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

$$\nabla_w R_{sq}(\vec{w}) = \frac{d}{d\vec{w}} R_{sq}(\vec{w})$$

$$= 2X^TX\vec{w} - 2X^T\vec{y}$$

$$\boxed{(X^TX)\vec{w} = X^T\vec{y}}$$

DSC 40A

Lecture 08

Least Squares Regression, pt. IV

Last Time

- How do we make predictions using multiple features?
- Assume a linear decision rule:

  H(experience, GPA, # internships) =

  $w_0 + w_1 \times \text{(experience)} + w_2 \times \text{(GPA)} + w_3 \times \text{(\# of internships)}$
- In general:

$$H(x_1, \dots, x_d) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d$$

Feature Vectors

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

Feature Vectors

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

- **Actually, we should include w_0 in $\vec{w}$...**

Augmented Feature Vectors

- The **augmented feature vector** $\text{Aug}(\vec{x})$ is the vector obtained by adding a 1 to the front of $\vec{x}$:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \qquad \text{Aug}(\vec{x}) = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \qquad \vec{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

- Then:

$$\begin{aligned} H(x_1, \ldots, x_d) &= w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d \\ &= \text{Aug}(\vec{x}) \cdot \vec{w} \end{aligned}$$

Last Time

- We want to fit a decision rule of the form $H(\vec{x}) = \text{Aug}(\vec{x}) \cdot \vec{w}$.
- Minimize **mean squared error:**

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

Rewriting the Mean Squared Error

- Define the **design matrix**:

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

- And the vector of **observations**: $\vec{y} = (y_1, \dots, y_n)^T$

Rewriting the Mean Squared Error

- Then:

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

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

- Today's goal: find the $\vec{w}$ that minimizes the MSE.

Minimizing the Mean Squared Error

- Our goal: minimize the function:

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

- Strategy:
  1. Take partial derivatives,

$$\frac{\partial R_{\text{sq}}}{\partial w_0}(\vec{w}), \quad \frac{\partial R_{\text{sq}}}{\partial w_1}(\vec{w}), \quad \frac{\partial R_{\text{sq}}}{\partial w_2}(\vec{w}), \quad \dots \quad \frac{\partial R_{\text{sq}}}{\partial w_d}(\vec{w})$$

  2. Set each equal to zero and solve for $w_0, w_1, \dots, w_d$.

Minimizing the MSE: Gradient Edition

- The vector of partial derivatives is called the **gradient**:

$$\left(\frac{\partial R_{\text{sq}}}{\partial w_0}(\vec{w}), \quad \frac{\partial R_{\text{sq}}}{\partial w_1}(\vec{w}), \quad \frac{\partial R_{\text{sq}}}{\partial w_2}(\vec{w}), \quad \ldots, \quad \frac{\partial R_{\text{sq}}}{\partial w_d}(\vec{w})\right)^T$$

- Written: $\nabla_{\vec{w}} R_{\text{sq}}(\vec{w})$ or $\frac{dR_{\text{sq}}}{d\vec{w}}(\vec{w})$

- Strategy:
  1. Compute the gradient of $R_{\text{sq}}(\vec{w})$.
  2. Set it to zero and solve for $\vec{w}$.

Gradients Review

Computing Gradients

When computing $\frac{df}{d\vec{x}}(\vec{x})$:

- Before: make sure that f takes in vectors, outputs scalars.
  - **Example**: $\frac{d}{d\vec{x}}\left[A\vec{x}\right]$
  - **Example**: $\frac{d}{d\vec{x}}\left[\vec{x}\cdot\vec{x}\right]$, $\frac{d}{d\vec{x}}\left[\vec{x}^T A^T A\vec{x}\right]$
- After: make sure your result is a vector.

Finding the Gradient: Strategy #1

Example: Find $\frac{d}{d\vec{x}}\left[\vec{a} \cdot \vec{x}\right]$ where $\vec{x}$ and $\vec{a}$ have d elements.

1. "Unpack" all matrix multiplications/dot products
   - $\vec{a} \cdot \vec{x} =$

Finding the Gradient: Strategy #1

Example: Find $\frac{d}{d\vec{x}}\left[\vec{a}\cdot\vec{x}\right]$ where $\vec{x}$ and $\vec{a}$ have d elements.

1. "Unpack" all matrix multiplications/dot products
   - $\vec{a}\cdot\vec{x} = a_1 x_1 + a_2 x_2 + \ldots + a_d x_d$

2. Take partial derivatives (perhaps with arbitrary index):

$$\frac{\partial}{\partial x_1}\left[a_1 x_1 + a_2 x_2 + \ldots + a_d x_d\right] =$$

$$\frac{\partial}{\partial x_2}\left[a_1 x_1 + a_2 x_2 + \ldots + a_d x_d\right] =$$

$$\vdots$$

$$\frac{\partial}{\partial x_d}\left[a_1 x_1 + a_2 x_2 + \ldots + a_d x_d\right] =$$

Finding the Gradient: Strategy #1

3. Pack partial derivatives into a gradient vector:

$$\frac{d}{d\vec{x}}\left[\vec{a}\cdot\vec{x}\right] = \left(a_1, a_2, \ldots, a_d\right)^T$$

4. Simplify:

$$\left(a_1, a_2, \ldots, a_d\right)^T = \vec{a}$$

- So $\frac{d}{d\vec{x}}\left[\vec{a}\cdot\vec{x}\right] = \vec{a}$
- Check: **result is a vector**.

Finding the Gradient: Strategy #1

- **Pro**: Always works, straightforward
- **Con**: Unpacking everything can get messy

Example

Show that $\frac{d}{d\vec{x}}\left[\vec{x}^T A^T A\vec{x}\right] = 2A^T A\vec{x}$, where A is $n \times d$ and $\vec{x}$ is $n \times 1$.

- Check: **it is a scalar**

1. After unpacking: $\vec{x}^T A^T A\vec{x} = \sum_{i=1}^{n} \left(\sum_{j=1}^{d} A_{ij} x_j \right)^2$

2. Take partial derivatives:

$$\frac{\partial}{\partial x_1}\left[\sum_{i=1}^{n} \left(\sum_{j=1}^{d} A_{ij} x_j \right)^2 \right] = \sum_{i=1}^{n} \sum_{j=1}^{d} A_{i1} A_{ij} x_j$$

Example

3. Pack into a gradient vector:

$$\frac{d}{d\vec{x}}\left[\vec{x}^T A^T A \vec{x}\right] = \begin{pmatrix} \sum_{i=1}^n \sum_{j=1}^d A_{i1} A_{ij} x_j \\ \sum_{i=1}^n \sum_{j=1}^d A_{i2} A_{ij} x_j \\ \vdots \\ \sum_{i=1}^n \sum_{j=1}^d A_{id} A_{ij} x_j \end{pmatrix}$$

4. Somehow simplify this to $A^T A \vec{x}$...

Finding the Gradient: Strategy #2

Chain Rule: If $f : \mathbb{R} \to \mathbb{R}$, and $g : \mathbb{R}^d \to \mathbb{R}$, then:

$$\frac{d}{d\vec{x}} f(g(\vec{x})) = \frac{df}{dg}\frac{dg}{d\vec{x}}$$

Example: What is $\frac{d}{d\vec{x}}\left[(\vec{a}\cdot\vec{x})^2\right]$?

- $f(g) =$
- $g(\vec{x}) =$
- $\frac{d}{d\vec{x}}\left[(\vec{a}\cdot\vec{x})^2\right] =$

Finding the Gradient: Strategy #2

1. Unpack until we can use chain rule, but no more.

2. Use the chain rule.

3. Simplify.

Recall

Suppose A is $n \times d$.
Let $\vec{A}_{i*}$ denotes its ith row. Then:

$$A\vec{x} = \begin{pmatrix} \vec{A}_{1*} \cdot \vec{x} \\ \vec{A}_{2*} \cdot \vec{x} \\ \vdots \\ \vec{A}_{n*} \cdot \vec{x} \end{pmatrix}$$

Let $\vec{A}_{*j}$ denotes its jth column, then:

$$A\vec{x} = \vec{A}_{*1}x_1 + \vec{A}_{*2}x_2 + \ldots + \vec{A}_{*d}x_d$$

Finding the Gradient: Strategy #2

Show that $\frac{d}{d\vec{x}}\left[\vec{x}^T A^T A\vec{x}\right] = 2A^T A\vec{x}$, where A is $n \times d$ and $\vec{x}$ is $n \times 1$.

1. Unpack $\vec{x}^T A^T A\vec{x} =$

Finding the Gradient: Strategy #2

Show that $\frac{d}{d\vec{x}}\left[\vec{x}^T A^T A\vec{x}\right] = 2A^T A\vec{x}$, where A is $n \times d$ and $\vec{x}$ is $n \times 1$.

2. Use chain rule:

Finding the Gradient: Strategy #2

Show that $\frac{d}{d\vec{x}}\left[\vec{x}^T A^T A \vec{x}\right] = 2A^T A\vec{x}$, where A is $n \times d$ and $\vec{x}$ is $n \times 1$.

3. Show that this $= 2A^T A\vec{x}$.

Back to Regression...

Minimizing the MSE

- We want to compute:

$$\frac{d}{d\vec{w}}\left[R_{\text{sq}}(\vec{w})\right] = \frac{d}{d\vec{w}}\left[\|X\vec{w} - \vec{y}\|^2\right]$$

- Step 1: Rewrite squared norm using dot product. Recall:

$$\begin{aligned}
(A+B)^T &= A^T + B^T \\
(AB)^T &= B^T A^T \\
\vec{u}\cdot\vec{v} &= \vec{v}\cdot\vec{u} \\
(\vec{u}+\vec{v})\cdot(\vec{w}+\vec{z}) &= \vec{u}\cdot\vec{w} + \vec{u}\cdot\vec{x} + \vec{v}\cdot\vec{w} + \vec{v}\cdot\vec{z} \\
\|\vec{u}\|^2 &= \vec{u}\cdot\vec{u}
\end{aligned}$$

Step 1: Rewriting squared norm

$$
\begin{aligned}
\|X\vec{w} - \vec{y}\|^2 &= \\
&= \\
&= \\
&=
\end{aligned}
$$

Step 2: Take gradients

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

$$=$$

The Normal Equations

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

$$2X^TX\vec{w} - 2X^T\vec{y} = 0 \implies X^TX\vec{w} = X^T\vec{y}$$

- This is a system of equations in matrix form, called the **normal equations**.

- Solution[1]: $\vec{w} = (X^TX)^{-1}X^T\vec{y}$.

[1] Don't actually compute inverse! Use Gaussian elimination.

Regression with Multiple Features

- We want to find $\vec{w}$ which minimizes $\|X\vec{w} - \vec{y}\|^2$.

- The answer: $\vec{w} = (X^T X)^{-1} X^T \vec{y}$.