Lecture 12

Multiple Linear Regression

DSC 40A, Fall 2025

Recap: Regression and linear algebra

Regression and linear algebra (Solution 1)

- Define the **design matrix** $X \in \mathbb{R}^{n\times 2}$, **observation vector** $\vec{y} \in \mathbb{R}^n$, and parameter vector $\vec{w} \in \mathbb{R}^2$ as:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \begin{matrix} \to \text{intercept} \\ \to \text{slope} \end{matrix}$$

- How do we make the hypothesis vector, $\vec{h} = X\vec{w}$, as close to $\vec{y}$ as possible? Use the parameter vector $\vec{w}^*$:

$$\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$$

- Solution: We chose $\vec{w}^*$ so that $\vec{h}^* = X\vec{w}^*$ is **the projection of $\vec{y}$ onto the span of the columns of the design matrix,** X and minimized the length of the projection error $\|\vec{e}\| = \|\vec{y} - X\vec{w}\|$.

Regression and linear algebra (Solution 2)

- Define the **design matrix** $X \in \mathbb{R}^{n \times 2}$, **observation vector** $\vec{y} \in \mathbb{R}^n$, and parameter vector $\vec{w} \in \mathbb{R}^2$ as:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

- How do we minimize the mean squared error $R_{\text{sq}}(\vec{w}) = \frac{1}{n}||\vec{y} - X\vec{w}||^2$? Using calculus the optimal paramter vector $\vec{w}^*$ is:

$\nabla R_{sq}(\vec{w}) = 0$

$$\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$$

- Solution: we computed the gradient of $R_{\text{sq}}(\vec{w})$, set it to zero and solved for $\vec{w}$.

Multiple linear regression

	$x^{(1)}$ departure_hour	$x^{(2)}$ day_of_month	y minutes
0	10.816667	15	68.0
1	7.750000	16	94.0
2	8.450000	22	63.0
3	7.133333	23	100.0
4	9.150000	30	69.0
...	...	...	...

commute

So far, we've fit **simple** linear regression models, which use only **one** feature (`'departure_hour'`) for making predictions.

The setup

- Suppose we have the following dataset.

commute time

	departure_hour	day_of_month	minutes
row	$x^{(1)}$	$x^{(2)}$	y
1	8.45	22	63.0
2	8.90	28	89.0
3	8.72	18	89.0

- We can represent each day with a **feature vector**, $\vec{x}$:

$$x_1 = \begin{bmatrix} 8.45 \\ 22 \end{bmatrix} \qquad y_1 = 63 \qquad H(\vec{x}_i) = y_i$$

$$x_2 = \begin{bmatrix} 8.90 \\ 28 \end{bmatrix} \qquad y_2 = 89$$

Finding the optimal parameters

- To find the optimal parameter vector, $\vec{w}^*$, we can use the **design matrix** $X \in \mathbb{R}^{n \times 3}$ and **observation vector** $\vec{y} \in \mathbb{R}^n$:

$$X = \begin{bmatrix} 1 & \text{departure hour}_1 & \text{day}_1 \\ 1 & \text{departure hour}_2 & \text{day}_2 \\ \dots & \dots & \dots \\ 1 & \text{departure hour}_n & \text{day}_n \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} \text{commute time}_1 \\ \text{commute time}_2 \\ \vdots \\ \text{commute time}_n \end{bmatrix}$$

- Then, all we need to do is solve the **normal equations**:

$$X^T X \vec{w}^* = X^T \vec{y}$$

$$\vec{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

If $X^T X$ is invertible, we know the solution is:

$$\vec{w}^* = \underbrace{(X^T X)}_{3 \times 3}{}^{-1} X^T \vec{y}$$

Notation for multiple linear regression

X_1 – 1st data point $\in \mathbb{R}^2$

$X^{(1)}$ – 1st feature $\in \mathbb{R}^n$

- We will need to keep track of multiple features for every individual in our dataset.
  - In practice, we could have hundreds or thousands of features!
- As before, subscripts distinguish between individuals in our dataset. We have n individuals, also called **training examples**.
- Superscripts distinguish between **features**. We have d features. $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$

$$\text{departure hour: } x^{(1)} \in \mathbb{R}^n$$

$$\text{day of month: } x^{(2)} \in \mathbb{R}^n$$

Think of $x^{(1)}$, $x^{(2)}$, ... as new variable names, like new letters.

not exponent!

$x^2 = x \cdot x$

$X^{(7)} \to$ 7th column
7th feature

$X_{21} \to$ 21st data point

Augmented feature vectors

- The **augmented feature vector** $\text{Aug}(\vec{x})$ is the vector obtained by adding a 1 to the front of feature vector $\vec{x}$:

$$\vec{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(d)} \end{bmatrix} \qquad \text{Aug}(\vec{x}) = \begin{bmatrix} 1 \\ x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(d)} \end{bmatrix} \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

departure time

day of the month

temperture

- Then, our hypothesis function is:

$$H(\vec{x}) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_d x^{(d)}$$
$$= \vec{w} \cdot \text{Aug}(\vec{x})$$

$\in \mathbb{R}$

$(d+1)\times 1$ $(d+1)\times 1$

The general problem

- We have n data points, $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)$, where each $\vec{x}_i$ is a feature vector of d features:

$$\vec{x}_i = \begin{bmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(d)} \end{bmatrix} \in \mathbb{R}^d$$

- We want to find a good linear hypothesis function:

$$\begin{aligned} H(\vec{x}) &= w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_d x^{(d)} \\ &= \vec{w} \cdot \text{Aug}(\vec{x}) \end{aligned}$$

How to find $\vec{w}^* = (w_0, w_1, \ldots, w_d)$

SLR model

$(x_1, y_1), \ldots (x_n, y_n)$

$x_i \in \mathbb{R}$

scalar

The general solution

- Define the **design matrix** $X \in \mathbb{R}^{n \times (d+1)}$ and **observation vector** $\vec{y} \in \mathbb{R}^n$:

data point 1

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(d)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(d)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(d)} \end{bmatrix} = \begin{bmatrix} \text{Aug}(\vec{x_1})^T \\ \text{Aug}(\vec{x_2})^T \\ \vdots \\ \text{Aug}(\vec{x_n})^T \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$n \times (d+1)$

- Then, solve the **normal equations** to find the optimal parameter vector, $\vec{w}^*$:

feature 1

$$X^T X \vec{w}^* = X^T \vec{y}$$

If $X^T X$ invertible

optimal $\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$

Terminology for parameters

- With d features, $\vec{w}$ has $d + 1$ entries.
- w_0 is the **bias**, also known as the **intercept**.
- $w_1, w_2, \ldots, w_d$ each give the **weight**, or **coefficient**, or **slope**, of a feature.

$$H(\vec{x}) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_d x^{(d)}$$

Interpreting parameters

Example: Predicting sales

- For each of 27 stores, we have:
  - net sales, $\vec{y}$
  - square feet,
  - inventory,
  - advertising expenditure,
  - district size, and
  - number of competing stores.

X 27 × 6

- **Goal**: Predict net sales given the other five features.
- To begin, we'll start trying to fit the hypothesis function to predict sales:

$$H(\text{square feet}, \text{competitors}) = w_0 + w_1 \cdot \underline{\text{square feet}} + w_2 \cdot \underline{\text{competitors}}$$

d=2

Question 🤔

Answer at q.dsc40a.com

$$H(\text{square feet}, \text{competitors}) = w_0 + w_1 \cdot \text{square feet} + w_2 \cdot \text{competitors}$$

bigger stores sell more

more competitors sell less

What will be the signs of w_1^* and w_2^*?

- A. w_1^* $+$ w_2^* $+$
- B. w_1^* $+$ w_2^* $-$
- A. w_1^* $-$ w_2^* $+$
- A. w_1^* $-$ w_2^* $-$

Let's find out! Follow along in this notebook.