DSC 40A

Theoretical Foundations of Data Science I

Last Time

- We used linear algebra to fit a prediction rule of the form

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

- Instead of using our formulas for w_0 and w_1, we can find these parameters by solving the **normal equations**:

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

where

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

In This Video

Can we change the form of the prediction rule to be nonlinear?

Recommended Reading

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

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.

- When our prediction rule is

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

the hypothesis vector $\vec{h}$ can be written

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

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.

- When our prediction rule is

$$H(x) = w_0 + w_1 x + w_2 x^2$$

the hypothesis vector $\vec{h}$ can be written

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

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.

- When our prediction rule is

$$H(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

the hypothesis vector $\vec{h}$ can be written

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

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.

- When our prediction rule is

$$H(x) = w_1 \frac{1}{x^2} + w_2 \sin x + w_3 e^x$$

the hypothesis vector $\vec{h}$ can be written

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

Minimizing the Mean Squared Error

- As long as the form of the prediction rule permits us to write $\vec{h} = X\vec{w}$ for some X and $\vec{w}$, the mean squared error is

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

- Regardless of the values of X and $\vec{w}$,

$$\begin{aligned} \frac{dR_{\text{sq}}}{d\vec{w}} &= 0 \\ \implies -2X^T\vec{y} + 2X^TX\vec{w} &= 0 \\ \implies X^TX\vec{w} &= X^T\vec{y}. \end{aligned}$$

- Solving the **normal equations** still works!

Linear in the Parameters

- We can fit rules like:

$$w_0 + w_1 x + w_2 x^2 \qquad w_1 e^{-x^2} + w_2 \cos(x + \pi) + w_3 \frac{\log 2x}{x}$$

- This includes arbitrary polynomials.

- We can't fit rules like:

$$e^{w_1 x} + w_0 \qquad \sin(w_1 x + w_0)$$

- We can have any number of parameters, but must be **linear in the parameters**.

Determining Function Form

- How do we know what form our prediction rule should take?
- Sometimes, we know from *theory*, using knowledge about what the variables represent and how they should be related.
- Other times, we make a guess based on the data.
- Generally, start with simpler functions first.

Question

Suppose you collect data on the height, or position, of a freefalling object at various times t_i. Which form should your prediction rule take to best fit the data?

a) constant, $H(t) = w_0$

b) linear, $H(t) = w_0 + w_1 t$

c) quadratic, $H(t) = w_0 + w_1 t + w_2 t^2$

d) no way to know without plotting the data

Example: Amdahl's Law

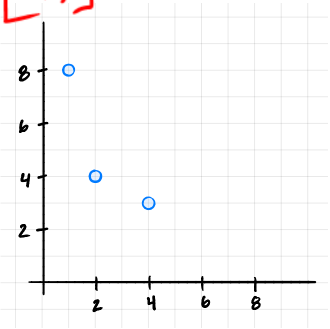
- Amdahl's Law relates the runtime of a program on p processors to the time to do the sequential and nonsequential parts on one processor.

$$H(p) = \frac{t_{\text{NS}}}{p} + t_{\text{S}}$$

- Collect data by timing a program with varying numbers of processor:

Processors	Time (Hours)
1	8
2	4
4	3

Example: fitting $H(x) = w_1 \cdot \frac{1}{x_i} + w_0$



x_i	y_i
1	8
2	4
4	3

$\vec{w} = \begin{bmatrix} 1 \\ 48/7 \end{bmatrix}$

theory

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

normal eqns $X^T X w = X^T y$

$y = \begin{bmatrix} 8 \\ 4 \\ 3 \end{bmatrix}$

$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1/2 & 1/4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1/2 \\ 1 & 1/4 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1/2 & 1/4 \end{bmatrix} \begin{bmatrix} 8 \\ 4 \\ 3 \end{bmatrix}$

$(X^T X)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1/2 & 1/4 \end{bmatrix} \begin{bmatrix} 8 \\ 4 \\ 3 \end{bmatrix}$

Example: Amdahl's Law

- We found: $t_{\mathsf{NS}} = \frac{48}{7} \approx 6.88, \quad t_{\mathsf{S}} = 1$
- Therefore our prediction rule is:

$$H(p) = \frac{t_{\mathsf{NS}}}{p} + t_{\mathsf{S}}$$

$$= \frac{6.88}{p} + 1$$

Demo

Summary

- Whenever our prediction rule is **linear in the parameters**, we can use the **normal equations**

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

  to find the parameters than minimize the mean squared error.

- This means we can use regression to fit prediction rules of many forms, including arbitrary polynomials.

- **Next time:** We'll use the normal equations to make predictions based on multiple features.