# Lecture 8 – Regression and Linear Algebra

**DSC 40A, Fall 2021 @ UC San Diego**

Suraj Rampure, with help from **many others**

# Announcements

- Midterm Exam on Thursday from 11:00AM-12:30PM.
    - Remote on Gradescope.
    - Absolutely no collaboration.
    - Make sure to read https://campuswire.com/c/GF82D3B2E/feed/248 for all the details.
- Review session **today** from 5-8PM in Pepper Canyon 109.
    - Brief conceptual overview, as well as review of Homeworks 1-3.
    - Will be podcasted.
- **Please fill out Survey 3!**
    - Will close tomorrow night.
- The OH schedule is now updated; we have many more OH on Tuesday and Wednesday, and no OH on Thursday or Friday. **Discussion is replaced with office hours.**

# Midterm preparation

- Review the solutions to previous homeworks and groupworks.
  - Homework 3 solutions are now up.
- Identify which concepts are still iffy. Re-watch lecture, post on Campuswire, come to office hours.
- Look at the past exams at `https://dsc40a.com/resources`.
  - Walkthrough exists for SP20.
- Study in groups.
- Make a “cheat sheet”.

# Agenda

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

# Linear algebra review

# Dot products

- The **dot product** of two vectors $\vec{u}$ and $\vec{v}$ in $\mathbb{R}^n$ is denoted by:

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

- Definition:

$$\vec{u} \cdot \vec{v} = \sum_{i=1}^{n} u_i v_i = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$$

- The result is a **scalar**!

$$\vec{u} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad \vec{v} = \begin{bmatrix} 7 \\ -1 \\ 0 \end{bmatrix} \qquad \vec{u} \cdot \vec{v} = 1 \cdot 7 + 2(-1) + 3(0)$$

$$= 7 - 2 = 5$$

# Properties of the dot product

- Commutative:

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

- Distributive:

$$\vec{u} \cdot (\vec{v} + \vec{w}) = \vec{u} \cdot \vec{v} + \vec{u} \cdot \vec{w}$$

# Matrix-vector multiplication

- Special case of matrix-matrix multiplication.

- The result is always a vector with the same number of rows as the matrix.

- One view: a "mixture" of the columns.

$$\begin{array}{c}\rightarrow\\\rightarrow\end{array}\underbrace{\begin{bmatrix}1 & 2 & 1\\3 & 4 & 5\end{bmatrix}}_{2\times3}\underbrace{\begin{bmatrix}a_1\\a_2\\a_3\end{bmatrix}}_{3\times1} = a_1\begin{bmatrix}1\\3\end{bmatrix} + a_2\begin{bmatrix}2\\4\end{bmatrix} + a_3\begin{bmatrix}1\\5\end{bmatrix} = \begin{bmatrix}\ \end{bmatrix}_{2\times1}$$

- Another view: a dot product with the rows.

$$\begin{bmatrix}1 & 2 & 1\\3 & 4 & 5\end{bmatrix}\begin{bmatrix}a_1\\a_2\\a_3\end{bmatrix} = \begin{bmatrix}1a_1+2a_2+1a_3\\3a_1+4a_2+5a_3\end{bmatrix}$$

## Discussion Question

If $A$ is an $m \times n$ matrix and $\vec{v}$ is a vector in $\mathbb{R}^n$, what are the dimensions of the product $\vec{v}^T A^T A \vec{v}$?

a) $m \times n$ (matrix)

b) $n \times 1$ (vector)

[✓ c) $1 \times 1$ (scalar)]

d) The product is undefined.

**To answer, go to `menti.com` and enter 22 77 26 8.**

$$\underbrace{v^T}_{1 \times n}\,\underbrace{\underbrace{A^T}_{n \times m}\,\underbrace{\underbrace{A}_{m \times n}\,\underbrace{v}_{n \times 1}}_{m \times 1}}_{n \times 1} \rightarrow \text{result is } 1 \text{ by } 1$$

(scalar)!

# Matrices and functions

- Suppose $A$ is an $m \times n$ matrix and $\vec{x}$ is a vector in $\mathbb{R}^n$.

- Then, the function $f(\vec{x}) = Ax$ is a linear function that maps elements in $\mathbb{R}^n$ to elements in $\mathbb{R}^m$.
  - The input to $f$ is a vector, and so is the output.

- **Key idea:** matrix-vector multiplication can be thought of as applying a linear function to a vector.

$$A_{m \times n} \quad \vec{x}_{n \times 1} \quad \longrightarrow \quad \text{result}_{m \times 1}$$

# Mean squared error, revisited

# 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:

$$R_{\text{sq}}(H) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \underset{\uparrow\; w_0+w_1x_i}{H(x_i)})^2.$$

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

$$\|\vec{v}\|^2 = \vec{v}\cdot\vec{v} = v_1 \cdot v_1 + v_2 \cdot v_2 + v_3 \cdot v_3 + \dots + v_n \cdot v_n$$

$$= v_1^2 + v_2^2 + \dots + v_n^2$$

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

# Regression and linear algebra

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.

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

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

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

- 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.

$$\vec{e} = \begin{bmatrix} y_1 - H(x_1) \\ y_2 - H(x_2) \\ \vdots \\ y_n - H(x_n) \end{bmatrix}$$

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

# Regression and linear algebra

$$\vec{e} = \begin{bmatrix} y_1 - H(x_1) \\ y_2 - H(x_2) \\ \vdots \\ y_n - H(x_n) \end{bmatrix}$$

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.

- 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.

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

- We can rewrite the mean squared error as:

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

# 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.

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

$$\vec{h} = \begin{bmatrix} H(x_1) \\ H(x_2) \\ \vdots \\ H(x_n) \end{bmatrix} = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ \vdots \\ w_0 + w_1 x_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}}_{\vec{w}}$$

design matrix

parameter vector

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

# Rewriting the mean squared error

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

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

equivalent

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

- 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:

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

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

# Mean squared error, reformulated

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

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

  - 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

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

- **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:

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

function of a vector

- 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,

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

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

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

$$R_{sq}(\vec{w}) \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$R_{sq}(w_0, w_1)$$

[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:

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

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 $\frac{d}{d\vec{x}} f(\vec{x})$?

$$f(\vec{x}) = \vec{a} \cdot \vec{x} = a_1 x_1 + a_2 x_2 + a_3 x_3 + \dots + a_n x_n$$

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

$$\frac{\partial f}{\partial x_2} = a_2$$

$$\vdots$$

$$\frac{\partial f}{\partial x_n} = a_n$$

$$\frac{d}{d\vec{x}} f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \vec{a}$$

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

# 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

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

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

## Discussion Question

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

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 22 77 26 8.**

# Rewriting mean squared error

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

$$(AB)^T = B^T A^T$$

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

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

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

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

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

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

$$\underbrace{\vec{w}^T X^T} X\vec{w}$$

(fixed after lecture)

# Rewriting mean squared error

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

# Compute the gradient

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

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

side note

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

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

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

$$= \vec{X}^Ty\cdot\vec{w}$$

$$(AB)^T = B^TA^T$$

# Compute the gradient

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

- $\frac{d}{d\vec{w}}\left(\vec{y}\cdot\vec{y}\right) = 0.$
  - Why? $\vec{y}$ is a constant with respect to $\vec{w}$.

- $\frac{d}{d\vec{w}}\left(\vec{2}X^T\vec{y}\cdot\vec{w}\right) = 2X^Ty.$
  - Why? We already showed $\frac{d}{d\vec{x}}\vec{a}\cdot\vec{x} = \vec{a}$.

- $\frac{d}{d\vec{w}}\left(\vec{w}^TX^TX\vec{w}\right) = 2X^TX\vec{w}.$
  - Why? See Homework 4.

# Compute the gradient

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

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

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

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

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

# The normal equations

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

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

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

$$A\vec{w} = b$$

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

- If $X^TX$ is invertible, the solution is

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

- 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**, $X^T X\vec{w} = X^T y$.
  - Closed-form solution:

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

# 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).