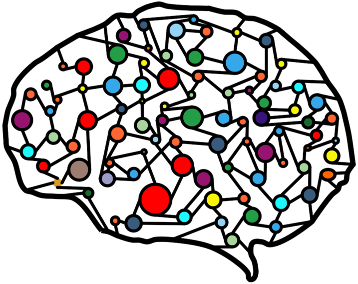
Lecture 10 – Linear Algebra and Regression



DSC 40A, Fall 2022 @ UC San Diego

Midterm study strategy

- Review the solutions to previous assignments.
- 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`.
- Study in groups.
- Make a “cheat sheet”.

Agenda

- Linear Algebra Review.
- Mean squared error, revisited

Linear algebra review

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).
- Thinking about linear regression in terms of **linear algebra** will allow us to find prediction rules that
    - use multiple features.
    - are non-linear.
- Before we dive in, let's review.

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

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) $\vec{u}^2$

To answer, go to `menti.com` and enter the code 4821 5997.

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.

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.

To answer, go to `menti.com` and enter the code 4821 5997.

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:

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

- 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

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

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.**

Summary

Summary, next time

- The correlation coefficient, r, measures the strength of the linear association between two variables x and y.

- We can re-write the optimal parameters for the linear prediction rule (under squared loss) as

$$w_1^* = r\frac{\sigma_y}{\sigma_x} \qquad w_0^* = \bar{y} - w_1^*\bar{x}$$

- We can then make predictions using $H^*(x) = w_0^* + w_1^* x$.

- We will need linear algebra in order to generalize regression to work with multiple features.

- **Next time**: Formulate linear regression in terms of linear algebra.

Summary

Summary

- We will need linear algebra in order to generalize regression to work with multiple features.
- 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**.