our prediction rule.

Updated goal

- Find the slope w_1^* and intercept w_0^* that minimize the MSE, $R_{\text{sq}}(w_0, w_1)$:

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

- Strategy: multivariable calculus.

Recall: the gradient

→ reg. math notation

- If $f(x, y)$ is a function of two variables, the **gradient** of f at the point (x_0, y_0) is a **vector** of **partial derivatives**:

$$\nabla f(x_0, y_0) = \begin{pmatrix} \frac{\partial f}{\partial x}(x_0, y_0) \\ \frac{\partial f}{\partial y}(x_0, y_0) \end{pmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

> system of eqns

- **Key Fact #1**: The derivative is to the tangent line as the gradient is to the tangent plane.

- **Key Fact #2**: The gradient points in the direction of the biggest increase.

- **Key Fact #3**: The gradient is zero at critical points.

Minimizing multivariable functions

- From calculus, to optimize a multivariable differentiable function:
  1. Calculate the gradient vector, or vector of partial derivatives.

  2. Set the gradient equal to to 0 (that is, the zero vector).

  3. Solve the resulting system of equations.

Example

Discussion Question

Find the point at which the function

$$f(x, y) = x^2 + y^2 - 2x - 4y$$

is minimized.

$\frac{\partial f}{\partial x} = 2x - 2 = 0$ gradient vector $= \begin{bmatrix} 2x-2 \\ 2y-4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$\frac{\partial f}{\partial y} = 2y - 4 = 0$ $2x-2=0$ $2y-4=0$

$x=1$ $y=2$

Summary

Summary, next time

- We introduced the linear prediction rule, $H(x) = w_0 + w_1 x$.
- To determine the best linear prediction rule, we'll use the squared loss and choose the one that minimizes the empirical risk, or mean squared error:

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

- **Next time**: We'll use calculus to minimize the mean squared error and find the best linear prediction rule.
  - Spoiler alert: it's the regression line, as we saw in DSC 10.