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.

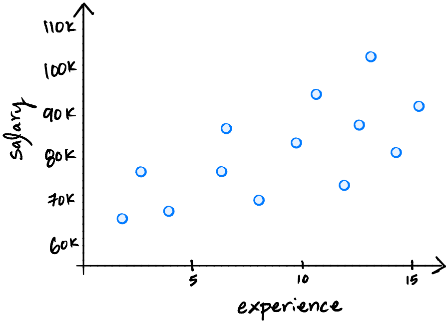
Finding the best prediction rule

- **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

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

- 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: $H(x) = w_0$.

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

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

- 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, $R_{\text{sq}}(w_0, w_1)$:

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

- Strategy: multivariable calculus.

Recall: the gradient

- 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**:

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

- **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.

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

Discussion Question

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

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

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

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

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

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

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

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

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

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

Strategy

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

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^*

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

Solve for w_1^*

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

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

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

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.

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

Proof:

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:

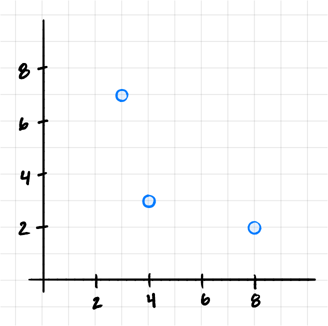
$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad w_0^* = \bar{y} - w_1\bar{x}$$

  - 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



$\bar{x} =$

$\bar{y} =$

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

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

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				
4	3				
8	2				

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.