DSC 40A

Theoretical Foundations of Data Science I

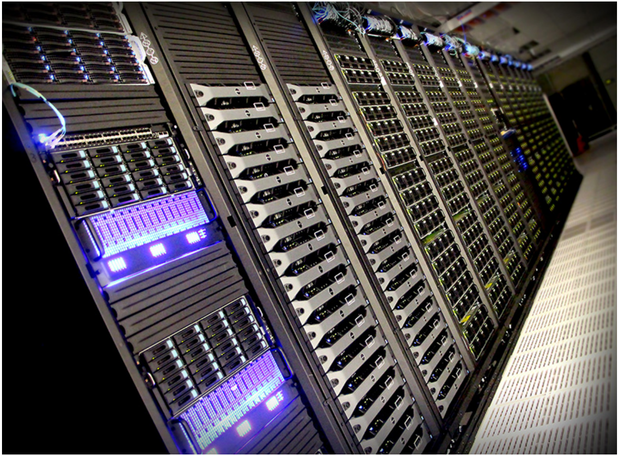
In This Video

Can we use linear regression to fit nonlinear functions to data?

Recommended Reading

Course Notes: Chapter 2, Section 1

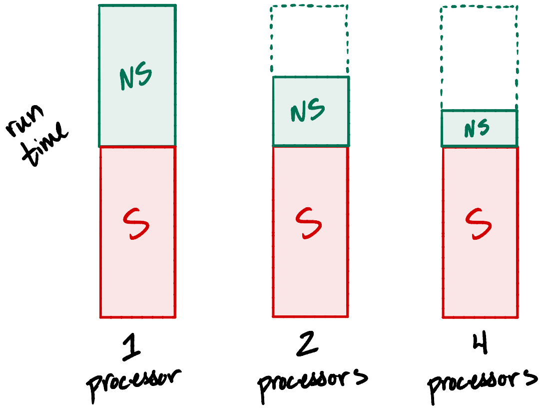
Example: Parallel Processing



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:

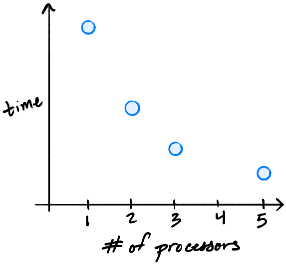
$$T(p) = t_S + \frac{t_{NS}}{p}$$

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 prediction rule $H(p) = \frac{t_{NS}}{p} + t_S$ by minimizing MSE.

General Problem

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

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

Using definition of H:

Minimizing MSE

- Take partial derivatives, 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

- Take partial derivatives, set to zero, solve. You'll find:

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

- Define

$$z_i = \frac{1}{x_i}, \qquad \bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}$$

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

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:

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

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

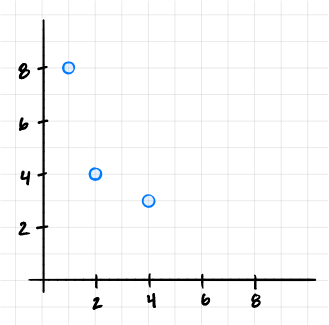
Example: Amdahl's Law

- We have timed our program:

Processors	Time (Hours)
1	8
2	4
4	3

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

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



$\bar{z} =$

$\bar{y} =$

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

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

x_i	y_i	z_i	$(z_i - \bar{z})$	$(y_i - \bar{y})$	$(z_i - \bar{z})(y_i - \bar{y})$	$(z_i - \bar{z})^2$
1	8					
2	4					
4	3					

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:

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

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

Linear in the Parameters

- 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:

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

- Has to be **linear in the parameters**, or linear as a function of w_1, w_0.

Transformations

- Try rewriting functions to see if they can be expressed as linear functions in new variables.
- **Example**

$$H(x) = c_0 x^{c_1}$$

Transformations

$$y = c_0 x^{c_1}$$
$$\log y = \log c_0 + c_1 \log x$$

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

$$w_0 = \frac{1}{n} \sum_{i=1}^{n} \log y_i - w_1 \cdot \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

General Strategy

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

1. Create a new data set $(z_1, v_1), \ldots, (z_n, v_n)$, where $z_i = f(x_i)$ and $v_i = g(y_i)$.

2. Fit $v = w_1 z + w_0$ using familiar least squares solutions:

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

   where $\bar{z}$ is the mean of the z_i's, $\bar{v}$ is the mean of the v_i's.

3. If necessary, use w_0 and w_1 to find the parameters of the original prediction rule.

Summary

- We can sometimes fit nonlinear functions to data by thinking of these non-linear functions as linear functions in new variables.
- **Next Time:** Using linear algebra to do regression helps us fit even more non-linear functions to data and allows us to make predictions based on multiple features.
- E.g., experience, highest education level, GPA, number of internships, etc.