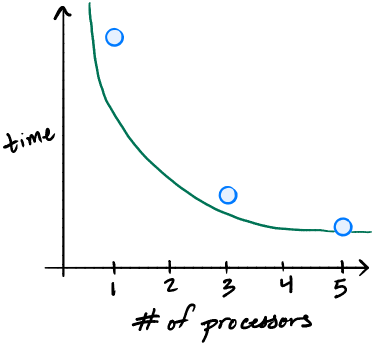


DSC 40A

Lecture 07

Least Squares Regression, pt. II

Last Time

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

- Minimize mean absolute error?

$$\frac{1}{n}\sum_{i=1}^{n}\left|H(x_i) - y_i\right|$$

- **Not differentiable**, instead minimize **mean squared error**:

$$\frac{1}{n}\sum_{i=1}^{n}\left(H(x_i) - y_i\right)^2$$

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

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

Last Time

- **Goal**: find w_1 and w_0 which minimize MSE:

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

- **Strategy**: Take derivatives $\partial R_{\text{sq}}/\partial w_1$ and $\partial R_{\text{sq}}/\partial w_0$, set to zero, solve.

- We found:

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

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

Today

- Solve these equations to find the **least squares solutions**.
- See how to easily fit non-linear trends, too.

Strategy

$$0 = \frac{2}{n} \sum_{i=1}^{n} \left((w_1 x_i + w_0) - y_i\right) x_i \qquad 0 = \frac{2}{n} \sum_{i=1}^{n} \left((w_1 x_i + w_0) - y_i\right)$$

1. Solve for w_0 in second equation.
2. Plug solution for w_0 into first equation, solve for w_1.

Solve for w_0

$$0 = \frac{2}{n}\sum_{i=1}^{n}\left((w_1 x_i + w_0) - y_i\right)$$

Solve for w_0

$$0 = \frac{2}{n} \sum_{i=1}^{n} \left((w_1 x_i + w_0) - y_i \right)$$

Key Fact

- Define

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Then

$$\sum_{i=1} (x_i - \bar{x}) = 0 \qquad \sum_{i=1} (y_i - \bar{y}) = 0$$

Solve for w_1

$$0 = \frac{2}{n} \sum_{i=1}^{n} \left((w_1 x_i + w_0) - y_i \right) x_i \qquad w_0 = \bar{y} - w_1 \bar{x}$$

Solve for w_1

$$0 = \frac{2}{n} \sum_{i=1}^{n} \left((w_1 x_i + w_0) - y_i \right) x_i \qquad w_0 = \bar{y} - w_1 \bar{x}$$

Least Squares Solutions

- The **lease squares solutions** for the slope w_1 and intercept w_0 are:

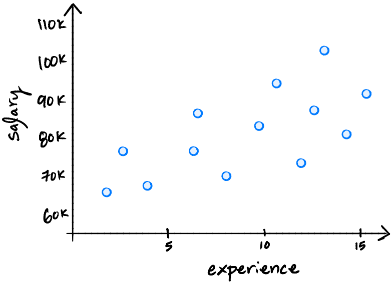
$$w_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad w_0 = \bar{y} - w_1\bar{x}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Interpretation of Slope

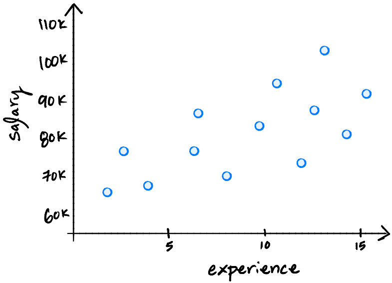
$$w_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



- What is the sign of $(x_i - \bar{x})(y_i - \bar{y})$ when:
  - $x_i > \bar{x}$ and $y_i > \bar{y}$?
  - $x_i < \bar{x}$ and $y_i < \bar{y}$?
  - $x_i > \bar{x}$ and $y_i < \bar{y}$?
  - $x_i < \bar{x}$ and $y_i > \bar{y}$?

Interpretation of Intercept

$$w_0 = \bar{y} - w_1 \bar{x}$$



- What is $H(\bar{x})$?

Discussion Question

We fit a linear prediction rule for salary given years of experience. Then everyone gets a $5,000 raise. Which of these happens?

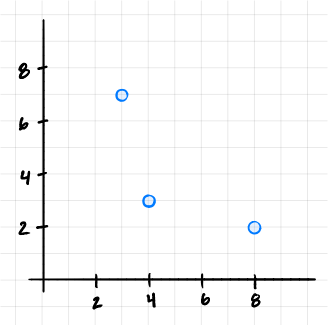
a) slope increases, intercept increases

b) slope decreases, intercept increases

c) slope stays same, intercept increases

d) slope stays same, intercept stays same

Example



$\bar{x} =$

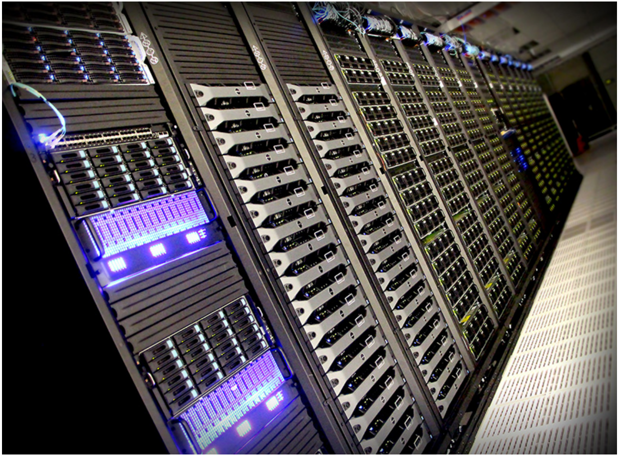
$\bar{y} =$

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

$$w_0 = \bar{y} - w_1 \bar{x}$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
3	7				
4	3				
8	2				

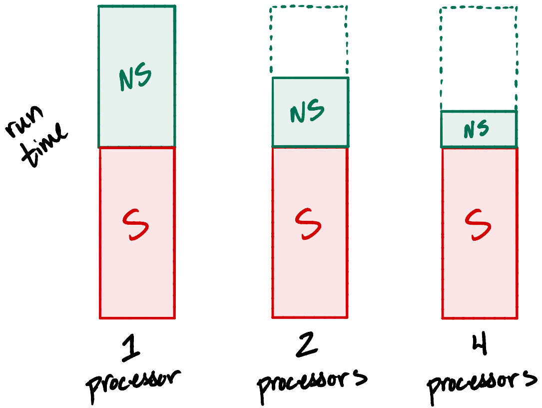
Example: Parallel Processing



Problem

- Some parts of a program are necessarily **sequential**.
- E.g., downloading the data must happen before analysis.
- More processors do not speed up **sequential** code.
- But they do speed up **non-sequential** code.

Speedup



Amdahl's Law

The time T it takes to run a program on p processors is:

$$T(p) = t_{\mathrm{S}} + \frac{t_{\mathrm{NS}}}{p}$$

where t_{S} and t_{NS} are the time it takes the sequential and non-sequential parts to run on one processor, respectively.

Amdahl's Law

The time T it takes to run a program on p processors is:

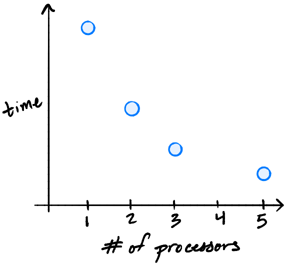
$$T(p) = t_\mathrm{S} + \frac{t_\mathrm{NS}}{p}$$

where t_S and t_NS are the time it takes the sequential and non-sequential parts to run on one processor, respectively.

Problem: we don't know t_S and t_NS.

Fitting Amdahl's Law

- **Solution**: we will learn t_S and t_{NS} from data.
- Run with varying number of processors, record total time:



- Find decision rule $H(p) = \dfrac{t_{NS}}{p} + t_S$ by minimizing MSE.

General Problem

- Given data $(x_1, y_1), \dots, (x_n, y_n)$.
- Fit a **non-linear** rule $H(x) = w_1 \cdot \frac{1}{x} + w_0$ by minimizing MSE:

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

Using definition of H:

Minimizing MSE

- Take derivatives, you'll find:

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

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

Minimizing MSE

- Set to zero, solve. You'll find:

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

Minimizing MSE

- Set to zero, solve. You'll find:

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

- Define $z_i = \frac{1}{x_i}$, $\quad \bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}$. Then:

$$w_1 = \qquad\qquad\qquad w_0 =$$

Fitting Non-Linear Trends

To fit a prediction rule of the form $H(x) = w_1 \cdot \frac{1}{x} + w_0$:

1. Create a new data set $(z_1, y_1), \ldots, (z_n, y_n)$, where $z_i = \frac{1}{x_i}$.

2. Fit $H(z) = w_1 z + w_0$ using familiar least squares solutions:

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

3. Use w_1 and w_0 in original decision rule, $H(x)$.

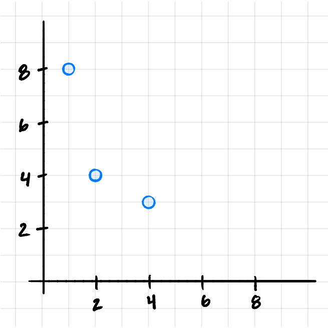
Example: Amdahl's Law

- We have timed our program:

Processors	Time (Hours)
1	8
2	4
4	3

- Fit prediction rule: $H(p) = \dfrac{t_{\mathrm{NS}}}{p} + t_{\mathrm{S}}$

Example: fitting $H(x) = w_1 \cdot \frac{1}{x_i} + x_0$



$\bar{z} =$

$\bar{y} =$

$$w_1 = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n} (z_i - \bar{z})^2} =$$

$$w_0 = \bar{y} - w_1 \bar{z}$$

x_i	z_i	y_i	$(z_i - \bar{z})$	$(y_i - \bar{y})$	$(z_i - \bar{z})(y_i - \bar{y})$	$(z_i - \bar{z})^2$
3	7					
4	3					
8	2					

Example: Amdahl's Law

- We found: $t_{\text{NS}} = \frac{48}{7} \approx 6.88, \quad t_{\text{S}} = 1$
- Our prediction rule:

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

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

Fitting Non-Linear Trends

To fit a prediction rule of the form $H(x) = w_1 \cdot f(x) + w_0$:

1. Create a new data set $(z_1, y_1), \ldots, (z_n, y_n)$, where $z_i = f(x_i)$.
2. Fit $H(z) = w_1 z + w_0$ using familiar least squares solutions:

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

3. Use w_1 and w_0 in original decision rule, $H(x)$.

Fitting Non-Linear Trends

- We can fit rules like:

$$w_1 x + w_0 \qquad w_1 \cdot \frac{1}{x} + w_0 \qquad w_1 x^2 + w_0 \qquad w_1 e^x + w_0$$

- We can't fit rules like:

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

- Can fit as long as **linear** function of w_1, w_0.

What's Left?

- How do we make predictions with lots of features?
- E.g., experience, age, GPA, number of internships, etc.