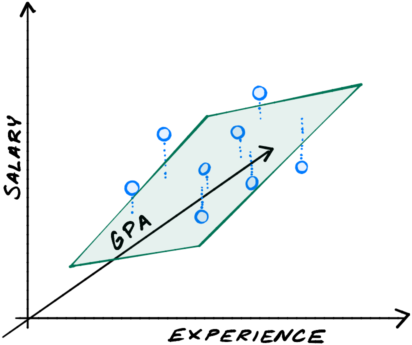


DSC 40A

Lecture 08

Least Squares Regression, pt. III

Announcements

- The midterm is Tuesday, in lecture.
- Covers Lectures 01 through 07 (this Tuesday).
- Concepts:
  - loss functions and ERM, gradient descent, convexity, least squares regression, etc.
- Core Skills:
  - partial derivatives, working with summations, chains of inequalities, etc.
- Best study device: homeworks and discussion worksheets.

- Cheat Sheet

Last Time

- **Goal**: Find prediction rule $H(x)$ for predicting salary given years of experience.

- To avoid **overfitting**, use linear prediction rule:

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

- We want w_1 and w_0 to minimize the mean squared error:

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

Last Time

- Take derivatives, set to zero, solve:

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

Today

- How do we predict salary given **multiple** features?
  - years of experience, number of internships, GPA, etc.

- We'll need to use some linear algebra...

Basic Linear Algebra Review

Matrices

An $m \times n$ **matrix** is a table of numbers with m rows, n columns:

- Example: 2×3 matrix:

$$\begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{pmatrix}$$

- Example: 3×3 “square” matrix:

$$\begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}$$

- Example: 3×1 “column”:

$$\begin{pmatrix} m_{11} \\ m_{21} \\ m_{31} \end{pmatrix}$$

Matrix Notation

- We use upper-case letters for matrices.

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

- Sometimes use subscripts to denote particular elements: $A_{13} = 3, A_{21} = 4$

- A^T denotes the transpose of A:

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

Matrix Addition and Scalar Multiplication

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

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

- Scalar multiplication occurs elementwise, too:

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

Matrix-Matrix Multiplication

- We can multiply two matrices A and B only if # cols in A is equal to # rows in B

$(m \times \cancel{n})(\cancel{n} \times p)$

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

- The low-level definition. the ij entry of the product is:

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

Matrix-Matrix Multiplication Example

$$A = \overset{2\times 3}{\begin{pmatrix} 1 & 2 & 1 \\ 3 & 4 & 5 \end{pmatrix}} \qquad B = \overset{3\times 2}{\begin{pmatrix} 3 & 6 \\ 1 & 3 \\ 4 & 8 \end{pmatrix}}$$

- What is the size of AB?

  2×2

- What is $(AB)_{12}$?

$$1 \cdot 6 + 2 \cdot 3 + 1 \cdot 8 = 6 + 6 + 8$$
$$= 20$$

Matrix-Matrix Multiplication Properties

- Distributive: $A(B + C) = AB + AC$
- Associative: $(AB)C = A(BC)$
- **Not commutative in general**: $AB \neq BA$

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

Identity Matrices

- The $n \times n$ **identity matrix** I has ones along the diagonal:

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- If A is $n \times m$, then $IA = A$.

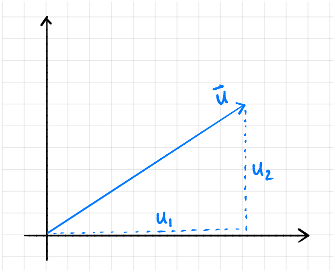
- If B is $m \times n$, then $BI = B$.

Vectors

- An d-**vector** is an $d \times 1$ matrix.
- Often use arrow, lower-case letters to denote: $\vec{x}$.
- Often write $\vec{x} \in \mathbb{R}^d$ to say $\vec{x}$ is a d vector.
- Example. A 4-vector:
$$\begin{pmatrix} 2 \\ 1 \\ 5 \\ -3 \end{pmatrix}$$
- Vector addition and scalar multiplication are also elementwise.

Geometric Meaning of Vectors

- A vector $\vec{u} = (u_1, \ldots, u_d)^T$ is an arrow to the point $(u_1, \ldots, u_d)$:



$\vec{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$

- The length, or **norm**, of $\vec{u}$ is $\|\vec{u}\| = \sqrt{u_1^2 + u_2^2 + \ldots + u_d^2}$.

- A **unit vector** is a vector of norm 1.

Dot Products

- The **dot product** of two d-vectors $\vec{u}$ and $\vec{v}$ is:

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

- Using low-level matrix multiplication definition:

$$\vec{u} \cdot \vec{v} = \sum_{i=1}^{n} u_i v_i$$

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

$$\vec{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \qquad \vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

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

$$= \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = u_1 v_1 + u_2 v_2 + u_3 v_3$$

Dot Product Example

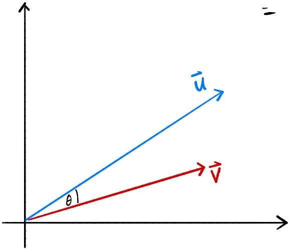
$$\vec{u} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \qquad \vec{v} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} \qquad \vec{u} \cdot \vec{v} = 4 + 10 + 18$$

$$= 32$$

Geometric Interpretation of Dot Product

- $\vec{u} \cdot \vec{v} = \|\vec{u}\| \|\vec{v}\| \cos\theta$.

$$\vec{u} \cdot \vec{u} = \|\vec{u}\|^2 \cos 0$$
$$= \|\vec{u}\|^2$$



Discussion Question

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

$\vec{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$ $\vec{u} \cdot \vec{u} = u_1^2 + u_2^2 + u_3^2$ $\|\vec{u}\|$

$\sqrt{\vec{u} \cdot \vec{u}} = \sqrt{u_1^2 + u_2^2 + u_3^2}$

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

Properties of the Dot Product

- Commutative: $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$
- Distributive: $\vec{u} \cdot (\vec{v} + \vec{w}) = \vec{u} \cdot \vec{v} + \vec{u} \cdot \vec{w}$
- Linear: $\vec{u} \cdot (\alpha\vec{v} + \beta\vec{w}) = \alpha\vec{u} \cdot v + \beta\vec{u} \cdot \vec{w}$

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.

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

Matrices and Functions

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

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

- Matrix multiplication evaluates that function.

For more, see www.dsc40a.com

Today

- How do we predict salary given **multiple** features?
  - years of experience, number of internships, GPA, etc.

Using Multiple Features

- We believe salary is a function of experience *and* GPA.

- I.e., there is a function H so that:

$$\text{salary} \approx H(\text{years of experience}, \text{GPA})$$

- Recall: H is a **prediction rule**.

- **Our goal**: find a good prediction rule, H.

Example Prediction Rules

$$H_1(\text{experience}, \text{GPA}) = \$40{,}000 \times \frac{\text{GPA}}{4.0} + \$2,000 \times (\text{experience})$$

$$H_2(\text{experience}, \text{GPA}) = \$60{,}000 \times 1.05^{(\text{experience}+\text{GPA})}$$

$$H_3(\text{experience}, \text{GPA}) = \sin(\text{GPA}) + \cos(\text{experience})$$

Linear Prediction Rule

- We'll restrict ourselves to **linear** prediction rules:

$$H(\text{experience}, \text{GPA}) = w_0 + w_1 \times (\text{experience}) + w_2 \times (\text{GPA})$$

- Can add more features, too[1]:

$$H(\text{experience}, \text{GPA}, \#\text{ internships}) =$$
$$w_0 + w_1 \times (\text{experience}) + w_2 \times (\text{GPA})$$
$$+ w_3 (\#\text{ of internships})$$

- Interpretation of w_i: the *weight* of feature x_i.

[1]In practice, might use tens, hundreds, even thousands of features.

Feature Vectors

- In general, if $x_1, \ldots, x_d$ are d features:

$$H(x_1, \ldots, x_d) = w_0 + \underbrace{w_1 x_1 + w_2 x_2 + \ldots + w_d x_d}_{\vec{x} \cdot \vec{w}}$$

- Nicer to pack into a **feature vector** and **parameter vector**:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \qquad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

- Then:

$$H(\vec{x}) = w_0 + \vec{w} \cdot \vec{x}$$

Example

- Recall the prediction rule:

$$H_1(\text{experience}, \text{GPA}) = \$40{,}000 \times \frac{\text{GPA}}{4.0} + \$2{,}000 \times (\text{experience})$$

- This is linear. If x_1 is experience, x_2 is GPA, then:

$$\vec{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 2{,}000 \\ 10{,}000 \end{pmatrix} \qquad w_0 = 0$$

- Our prediction for someone with 2 years experience, 3.0 GPA:

$$\vec{x} = \begin{pmatrix} 2 \\ 3.0 \end{pmatrix} \qquad H(\vec{x}) = w_0 + \vec{w} \cdot \vec{x} = 0 + \begin{pmatrix} 2000 & 10000 \end{pmatrix} \begin{pmatrix} 2 \\ 3.0 \end{pmatrix} = 4000 + 30000 = 34000$$

The Data

- For each person, collect 3 features, plus salary:

Person #	Experience	GPA	# Internships	Salary	
1	3	3.7	1	85,000	$= y_1$
2	6	3.3	2	95,000	$= y_2$
3	10	3.1	3	105,000	$= y_3$

- We represent each person with a **data vector**:

$$\vec{x}^{(1)} = \begin{pmatrix} 3 \\ 3.7 \\ 1 \end{pmatrix}, \qquad \vec{x}^{(2)} = \begin{pmatrix} 6 \\ 3.3 \\ 2 \end{pmatrix}, \qquad \vec{x}^{(3)} = \begin{pmatrix} 10 \\ 3.1 \\ 3 \end{pmatrix}$$

Notation

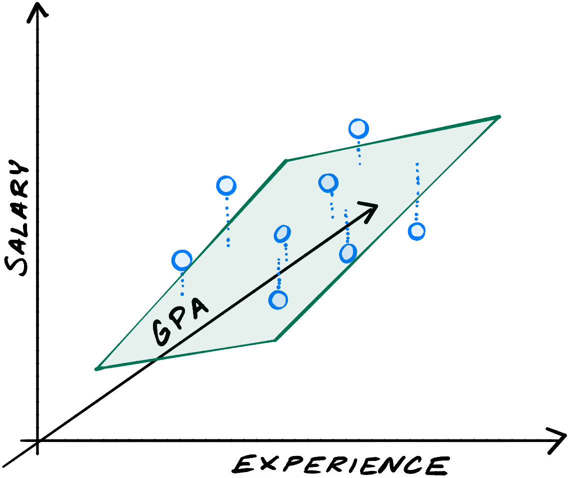
- $\vec{x}^{(i)}$ is the ith data vector.

- $x_j^{(i)}$ is the jth feature in the ith data vector.

- If there are d features:

$$\vec{x}^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_d^{(i)} \end{pmatrix}$$

Geometric Interpretation



The General Problem

- We have n data points (or **training examples**):
  $\left(\vec{x}^{(1)}, y_1\right), \ldots, \left(\vec{x}^{(n)}, y_n\right)$

  salary of 1st person

  feature vector for 1st person

- We want to find a good linear prediction rule:

$$H(\vec{x}) = w_0 + \vec{w} \cdot \vec{x}$$

- To do so, we'll minimize the mean squared error:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} \left(H(\vec{x}) - y_i\right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\left(w_0 + \vec{w} \cdot \vec{x}^{(i)}\right) - y_i\right)^2$$

The Risk

- With d features, we have $d + 1$ parameters: $w_0, w_1, \ldots, w_d$.
- The risk $R_{\text{sq}}(\vec{w})$ is a function from $\mathbb{R}^{d+1}$ to $\mathbb{R}^1$.
- It is a $(d + 1)$-dimensional hypersurface.
- **No hope of visualizing it directly when $d \geq 2$.**

Rewriting the Mean Squared Error

- Let $\vec{e}$ be such that e_i is the (signed) error on ith example:

$$e_i = (w_0 + \vec{w} \cdot \vec{x}^{(i)}) - y_i$$

- Then:

$$\begin{aligned} R_{\text{sq}}(\vec{w}) &= \frac{1}{n} \sum_{i=1}^{n} \left(\left(w_0 + \vec{w} \cdot \vec{x}^{(i)} \right) - y_i \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} e_i^2 \end{aligned}$$

Rewriting the Mean Squared Error

- Let $\vec{e}$ be such that e_i is the (signed) error on ith example:

$$e_i = (w_0 + \vec{w} \cdot \vec{x}^{(i)}) - y_i$$

- Then:

$$\begin{aligned} R_{\text{sq}}(\vec{w}) &= \frac{1}{n} \sum_{i=1}^{n} \left(\left(w_0 + \vec{w} \cdot \vec{x}^{(i)} \right) - y_i \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} e_i^2 \\ &= \frac{1}{n} \vec{e} \cdot \vec{e} \\ &= \frac{1}{n} \|\vec{e}\|^2 \end{aligned}$$

Rewriting the Mean Squared Error

- Define $\vec{y} = (y_1, \dots, y_n)^T$. Then:

$$\vec{e} = \begin{pmatrix} (w_0 + \vec{w} \cdot \vec{x}^{(1)}) - y_1 \\ (w_0 + \vec{w} \cdot \vec{x}^{(2)}) - y_2 \\ \vdots \\ (w_0 + \vec{w} \cdot \vec{x}^{(n)}) - y_n \end{pmatrix} = \underbrace{\begin{pmatrix} w_0 + \vec{w} \cdot \vec{x}^{(1)} \\ w_0 + \vec{w} \cdot \vec{x}^{(2)} \\ \vdots \\ w_0 + \vec{w} \cdot \vec{x}^{(n)} \end{pmatrix}}_{\vec{h}} - \vec{y}$$

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

- $\vec{h}$ is the vector of predictions.

Rewriting the Mean Squared Error

- So far: $R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{e}\|^2$, and $\vec{e} = \vec{h} - \vec{y}$.
- Therefore:

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

- $\vec{w}$ is hidden inside of $\vec{h}$, let's pull it out.

Rewriting the Mean Squared Error

- Define the **design matrix** X: $n \times (d+1)$

2nd feature of 1st person

$$X = \begin{pmatrix} 1 & \vec{x}^{(1)} \longrightarrow \\ 1 & \vec{x}^{(2)} \longrightarrow \\ \vdots & \vdots \\ 1 & \vec{x}^{(n)} \longrightarrow \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_d^{(n)} \end{pmatrix}$$

- Then $\vec{h} = X\vec{w}$.

$$\underbrace{\begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix}}_{\vec{w}} = \begin{pmatrix} w_0 + w_1 x_1^{(1)} + w_2 x_2^{(1)} \\ w_0 + w_1 x_1^{(2)} + w_2 x_2^{(2)} \\ w_0 + w_1 x_1^{(3)} + w_2 x_2^{(3)} \end{pmatrix} = \begin{pmatrix} H(\vec{x}^{(1)}) \\ H(\vec{x}^{(2)}) \\ H(\vec{x}^{(3)}) \end{pmatrix} = \vec{h}$$

Rewriting the Mean Squared Error

- The mean squared error is:

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

  where X is the **design matrix** containing the data, $\vec{w}$ is the **parameter vector**, and $\vec{y}$ is the vector of **observations** (or right answers).

- To minimize MSE: take derivative (gradient), set to zero, solve.