DSC 40A

Theoretical Foundations of Data Science I

Last Time

- We saw how to fit certain nonlinear functions to data by thinking of them as linear functions in new variables.

In This Video

We show how to think of linear regression in another way using linear algebra. This will eventually allow us to generalize our results to fit more nonlinear functions to data as well as make predictions based on multiple features.

Recommended Reading

Course Notes: Chapter 2, Section 2
Review: Linear Algebra Textbook

Quick Linear Algebra 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

product

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

match

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

- The ij entry of the product is:

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

$$\underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}}_{3\times 3} \underbrace{\begin{bmatrix} 1 & 0 \\ 2 & 3 \\ 1 & 1 \end{bmatrix}}_{3\times 2} = \underbrace{\begin{bmatrix} 8 & 9 \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix}}_{3\times 2}$$

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

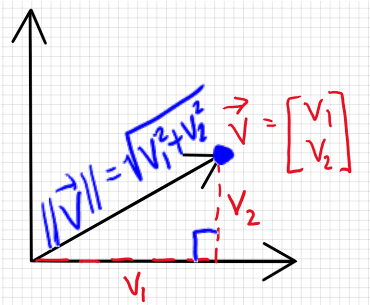
- An **vector** in $\mathbb{R}^n$ is an $n \times 1$ matrix.
- We use lower-case letters for vectors.

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

- Vector addition and scalar multiplication occur elementwise.

Geometric Meaning of Vectors

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



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

Dot Products

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

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

- Using low-level matrix multiplication definition:

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

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \underbrace{\begin{bmatrix} u_1 & u_2 & \ldots & u_n \end{bmatrix}}_{1 \times n} \underbrace{\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}}_{n \times 1}$$

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) ~~$\vec{u}^2$~~

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n \times 1} \quad [\;\;]_{n \times 1}$$

don't match

$$\|\vec{u}\| = \sqrt{u_1^2 + u_2^2 + \ldots + u_n^2}$$

$$\|\vec{u}\|^2 = u_1^2 + u_2^2 + \ldots + u_n^2 = \vec{u} \cdot \vec{u}$$

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

Properties of the Dot Product

- Commutative:

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

$= u_1 v_1 + u_2 v_2 + \dots + u_n v_n$

- 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.

$$\underbrace{\begin{bmatrix} 1 & 2 & 1 \\ 3 & 4 & 5 \end{bmatrix}}_{2\times 3} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}}_{3\times 1} = \begin{bmatrix} \; \end{bmatrix}$$

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 $\vec{v}^T A^T A \vec{v}$?

$$= (Av)^T Av$$
$$= (Av) \cdot Av$$
$$= \|Av\|^2$$

a) $m \times n$ (matrix)

b) $n \times 1$ (vector)

c) 1×1 (scalar)

d) The product is undefined.

$$\underbrace{\underset{1\times n}{v^T}\ \underset{n\times m}{A^T}\ \underset{m\times n}{A}\ \underset{n\times 1}{v}}_{\text{result}}$$

Matrices and Functions

- Matrix-vector multiplication takes in a vector, outputs a vector.

- An $m \times n$ matrix is an encoding of a linear function mapping $\mathbb{R}^n$ to $\mathbb{R}^m$.

- Matrix multiplication evaluates that function on a given vector.

$$\begin{bmatrix} \text{matrix} \end{bmatrix}_{m\times n} \begin{bmatrix} \text{in } \mathbb{R}^n \end{bmatrix}_{n\times 1} = \begin{bmatrix} \text{in } \mathbb{R}^m \end{bmatrix}_{m\times 1}$$

Back to Regression

Regression and Linear Algebra

- We chose the parameters for our prediction rule

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

by minimizing the 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!

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

$$R_{sq}(H) = \frac{1}{n} \left\| \begin{bmatrix} y_1 - H(x_1) \\ y_2 - H(x_2) \\ \vdots \\ y_n - H(x_n) \end{bmatrix} \right\|^2 = \frac{1}{n} \|\vec{e}\|^2 \text{ where } \vec{e} = \begin{bmatrix} y_1 - H(x_1) \\ y_2 - H(x_2) \\ \vdots \\ y_n - H(x_n) \end{bmatrix}$$

Regression and Linear Algebra

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

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

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

$$\begin{bmatrix} H(x_1) \\ \vdots \\ H(x_n) \end{bmatrix}$$

- 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.

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

Regression and Linear Algebra

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

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

actual predicted

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

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

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

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

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:

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

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

Summary

- The mean squared error is:

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

  where X is the **design matrix** containing the data, $\vec{w}$ is the **parameter vector**, and $\vec{y}$ is the **observation vector**.

- **Next time:** We minimize this function using calculus.

- Soon, we'll extend these results to more interesting cases:
  - more nonlinear functions,
  - multiple predictor variables.