Lecture 8 – Simple Linear Regression

DSC 40A, Fall 2022 @ UC San Diego

Announcements

- Groupwork 2 is due **Today at 23:59pm.**
- HW 2 is due **Friday 10/14 at 2:00pm.**
- Midterm: 10/28 during class time.
  - Friday, 3-4PM, 4-5 PCYYNH 122.

Recap: Prediction Rule

Agenda

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

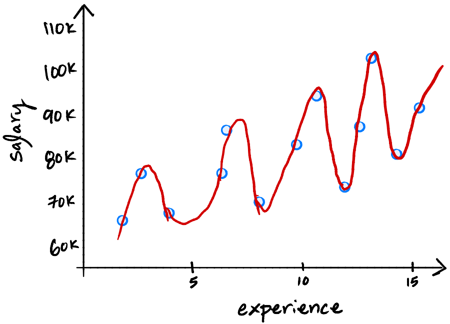
Finding the best prediction rule

- **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

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

- There's a problem.


110K
100K
90K
80K
70K
60K
salary
5
10
15
experience

Problem

- We can make mean squared error very small, even zero!
- But the function will be weird.
- This is called **overfitting**.
- 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.
- Examples:
  - Linear: $H(x) = w_0 + w_1 x$.
  - Quadratic: $H(x) = w_0 + w_1 x_1 + w_2 x^2$.
  - Exponential: $H(x) = w_0 e^{w_1 x}$.
  - Constant: $H(x) = w_0$.

h

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 $H(x) = w_0 + w_1 x$.

    slope

    $y = mx + b$

  - They are defined by a slope (w_1) and intercept (w_0).
- That is, H^* should be the linear function that minimizes

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

- This problem is called **least squares regression**.
  - "Simple linear regression" refers to linear regression with a single predictor variable.

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 $H(x_i) = w_0 + w_1 x_i$.

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

two

$R(h) \to h^*$

- Now R_{sq} is a function of w_0 and w_1.
- We call w_0 and w_1 **parameters**. of our model.
  - Parameters define our prediction rule.

Updated goal

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

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

- Strategy: multivariable calculus.

Recall: the gradient

x X

- 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**:

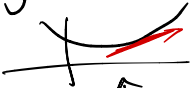
$f(x,y) = x^3 + xy + y^2$

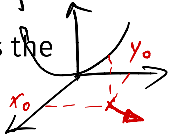
$\frac{\partial f}{\partial x} = 3x^2 + y + 0 = 0$

X

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

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





- **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.

Strategy

To minimize $R(w_0, w_1)$: compute the gradient, set it equal to zero, and solve.

$$\frac{\partial R_{sq}}{\partial w_0} = 0$$

$\Longrightarrow$ solve it!

$$\frac{\partial R_{sq}}{\partial w_1} = 0$$

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

Discussion Question

Choose the expression that equals $\dfrac{\partial R_{\text{sq}}}{\partial w_0}$.

a) $\displaystyle \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)$

b) $\displaystyle -\frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)$

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

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

Go to `menti.com` and enter the code 4821 5997.

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

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

$$= \frac{1}{n}\sum_{i=1}^{n} 2\left(\overset{0}{y_i} - \underbrace{(\overset{1}{w_0} + \overset{0}{w_1} x_i)}\right)^{2}(-1)$$

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

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

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

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

Strategy

$$-\frac{2}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right) = 0 \qquad -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right) x_i = 0$$

1. Solve for w_0 in first equation.
   - The result becomes w_0^*, since it is the "best intercept".

2. Plug w_0^* into second equation, solve for w_1.
   - The result becomes w_1^*, since it is the "best slope".

Solve for w_0^*

$$\cancel{\left(-\frac{n}{2}\right)}\cancel{\left(-\frac{2}{n}\right)}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i)) = 0 \quad \times\left(-\frac{n}{2}\right) \Rightarrow$$

years of exp $\to$ (x_i, y_i) $\leftarrow$ salary

$$w_0^* = \bar{y} - w_1 \bar{x}$$

$$\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right) = 0$$

$w_0 + w_0 + \cdots$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} w_0 - \sum_{i=1}^{n} w_1 x_i = 0 \Rightarrow$$

$$\sum_{i=1}^{n} y_i - n w_0 - w_1 \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow n w_0 = \sum_{i=1}^{n} y_i - w_1 \sum_{i=1}^{n} x_i$$

$$\Rightarrow w_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - w_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

Solve for w_1^*

$$\times\frac{-n}{2}\qquad -\frac{\cancel{2}}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)x_i = 0 \times\left(\frac{-n}{2}\right)$$

$$\sum_{i=1}^{n}\left(y_i - (\bar{y} - w_1\bar{x} + w_1 x_i)\right)x_i = 0$$

$$\sum_{i=1}^{n}\left[(y_i - \bar{y}) - w_1(x_i - \bar{x})\right]x_i = 0$$

$$\sum_{i=1}^{n}(y_i - \bar{y})x_i = w_1\sum_{i=1}^{n}(x_i - \bar{x})x_i$$

$$\Rightarrow \boxed{w_1^*} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}$$

Least squares solutions

- We've found that the values w_0^* and w_1^* that minimize the function $R_{sq}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$ are

1 2

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

slope intercept

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Let's re-write the slope w_1^* to be a bit more symmetric.

Key fact

The **sum of deviations from the mean** for any dataset is 0.

deviation

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \qquad \sum_{i=1}^{n}(y_i - \bar{y}) = 0$$

Proof:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$$

$$= \sum_{i=1}^{n} x_i - n\bar{x} = \underbrace{(n)\left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i}_{} - n\bar{x}$$

$$= n\bar{x} - n\bar{x} = 0$$

Equivalent formula for w_1^*

Claim

$$w_1^* = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Proof:

Least squares solutions

- The **least squares solutions** for the slope w_1^* and intercept w_0^* are:

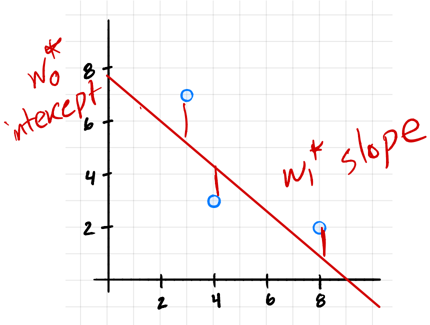
$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad\qquad w_0^* = \bar{y} - w_1\bar{x}$$

  - We also say that w_0^* and w_1^* are **optimal parameters**.

- To make predictions about the future, we use the prediction rule

$$H^*(x) = w_0^* + w_1^* x$$

Example



$$\bar{x} = \frac{3+4+8}{3} = 5$$

$$\bar{y} = 4$$

$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-11}{14}$$

$$w_0^* = \bar{y} - w_1\bar{x} = 4 - \left(\frac{-11}{14}\right)5 = 7.9\ldots$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
3	7	−2	3	−6	4
4	3	−1	−1	1	1
8	2	3	−2	−6	9
				−11	14

Summary

- We introduced prediction rule framework to incorporate features in our predictions.

- We introduced the linear prediction rule, $H(x) = w_0 + w_1 x$.

- To determine the best choice of slope (w_1) and intercept (w_0), we chose the squared loss function $(y_i - H(x_i))^2$ and minimized empirical risk $R_{sq}(w_0, w_1)$:

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

- After solving for w_0^* and w_1^* through partial differentiation, we have a prediction rule $H^*(x) = w_0^* + w_1^* x$ that we can use to make predictions about the future.