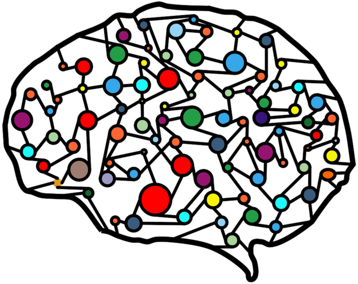
Lecture 9 – Regression and Linear Algebra



DSC 40A, Fall 2022 @ UC San Diego

Dr. Truong Son Hy, with help from **many others**

Announcements

- Look at the readings linked on the course website!
- Groupwork Relsease Day: Thursday afternoon
  Groupwork Submission Day: Monday midnight
  Homework Release Day: Friday after lecture
  Homework Submission Day: Friday before lecture
- See `dsc40a.com/calendar` for the Office Hours schedule.

Midterm study strategy

- Review the solutions to previous homeworks and groupworks.
- Re-watch lecture, post on Campuswire, come to office hours.
- Look at the past exams at `https://dsc40a.com/resources`.
- Study in groups.
- **Remember:** it's just an exam.

Agenda

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

Linear algebra review

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).
- Thinking about linear regression in terms of **linear algebra** will allow us to find prediction rules that use multiple features.
- Before we dive in, let's review.
- **There can be linear algebra on the midterm!!**

Matrices

- An $m \times n$ **matrix** is a table of numbers with m rows and n columns.
- We use upper-case letters for matrices.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

- A^T denotes the transpose of A:

$$A^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Matrix addition and scalar multiplication

- We can add two matrices only if they are the same size.
- Addition occurs elementwise:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 7 & 8 & 9 \\ -1 & -2 & -3 \end{bmatrix} = \begin{bmatrix} 8 & 10 & 12 \\ 3 & 3 & 3 \end{bmatrix}$$

- Scalar multiplication occurs elementwise, too:

$$2 \cdot \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix}$$

Matrix-matrix multiplication

- We can multiply two matrices A and B only if

  $$\#\text{ columns in } A = \#\text{ rows in } B.$$

- If A is $m \times n$ and B is $n \times p$, the result is $m \times p$.
  - This is **very useful**.

- The ij entry of the product is:

  $$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

Some matrix properties

- Multiplication is Distributive:

$$A(B + C) = AB + AC$$

- Multiplication is Associative:

$$(AB)C = A(BC)$$

- Multiplication is **not commutative**:

$$AB \neq BA$$

- Transpose of sum:

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

- Transpose of product:

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

Vectors

- An **vector** in $\mathbb{R}^n$ is an $n \times 1$ matrix.

- We use lower-case letters for vectors.

$$\vec{v} = \begin{bmatrix} 2 \\ 1 \\ 5 \\ -3 \end{bmatrix}$$

- Vector addition and scalar multiplication occur elementwise.

Geometric meaning of vectors

- A vector $\vec{v} = (v_1, \dots, v_n)^T$ is an arrow to the point $(v_1, \dots, v_n)$ from the origin.

- The **length**, or **norm**, of $\vec{v}$ is $\|\vec{v}\| = \sqrt{v_1^2 + v_2^2 + \dots + v_n^2}$.

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**!

- Sometimes, we can use the notation $\langle , \rangle$ for the dot product:
$$\vec{u} \cdot \vec{v} = \langle \vec{u}, \vec{v} \rangle$$

Discussion Question

Which of these is another expression for the length of $\vec{u}$?

a) $\vec{u} \cdot \vec{u}$
b) $\sqrt{\vec{u}^2}$
c) $\sqrt{\vec{u} \cdot \vec{u}}$
d) $\vec{u}^2$

Discussion Question

Which of these is another expression for the length of $\vec{u}$?

a) $\vec{u} \cdot \vec{u}$

b) $\sqrt{\vec{u}^2}$

c) $\sqrt{\vec{u} \cdot \vec{u}}$

d) $\vec{u}^2$

Answer: C

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.
- Result is always a vector with same number of rows as the matrix.
- One view: a "mixture" of the columns.

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

- Another view: a dot product with the rows.

Matrix-matrix multiplication & Dot product

Given two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, the matrix multiplication of A and B is a matrix $AB \in \mathbb{R}^{m \times p}$ with each element at row i-th and column j-th defined as:

$$(AB)_{ij} = \sum_{k} A_{ik} B_{kj}$$

How can it be related to dot product?

Matrix-matrix multiplication & Dot product

Given two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, the matrix multiplication of A and B is a matrix $AB \in \mathbb{R}^{m \times p}$ with each element at row i-th and column j-th defined as:

$$(AB)_{ij} = \sum_k A_{ik} B_{kj}$$

How can it be related to dot product?
First, rewrite it by using the transpose of B:

$$(AB)_{ij} = \sum_k A_{ik} B^T_{jk}$$

Matrix-matrix multiplication & Dot product

Let denote $X_{i,:}$ and $X_{:,j}$ as the i-th row and j-th column of a matrix X, respectively. We have:

$$(AB)_{ij} = \langle A_{i,:}, B_{j,:}^T \rangle = \langle A_{i,:}, B_{:,j} \rangle$$

Therefore, $(AB)_{ij}$ is the dot product of the i-th row of A and the j-th column of B.

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) 1×1 (scalar)
d) The product is undefined.

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) 1×1 (scalar)
d) The product is undefined.

Answer: C

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.

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:

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

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

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.

- 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.

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.

- 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.

- We can rewrite the mean squared error as:

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

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

$$\vec{h} = \begin{bmatrix} H(x_1) \\ H(x_2) \\ \vdots \\ H(x_n) \end{bmatrix} = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ \vdots \\ w_0 + w_1 x_n \end{bmatrix} =$$

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

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

Rewriting the mean squared error

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

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

- 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:

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

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

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

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

- **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

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

- Then we'll prove it ourselves by hand.

[1] assuming X^TX 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,

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

  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!

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

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

Keep in mind that $\frac{d}{d\vec{x}} f(\vec{x})$ is a vector of length n in which the i-th element is $\left[\frac{d}{d\vec{x}} f(\vec{x})\right]_i = \frac{\partial f}{\partial x_i}$. We have:

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

Keep in mind that $\frac{d}{d\vec{x}} f(\vec{x})$ is a vector of length n in which the i-th element is $\left[\frac{d}{d\vec{x}} f(\vec{x})\right]_i = \frac{\partial f}{\partial x_i}$. We have:

$$\frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i}(\vec{a} \cdot \vec{x}) = \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n} a_j \cdot x_j \right) =$$

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 $\frac{d}{d\vec{x}} f(\vec{x})$?
Keep in mind that $\frac{d}{d\vec{x}} f(\vec{x})$ is a vector of length n in which the i-th element is $\left[\frac{d}{d\vec{x}} f(\vec{x})\right]_i = \frac{\partial f}{\partial x_i}$. We have:

$$\frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i}(\vec{a} \cdot \vec{x}) = \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n} a_j \cdot x_j\right) = \sum_{j=1}^{n} a_j \cdot \frac{\partial x_j}{\partial x_i}$$

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

Keep in mind that $\frac{d}{d\vec{x}} f(\vec{x})$ is a vector of length n in which the i-th element is $\left[\frac{d}{d\vec{x}} f(\vec{x})\right]_i = \frac{\partial f}{\partial x_i}$. We have:

$$\frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i}(\vec{a} \cdot \vec{x}) = \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n} a_j \cdot x_j\right) = \sum_{j=1}^{n} a_j \cdot \frac{\partial x_j}{\partial x_i}$$

If $i \neq j$ then $\frac{\partial x_j}{\partial x_i} = 0$, otherwise 1. Thus: $\frac{\partial f}{\partial x_i} = a_i$. Therefore:

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

Goal

- We want to minimize the mean squared error:

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

- Strategy:
  1. Compute the gradient of $R_{sq}(\vec{w})$.
  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

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

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

Rewriting mean squared error

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

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

Answer: C

Rewriting mean squared error

Because $\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle} = \sqrt{\vec{x}^T \vec{x}}$, we have:

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

Rewriting mean squared error

Because $\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle} = \sqrt{\vec{x}^T \vec{x}}$, we have:

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

We have:

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

Rewriting mean squared error

Because $\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle} = \sqrt{\vec{x}^T \vec{x}}$, we have:

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

We have:

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

Rewriting mean squared error

Because $\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle} = \sqrt{\vec{x}^T \vec{x}}$, we have:

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

We have:

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

Keep in mind that $\vec{w}^T X^T \vec{y} = \vec{y}^T X \vec{w}$ is a symmetric 1×1 matrix or scalar.

Rewriting mean squared error

Because $\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle} = \sqrt{\vec{x}^T \vec{x}}$, we have:

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

We have:

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

Keep in mind that $\vec{w}^T X^T \vec{y} = \vec{y}^T X\vec{w}$ is a symmetric 1×1 matrix or scalar. We can even write them down as $X^T \vec{y} \cdot \vec{w}$ where $\cdot$ is dot product. We finally have:

Rewriting mean squared error

Because $\|\vec{x}\| = \sqrt{\langle\vec{x},\vec{x}\rangle} = \sqrt{\vec{x}^T\vec{x}}$, we have:

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

We have:

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

Keep in mind that $\vec{w}^TX^T\vec{y} = \vec{y}^TX\vec{w}$ is a symmetric 1×1 matrix or scalar. We can even write them down as $X^T\vec{y} \cdot \vec{w}$ where $\cdot$ is dot product. We finally have:

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

Compute the gradient

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

Compute the gradient

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

- $\frac{d}{d\vec{w}}\left(\vec{y}\cdot\vec{y}\right) = 0.$
  - 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? Your homework!

Compute the gradient

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

Compute the gradient

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

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

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

The normal equations

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

$$
\begin{aligned}
-2X^T\vec{y} + 2X^TX\vec{w} &= 0 \\
\implies X^TX\vec{w} &= X^T\vec{y}
\end{aligned}
$$

- 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).