Lecture 11 – Regression and Linear Algebra

DSC 40A, Fall 2022 @ UC San Diego

Mahdi Soleymani, with help from **many others**

Agenda

- Formulate mean squared error in terms of linear algebra.
- Minimize mean squared error using linear algebra.

Wait... why do we need linear algebra?

- Soon, we'll want to make predictions using more than one feature (e.g. predicting salary using years of experience and GPA).
  - If the intermediate steps get confusing, think back to this overarching goal.
- Thinking about linear regression in terms of **linear algebra** will allow us to find prediction rules that
  - use multiple features.
  - are non-linear.
- **Let's start by expressing R_{sq} in terms of matrices and vectors.**

Regression and linear algebra

- We chose the parameters for our prediction rule

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

by finding the w_0^* and w_1^* that minimized mean squared error:

$$R_{\text{sq}}(H) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overbrace{H(x_i)}^{w_0 + w_1 x})^2.$$

- This is kind of like the formula for the length of a vector!

$\vec{v} = (v_1, \dots, v_n)$

$\|\vec{v}\|^2 = \vec{v} \cdot \vec{v} = v_1^2 + v_2^2 + \cdots + v_n^2$

Regression and linear algebra

$y_1, y_2, \ldots, y_n$

$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

Let's define a few new terms:

- The **observation vector** is the vector $\vec{y} \in \mathbb{R}^n$ with components y_i. This is the vector of observed/"actual" values.

$H(x_i)$

$\vec{h} = \begin{bmatrix} H(x_1) \\ H(x_2) \\ \vdots \\ H(x_n) \end{bmatrix}$

- The **hypothesis vector** is the vector $\vec{h} \in \mathbb{R}^n$ with components $H(x_i)$. This is the vector of predicted values.
- The **error vector** is the vector $\vec{e} \in \mathbb{R}^n$ with components $e_i = y_i - H(x_i)$. This is the vector of (signed) errors.

$$\vec{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - H(x_1) \\ \vdots \\ y_n - H(x_n) \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} H(x_1) \\ \vdots \\ H(x_n) \end{bmatrix}$$

$$= \vec{y} - \vec{h}$$

Regression and linear algebra

$||\vec{v}||^2 = \sum_{i=1}^{n} v_i^2$

Let's define a few new terms:

- The **observation vector** is the vector $\vec{y} \in \mathbb{R}^n$ with components y_i. This is the vector of observed/"actual" values.

- The **hypothesis vector** is the vector $\vec{h} \in \mathbb{R}^n$ with components $H(x_i)$. This is the vector of predicted values.

- The **error vector** is the vector $\vec{e} \in \mathbb{R}^n$ with components $e_i = y_i - H(x_i)$. This is the vector of (signed) errors.

$$\vec{e} = \begin{bmatrix} y_1 - H(x_1) \\ \vdots \\ y_n - H(x_n) \end{bmatrix}$$

- We can rewrite the mean squared error as:

Def of length of vector

$$R_{sq}(H) = \frac{1}{n}\sum_{i=1}^{n}(y_i - H(x_i))^2 = \frac{1}{n}||\vec{e}||^2 = \frac{1}{n}||\vec{y} - \vec{h}||^2.$$

$\vec{e} = \vec{y} - \vec{h}$

The hypothesis vector

- The **hypothesis vector** is the vector $\vec{h} \in \mathbb{R}^n$ with components $H(x_i)$. This is the vector of predicted values.

$H(x) = w_0 + w_1 x$

- The hypothesis vector $\vec{h}$ can be written

$$\vec{h} = \begin{bmatrix} H(x_1) \\ H(x_2) \\ \vdots \\ H(x_n) \end{bmatrix} = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ \vdots \\ w_0 + w_1 x_n \end{bmatrix} = \underbrace{\overbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}^{X}}_{\text{Design matrix}} \underbrace{\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}}_{\text{parameter vector}}$$

$$\begin{bmatrix} b_1 & b_2 \\ b_3 & b_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$\vec{h} = X \vec{w}$$

Rewriting the mean squared error

- Define the **design matrix** X to be the $n \times 2$ matrix

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

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

- Define the **parameter vector** $\vec{w} \in \mathbb{R}^2$ to be $\vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$.

- Then $\vec{h} = X\vec{w}$, so the mean squared error becomes:

$$\vec{h} = X\vec{w}$$

$$R_{\text{sq}}(H) = \frac{1}{n}||\vec{y} - \vec{h}||^2$$

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n}||\vec{y} - X\vec{w}||^2$$

Mean squared error, reformulated

- Before, our goal was to find the values of w_0 and w_1 that minimize

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

  - The results:

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

- **Now**, our goal is to find the vector $\vec{w}$ that minimizes

$$R_{sq}(\vec{w}) = \frac{1}{n}||\vec{y} - X\vec{w}||^2$$

$\hookrightarrow \vec{y}$

find $\vec{w}^*$ that minimizes $R_{sq}(\vec{w})$

- **Both versions of R_{sq} are equivalent.**

Spoiler alert...

- Goal: find the vector $\vec{w}$ that minimizes

$$R_{sq}(\vec{w}) = \frac{1}{n}||\vec{y} - X\vec{w}||^2$$

- Spoiler alert: the answer[1] is

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

- Let's look at this formula in action in a notebook.
- Then we'll prove it ourselves by hand.

[1] assuming $X^T X$ is invertible

Minimizing mean squared error, again

Some key linear algebra facts

If A and B are matrices, and $\vec{u}, \vec{v}, \vec{w}, \vec{z}$ are vectors:

- $(A + B)^T = A^T + B^T$
- $(AB)^T = B^T A^T$
- $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u} = \vec{u}^T \vec{v} = \vec{v}^T \vec{u}$
- $\|\vec{u}\|^2 = \vec{u} \cdot \vec{u}$
- $(\vec{u} + \vec{v}) \cdot (\vec{w} + \vec{z}) = \vec{u} \cdot \vec{w} + \vec{u} \cdot \vec{z} + \vec{v} \cdot \vec{w} + \vec{v} \cdot \vec{z}$

Goal

- We want to minimize the mean squared error:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$$

- Strategy: Calculus.

- **Problem:** This is a *function of a vector*. What does it even mean to take the derivative of $R_{\text{sq}}(\vec{w})$ with respect to a vector $\vec{w}$?

A function of a vector

- **Solution:** A function *of a vector* is really just a function *of multiple variables*, which are the components of the vector. In other words,

$$R_{\text{sq}}(\vec{w}) = R_{\text{sq}}(w_0, w_1, \ldots, w_d)$$

  where $w_0, w_1, \ldots, w_d$ are the entries of the vector $\vec{w}$.[2]

- We know how to deal with derivatives of multivariable functions: the gradient!

[2] In our case, $\vec{w}$ has just two components, w_0 and w_1. We'll be more general since we eventually want to use prediction rules with even more parameters.

The gradient with respect to a vector

- The **gradient of $R_{\text{sq}}(\vec{w})$ with respect to $\vec{w}$** is the vector of partial derivatives:

$$\nabla_{\vec{w}} R_{\text{sq}}(\vec{w}) = \frac{dR_{\text{sq}}}{d\vec{w}} = \begin{bmatrix} \frac{\partial R_{\text{sq}}}{\partial w_0} \\ \frac{\partial R_{\text{sq}}}{\partial w_1} \\ \vdots \\ \frac{\partial R_{\text{sq}}}{\partial w_d} \end{bmatrix}$$

where $w_0, w_1, \ldots, w_d$ are the entries of the vector $\vec{w}$.

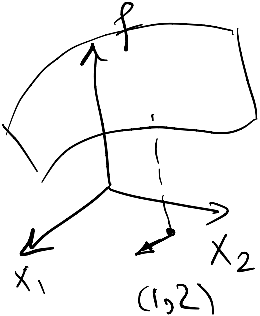
Example gradient calculation

Example: Suppose $f(\vec{x}) = \vec{a} \cdot \vec{x}$, where $\vec{a}$ and $\vec{x}$ are vectors in $\mathbb{R}^n$. What is $\frac{d}{d\vec{x}} f(\vec{x})$?

$$f(\vec{x}) = \vec{a} \cdot \vec{x} \rightarrow \text{scalar}$$

$$= a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

$$\frac{d f(\vec{x})}{d\vec{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \vec{a}$$



$$\frac{\partial f}{\partial x_1} = a_1$$

$$\frac{\partial f}{\partial x_2} = a_2 \qquad \vdots \qquad \frac{\partial f}{\partial x_n} = a_n$$

$$\frac{d(\vec{a} \cdot \vec{x})}{d\vec{x}} = \vec{a}$$

$$\frac{d(ax)}{dx} = a$$

Goal

- We want to minimize the mean squared error:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$$

- Strategy:
  1. Compute the gradient of $R_{\text{sq}}(\vec{w})$.
  2. Set it to zero and solve for $\vec{w}$.
     - The result is called $\vec{w}^*$.

- Let's start by rewriting the mean squared error in a way that will make it easier to compute its gradient.

Rewriting mean squared error

$\|\vec{v}\|^2 = \vec{v} \cdot \vec{v} = \vec{v}^T \vec{v}$

$$R_{sq}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2 = \frac{1}{n}(\vec{y} - X\vec{w}) \cdot (\vec{y} - X\vec{w})$$

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

Discussion Question

Which of the following is equivalent to $R_{sq}(\vec{w})$?

$\vec{v}$ $n \times 1$

$\vec{v}^T$ $1 \times n$ $\vec{v}$ $n \times 1$

a) $\frac{1}{n}(\vec{y} - X\vec{w}) \cdot (X\vec{w} - y)$

b) $\frac{1}{n}\sqrt{(\vec{y} - X\vec{w}) \cdot (y - X\vec{w})}$

c) $\frac{1}{n}(\vec{y} - X\vec{w})^T(y - X\vec{w})$

d) $\frac{1}{n}(\vec{y} - X\vec{w})(y - X\vec{w})^T$

To answer, go to `menti.com` **and enter 8482 5148.**

Rewriting mean squared error

$$R_{sq}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2 = \frac{1}{n}\left(\vec{y} - X\vec{w}\right)^T\left(\vec{y} - X\vec{w}\right)$$

$$= \frac{1}{n}\left(\vec{y}^T - (X\vec{w})^T\right)\left(\vec{y} - X\vec{w}\right)$$

$$= \frac{1}{n}\left(\vec{y}^T\vec{y} - \vec{y}^T X\vec{w} - (X\vec{w})^T\vec{y} + (X\vec{w})^T X\vec{w}\right)$$

same!

$$\vec{v}^T\vec{u} = \vec{v}\cdot\vec{u}$$

$$(X\vec{w})^T\vec{y} = \vec{w}^T X^T \vec{y} = \vec{w}^T(X^T\vec{y}) = \vec{w}\cdot(X^T\vec{y})$$

$$\underbrace{\vec{y}^T}\, X\vec{w} = (X^T\vec{y})\cdot\vec{w}$$

$$(X^T\vec{y})^T = \vec{y}^T(X^T)^T = \vec{y}\cdot\vec{X}$$

Rewriting mean squared error

$R_{\text{sq}}(\vec{w}) =$

Compute the gradient

$$\frac{dR_{\text{sq}}}{d\vec{w}} = \frac{d}{d\vec{w}}\left(\frac{1}{n}\left[\vec{y}\cdot\vec{y} - 2X^T\vec{y}\cdot\vec{w} + \vec{w}^T X^T X\vec{w}\right]\right)$$

$$= \frac{1}{n}\left[\underbrace{\frac{d}{d\vec{w}}(\vec{y}\cdot\vec{y})} - \underbrace{\frac{d}{d\vec{w}}\left(2X^T\vec{y}\cdot\vec{w}\right)} + \underbrace{\frac{d}{d\vec{w}}\left(\vec{w}^T X^T X\vec{w}\right)}\right]$$

Compute the gradient

$$\frac{dR_{\text{sq}}}{d\vec{w}} = \frac{d}{d\vec{w}}\left(\frac{1}{n}\left[\vec{y}\cdot\vec{y} - 2X^T\vec{y}\cdot\vec{w} + \vec{w}^T X^T X\vec{w}\right]\right)$$
$$= \frac{1}{n}\left[\frac{d}{d\vec{w}}(\vec{y}\cdot\vec{y}) - \frac{d}{d\vec{w}}\left(2X^T\vec{y}\cdot\vec{w}\right) + \frac{d}{d\vec{w}}\left(\vec{w}^T X^T X\vec{w}\right)\right]$$

- $\frac{d}{d\vec{w}}(\vec{y}\cdot\vec{y}) = 0.$
  - Why? $\vec{y}$ is a constant with respect to $\vec{w}$.
- $\frac{d}{d\vec{w}}\left(\vec{2}X^T\vec{y}\cdot\vec{w}\right) = 2X^T y.$
  - Why? We already showed $\frac{d}{d\vec{x}}\vec{a}\cdot\vec{x} = \vec{a}$.
- $\frac{d}{d\vec{w}}\left(\vec{w}^T X^T X\vec{w}\right) = 2X^T X\vec{w}.$
  - Why? Will see in HW4.

Compute the gradient

$$
\begin{aligned}
\frac{dR_{\text{sq}}}{d\vec{w}} &= \frac{d}{d\vec{w}}\left(\frac{1}{n}\left[\vec{y}\cdot\vec{y} - 2X^T\vec{y}\cdot\vec{w} + \vec{w}^TX^TX\vec{w}\right]\right) \\
&= \frac{1}{n}\left[\frac{d}{d\vec{w}}(\vec{y}\cdot\vec{y}) - \frac{d}{d\vec{w}}\left(2X^T\vec{y}\cdot\vec{w}\right) + \frac{d}{d\vec{w}}\left(\vec{w}^TX^TX\vec{w}\right)\right]
\end{aligned}
$$

$$= \frac{1}{n}\left[0 - 2X^T\vec{y} + 2X^TX\vec{w}\right] = 0$$

$$\Rightarrow 2X^T\vec{y} + 2X^TX\vec{w} = 0$$

$$\Rightarrow X^TX\vec{w} = X^T\vec{y}$$

The normal equations

- To minimize $R_{\text{sq}}(\vec{w})$, set its gradient to zero and solve for $\vec{w}$:

$$-2X^T\vec{y} + 2X^TX\vec{w} = 0$$

$$\implies X^TX\vec{w} = X^T\vec{y}$$

$A x = \vec{b}$

A^{-1}

- This is a system of equations in matrix form, called the **normal equations**.

- If X^TX is invertible, the solution is

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

- This is equivalent to the formulas for w_0^* and w_1^* we saw before!
  - Benefit – this can be easily extended to more complex prediction rules.

Side note — another proof

- We set out to minimize

$$R_{sq}(\vec{w}) = \frac{1}{n}||\vec{y} - X\vec{w}||^2$$

- We did it using multivariable calculus.
- There's another proof of this same fact that relies on knowledge of linear projections. We will not cover it in class and you are not responsible for it, but you can watch video 13.4 here if you're curious: `http://ds100.org/su20/lecture/lec13/`.

Summary

Summary

- We used linear algebra to rewrite the mean squared error for the prediction rule $H(x) = w_0 + w_1 x$ as

$$R_{sq}(\vec{w}) = \frac{1}{n}||\vec{y} - X\vec{w}||^2$$

  - X is called the **design matrix**, $\vec{w}$ is called the **parameter vector**, $\vec{y}$ is called the **observation vector**, and $\vec{h} = X\vec{w}$ is called the **hypothesis vector**.

- We minimized $R_{sq}(\vec{w})$ using multivariable calculus and found that the minimizing $\vec{w}$ satisfies the **normal equations**, $X^TX\vec{w} = X^Ty$.
  - Closed-form solution:

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

What's next?

- The whole point of reformulating linear regression in terms of linear algebra was so that we could generalize our work to more sophisticated prediction rules.
  - Note that when deriving the normal equations, we didn't assume that there was just one feature.
- Examples of the types of prediction rules we'll be able to fit soon:
  - $H(x) = w_0 + w_1 x + w_2 x^2$.
  - $H(x) = w_0 + w_1 \cos(x) + w_2 e^x$.
  - $H(x^{(1)}, x^{(2)}) = w_0 + w_1 x^{(1)} + w_2 x^{(2)}$.
    - e.g. Predicted Salary = $w_0 + w_1$(Years of Experience) + w_2(GPA).