CS 145 Discussion 2

Reminders

- HW1 out, due 10/19/2017 (Thursday)
- Group formations for course project due today (1 pt)
- Join Piazza (email: juwood03@ucla.edu)

Overview

- Linear Regression
- Z Score Normalization
- Multidimensional Newton’s Method
- Decision Tree
- Twitter Crawler
- Likelihood

Linear Regression

- Linear model to predict value of a variable *y* using features *x*

$$y = \boldsymbol{x}^T \boldsymbol{\beta} = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$$

- Least Square Estimation

$$J(\boldsymbol{\beta}) = \frac{1}{2}(\mathrm{X}\boldsymbol{\beta} - y)^T(X\boldsymbol{\beta} - y)$$

- Closed form solution

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T y$$

A ball is rolled down a hallway and its position is recorded at five different times. Use the table shown below to calculate

- Weights
- Predicted position at each given time and at time 12 seconds

Time (seconds)	Position (meters)
1	9
2	12
4	17
6	21
8	26

Step 1: Calculate Weights

- What are X and Y variables?
- What are the parameters for our problem?
- Calculating parameters

Time (seconds)	Position (meters)
1	9
2	12
4	17
6	21
8	26

Step 1: Calculate Weights

- What are X and Y variables?
  - Time (X_1) and Position(Y)
- What are the parameters for our problem?
  - $\hat{\beta}_1$ for time and $\hat{\beta}_0$ for intercept
- Calculating parameters

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$$

Time (seconds)	Position (meters)
1	9
2	12
4	17
6	21
8	26

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \qquad y = \begin{bmatrix} 9 \\ 12 \\ 17 \\ 21 \\ 26 \end{bmatrix}$$

Time (seconds)	Position (meters)
1	9
2	12
4	17
6	21
8	26

$$X^T X = ?$$

$$(X^T X)^{-1} = ?$$

$$X^T y = ?$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$= ?$$

$$\hat{\beta}_0 = ?$$
$$\hat{\beta}_1 = ?$$

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \qquad y = \begin{bmatrix} 9 \\ 12 \\ 17 \\ 21 \\ 26 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 5 & 21 \\ 21 & 121 \end{bmatrix}$$

$$(X^T X)^{-1} = ?$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\underbrace{ad-bc}_{\text{determinant}}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$X^T y = ?$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$= ?$$

$$\hat{\beta}_0 = ?$$
$$\hat{\beta}_1 = ?$$

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix}$$

$$y = \begin{bmatrix} 9 \\ 12 \\ 17 \\ 21 \\ 26 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 5 & 21 \\ 21 & 121 \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{164} \begin{bmatrix} 121 & -21 \\ -21 & 5 \end{bmatrix}$$

$$X^T y = ?$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$= ?$$

$$\hat{\beta}_0 = ?$$
$$\hat{\beta}_1 = ?$$

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \qquad y = \begin{bmatrix} 9 \\ 12 \\ 17 \\ 21 \\ 26 \end{bmatrix} \qquad X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 5 & 21 \\ 21 & 121 \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{164} \begin{bmatrix} 121 & -21 \\ -21 & 5 \end{bmatrix} \qquad X^T y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} 9 \\ 12 \\ 17 \\ 21 \\ 26 \end{bmatrix} = \begin{bmatrix} 85 \\ 435 \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$= ?$$

$$\hat{\beta}_0 = ?$$

$$\hat{\beta}_1 = ?$$

$$\mathrm{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \qquad y = \begin{bmatrix} 9 \\ 12 \\ 17 \\ 21 \\ 26 \end{bmatrix}$$

$$X^T\mathrm{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 5 & 21 \\ 21 & 121 \end{bmatrix}$$

$$(X^T\mathrm{X})^{-1} = \frac{1}{164}\begin{bmatrix} 121 & -21 \\ -21 & 5 \end{bmatrix}$$

$$X^T\mathrm{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} 9 \\ 12 \\ 17 \\ 21 \\ 26 \end{bmatrix} = \begin{bmatrix} 85 \\ 435 \end{bmatrix}$$

$$\hat{\beta} = (X^T\mathrm{X})^{-1}X^T\mathrm{y}$$

$$= \frac{1}{164}\begin{bmatrix} 121 & -21 \\ -21 & 5 \end{bmatrix}\begin{bmatrix} 85 \\ 435 \end{bmatrix}$$

$$= \begin{bmatrix} 7.012 \\ 2.378 \end{bmatrix}$$

$\hat{\beta}_0$ = 7.012

$\hat{\beta}_1$ = 2.378

Step 2: Predict positions

- Plug time values into linear regression equation

(Position) = 2.378 (Time) + 7.012

- Predicted value at time = 12 secs

Position = 2.378 * 12 + 7.012 = 35.548

- Matrix form to predict all other positions

$$\hat{y} = X\hat{\beta}$$

Step 2: Predict positions

- Plug time values into linear regression equation

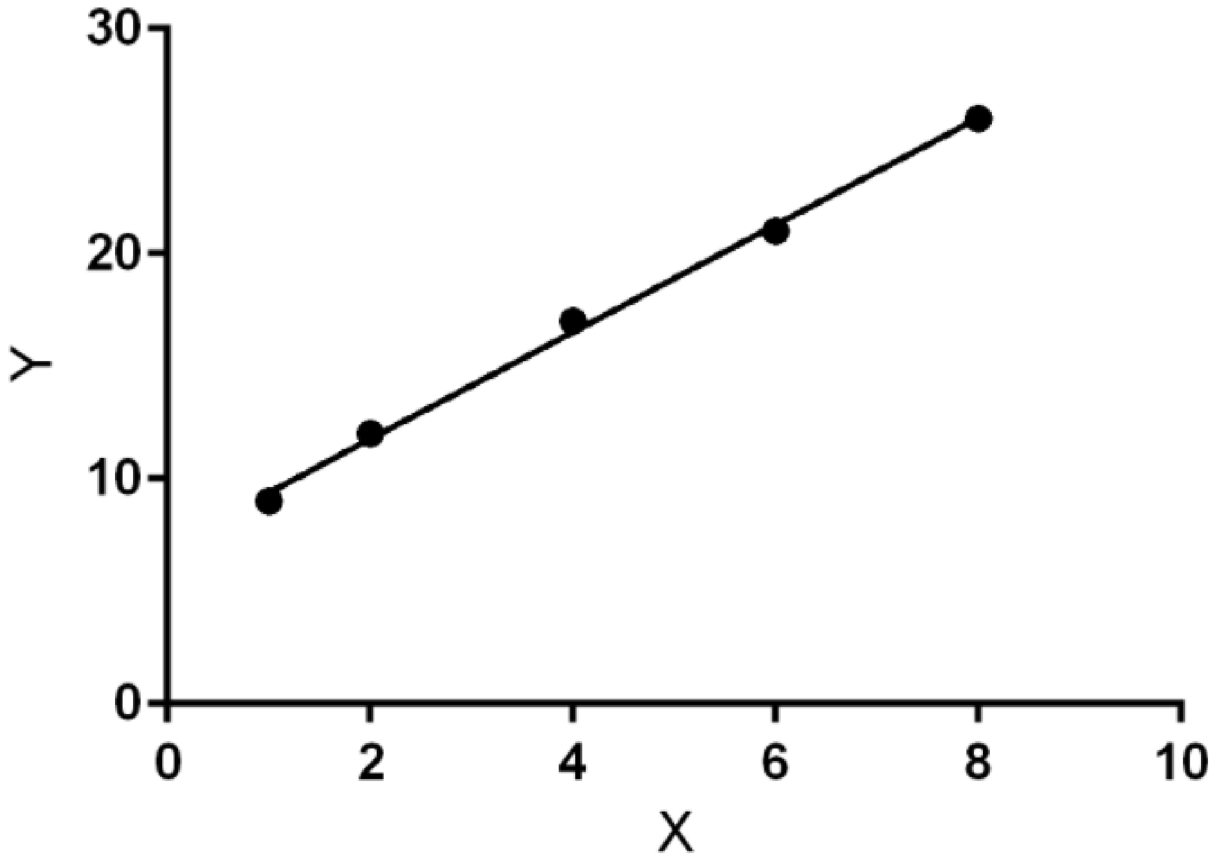
(Position) = 2.378 (Time) + 7.012

- Matrix form to predict all other positions

$$\hat{y} = X\hat{\beta}$$

$$\hat{y} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 7.012 \\ 2.378 \end{bmatrix} = \begin{bmatrix} 9.39 \\ 11.768 \\ 16.524 \\ 21.28 \\ 26.036 \end{bmatrix}$$

Plot



Z Score Normalization

- Why normalize features?
  - Different feature ranges such as [-1, 1] and [-100, 100] may negatively affect algorithm performance
  - Small change in bigger range can affect more than huge change in smaller range
- Z Score (Standard Score)

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

  - z_{ij} is the standard score for feature *j* of data point *i*
  - x_{ij} is the value of feature *j* of data point *i*
  - μ_j and σ_j are mean and standard deviation of feature *j*

Galaxy	Distance (Mpc)	Velocity (km/sec)
Virgo	15	1600
Ursa Minor	200	15,000
Corona Borealis	290	24,000
Bootes	520	40,000

- Normalize feature 'Distance'
- Compute Mean
  - $\mu_{dist} = \frac{1}{N}\sum_{i=1}^{N} x_{i.dist} = ?$
- Computer Standard Deviation
  - $\sigma_{dist} = \sqrt{\frac{\sum_{i=1}^{N}(x_{i.dist} - \mu_{dist})^2}{N-1}} = ?$

Galaxy	Distance (Mpc)	Velocity (km/sec)
Virgo	15	1600
Ursa Minor	200	15,000
Corona Borealis	290	24,000
Bootes	520	40,000

- Normalize feature ‘Distance’
- Compute Mean
  - $\mu_{dist} = \frac{1}{N}\sum_{i=1}^{N} x_{i.dist} = \frac{15+200+290+520}{4} = 256.25$
- Computer Standard Deviation
  - $\sigma_{dist} = \sqrt{\frac{\sum_{i=1}^{N}(x_{i.dist} - \mu_{dist})^2}{N-1}} = ?$

Galaxy	Distance (Mpc)	Velocity (km/sec)
Virgo	15	1600
Ursa Minor	200	15,000
Corona Borealis	290	24,000
Bootes	520	40,000

- Normalize feature ‘Distance’
- Compute Mean
  - $\mu_{dist} = \frac{1}{N}\sum_{i=1}^{N} x_{i.dist} = \frac{15+200+290+520}{4} = 256.25$
- Computer Standard Deviation
  - $\sigma_{dist} = \sqrt{\frac{\sum_{i=1}^{N}(x_{i.dist} - \mu_{dist})^2}{N}} = \sqrt{\frac{(15-256.25)^2+(200-256.25)^2+(290-256.25)^2+(520-256.25)^2}{4}} = 181.7063221$

Galaxy	Distance (Mpc)	Velocity (km/sec)
Virgo	15	1600
Ursa Minor	200	15,000
Corona Borealis	290	24,000
Bootes	520	40,000

- $\mu_{dist} = 256.25$
- $\sigma_{dist} = 181.7063221$
- Compute standard scores
  - $z_{virgo.dist} = \frac{x_{virgo.dist} - \mu_{dist}}{\sigma_{dist}} = ?$
  - $z_{ursa.dist} = \frac{x_{ursa.dist} - \mu_{dist}}{\sigma_{dist}} = ?$
  - $z_{corona.dist} = \frac{x_{corona.dist} - \mu_{dist}}{\sigma_{dist}} = ?$
  - $z_{bootes.dist} = \frac{x_{bootes.dist} - \mu_{dist}}{\sigma_{dist}} = ?$
- Similarly, other features like velocity can be standardized

Galaxy	Distance (Mpc)	Velocity (km/sec)
Virgo	15	1600
Ursa Minor	200	15,000
Corona Borealis	290	24,000
Bootes	520	40,000

- $\mu_{dist} = 256.25$
- $\sigma_{dist} = 181.7063221$
- Compute standard scores
  - $z_{virgo.dist} = \frac{x_{virgo.dist} - \mu_{dist}}{\sigma_{dist}} = \frac{15 - 256.25}{181.706} = -1.328$
  - $z_{ursa.dist} = \frac{x_{ursa.dist} - \mu_{dist}}{\sigma_{dist}} = \frac{200 - 256.25}{181.706} = -0.3096$
  - $z_{corona.dist} = \frac{x_{corona.dist} - \mu_{dist}}{\sigma_{dist}} = \frac{290 - 256.25}{181.706} = 0.186$
  - $z_{bootes.dist} = \frac{x_{bootes.dist} - \mu_{dist}}{\sigma_{dist}} = \frac{520 - 256.25}{181.706} = 1.452$
- Similarly, other features like velocity can be standardized

Multidimensional Newton's Method

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 L(\beta)}{\partial\beta\partial\beta^T}\right)^{-1} \frac{\partial L(\beta)}{\partial\beta}$$

- $x^{(0)} = [3, -1, 0]$
- $f(x_1, x_2, x_3) = (x_1 + 10x_2)^2 + 5(x_1 - x_3)^2 + (x_2 - 2x_3)^4$
- What is $f(x^{(0)})$?

Multidimensional Newton's Method

- $x^{(0)} = [3, -1, 0]$
- $f(x_1, x_2, x_3) = (x_1 + 10x_2)^2 + 5(x_1 - x_3)^2 + (x_2 - 2x_3)^4$
- What is $f(x^{(0)})$?
  - $(3 + 10 \cdot -1)^2 + 5 \cdot (3\ - 0)^2 + (-1\ - 2 \cdot 0)^4 = 95$

Multidimensional Newton's Method

- $x^{(0)} = [3, -1, 0]$
- $f(x_1, x_2, x_3) = (x_1 + 10x_2)^2 + 5(x_1 - x_3)^2 + (x_2 - 2x_3)^4$
- What is $f(x^{(0)})$?
  - $(3 + 10 \cdot -1)^2 + 5 \cdot (3\ - 0)^2 + (-1\ - 2 \cdot 0)^4 = 95$
- What is $\nabla f(x)$?

Multidimensional Newton's Method

- $x^{(0)} = [3, -1, 0]$
- $f(x_1, x_2, x_3) = (x_1 + 10x_2)^2 + 5(x_1 - x_3)^2 + (x_2 - 2x_3)^4$
- What is $f(x^{(0)})$?
  - $(3 + 10 \cdot -1)^2 + 5 \cdot (3 \ - 0)^2 + (-1 \ - 2 \cdot 0)^4 = 95$
- What is $\nabla f(x)$?
  - $\nabla f(x_1) = 2 \cdot (x_1 + 10 \cdot x_2) + 10 \cdot (x_1 - x_3)$
  - $\nabla f(x_2) = 20 \cdot (x_1 + 10 \cdot x_2) + 4 \cdot (x_2 - 2 \cdot x_3)^3$
  - $\nabla f(x_2) = -10 \cdot (x_1 - x_3) + -8 \cdot (x_2 - 2 \cdot x_3)^3$

Multidimensional Newton’s Method

- What is $\nabla f(x)$?
  - $\nabla f(x_1) = 2 \cdot (x_1 + 10 \cdot x_2) + 10 \cdot (x_1 - x_3)$
  - $\nabla f(x_2) = 20 \cdot (x_1 + 10 \cdot x_2) + 4 \cdot (x_2 - 2 \cdot x_3)^3$
  - $\nabla f(x_2) = -10 \cdot (x_1 - x_3) + -8 \cdot (x_2 - 2 \cdot x_3)^3$
- What is F(x)?

Multidimensional Newton's Method

- What is $\nabla f(x)$?
  - $\nabla f(x_1) = 2 \cdot (x_1 + 10 \cdot x_2) + 10 \cdot (x_1 - x_3)$
  - $\nabla f(x_2) = 20 \cdot (x_1 + 10 \cdot x_2) + 4 \cdot (x_2 - 2 \cdot x_3)^3$
  - $\nabla f(x_2) = -10 \cdot (x_1 - x_3) + -8 \cdot (x_2 - 2 \cdot x_3)^3$
- What is F(x)?
  - $\begin{bmatrix} 12 & 20 & -10 \\ 20 & 10 + 12 \cdot (x_2 - 2 \cdot x_3)^2 & -24 \cdot (x_2 - 2 \cdot x_3)^2 \\ -10 & -24 \cdot (x_2 - 2 \cdot x_3)^2 & 48 \cdot (x_2 - 2 \cdot x_3)^2 \end{bmatrix}$

Multidimensional Newton’s Method

- What is $F(x)$?
  - $\begin{bmatrix} 12 & 20 & -10 \\ 20 & 10 + 12 \cdot (x_2 - 2 \cdot x_3)^2 & -24 \cdot (x_2 - 2 \cdot x_3)^2 \\ -10 & -24 \cdot (x_2 - 2 \cdot x_3)^2 & 48 \cdot (x_2 - 2 \cdot x_3)^2 \end{bmatrix}$
- What is $\nabla f\left(x^{(0)}\right)$?

Multidimensional Newton's Method

- $x^{(0)} = [3, -1, 0]$
- What is $\nabla f(x)$?
  - $\nabla f(x_1) = 2 \cdot (x_1 + 10 \cdot x_2) + 10 \cdot (x_1 - x_3)$
  - $\nabla f(x_2) = 20 \cdot (x_1 + 10 \cdot x_2) + 4 \cdot (x_2 - 2 \cdot x_3)^3$
  - $\nabla f(x_2) = -10 \cdot (x_1 - x_3) + -8 \cdot (x_2 - 2 \cdot x_3)^3$
- What is $\nabla f\left(x^{(0)}\right)$?

Multidimensional Newton’s Method

- $x^{(0)} = [3, -1, 0]$
- What is $\nabla f(x)$?
  - $\nabla f(x_1) = 2 \cdot (x_1 + 10 \cdot x_2) + 10 \cdot (x_1 - x_3)$
  - $\nabla f(x_2) = 20 \cdot (x_1 + 10 \cdot x_2) + 4 \cdot (x_2 - 2 \cdot x_3)^3$
  - $\nabla f(x_2) = -10 \cdot (x_1 - x_3) + -8 \cdot (x_2 - 2 \cdot x_3)^3$
- What is $\nabla f\left(x^{(0)}\right)$?
  - $[16, -144, -22]$

Multidimensional Newton's Method

- What is $\nabla f\left(x^{(0)}\right)$?
  - $[16, -144, -22]$
- What is $F\left(x^{(0)}\right)$?

Multidimensional Newton's Method

- $x^{(0)} = [3, -1, 0]$
- What is $F(x)$?
  - $\begin{bmatrix} 12 & 20 & -10 \\ 20 & 10 + 12 \cdot (x_2 - 2 \cdot x_3)^2 & -24 \cdot (x_2 - 2 \cdot x_3)^2 \\ -10 & -24 \cdot (x_2 - 2 \cdot x_3)^2 & 48 \cdot (x_2 - 2 \cdot x_3)^2 \end{bmatrix}$
- What is $F\left(x^{(0)}\right)$?

Multidimensional Newton's Method

- $x^{(0)} = [3, -1, 0]$
- What is $F(x)$?
  - $\begin{bmatrix} 12 & 20 & -10 \\ 20 & 10 + 12 \cdot (x_2 - 2 \cdot x_3)^2 & -24 \cdot (x_2 - 2 \cdot x_3)^2 \\ -10 & -24 \cdot (x_2 - 2 \cdot x_3)^2 & 48 \cdot (x_2 - 2 \cdot x_3)^2 \end{bmatrix}$
- What is $F\left(x^{(0)}\right)$?
  - $\begin{bmatrix} 12 & 20 & -10 \\ 20 & 22 & -24 \\ -10 & -24 & 48 \end{bmatrix}$

Multidimensional Newton's Method

- What is $F\left(x^{(0)}\right)^{-1}$?
  - $\begin{bmatrix} -0.079 & 0.119 & 0.043 \\ 0.1192 & -0.0788 & -0.0145 \\ -0.043 & -0.0145 & 0.0225 \end{bmatrix}$
- What is $\nabla f\left(x^{(0)}\right)$?
  - $[16, -144, -22]$
- What is $F\left(x^{(0)}\right)^{-1} \cdot \nabla f\left(x^{(0)}\right)$?

Multidimensional Newton's Method

- What is $F\left(x^{(0)}\right)^{-1}$?
  - $\begin{bmatrix} -0.079 & 0.119 & 0.043 \\ 0.1192 & -0.0788 & -0.0145 \\ -0.043 & -0.0145 & 0.0225 \end{bmatrix}$
- What is $\nabla f\left(x^{(0)}\right)$?
  - $[16, -144, -22]$
- What is $F\left(x^{(0)}\right)^{-1} \cdot \nabla f\left(x^{(0)}\right)$?
  - $[-19.384, 13.576, 2.291]$

Multidimensional Newton's Method

1. Guess $x^{(0)}$
2. Get $\nabla f(x)$
3. Get F(x)
4. $n = 0$
5. Calculate $\nabla f\left(x^{(n)}\right)$
6. Calculate $F\left(x^{(n)}\right)$
7. Calculate $F\left(x^{(n)}\right)^{-1}$
8. $x^{(n+1)} = x^{(n)} - F\left(x^{(n)}\right)^{-1} \cdot \nabla f\left(x^{(n)}\right)$
9. n = n + 1

Multidimensional Newton’s Method

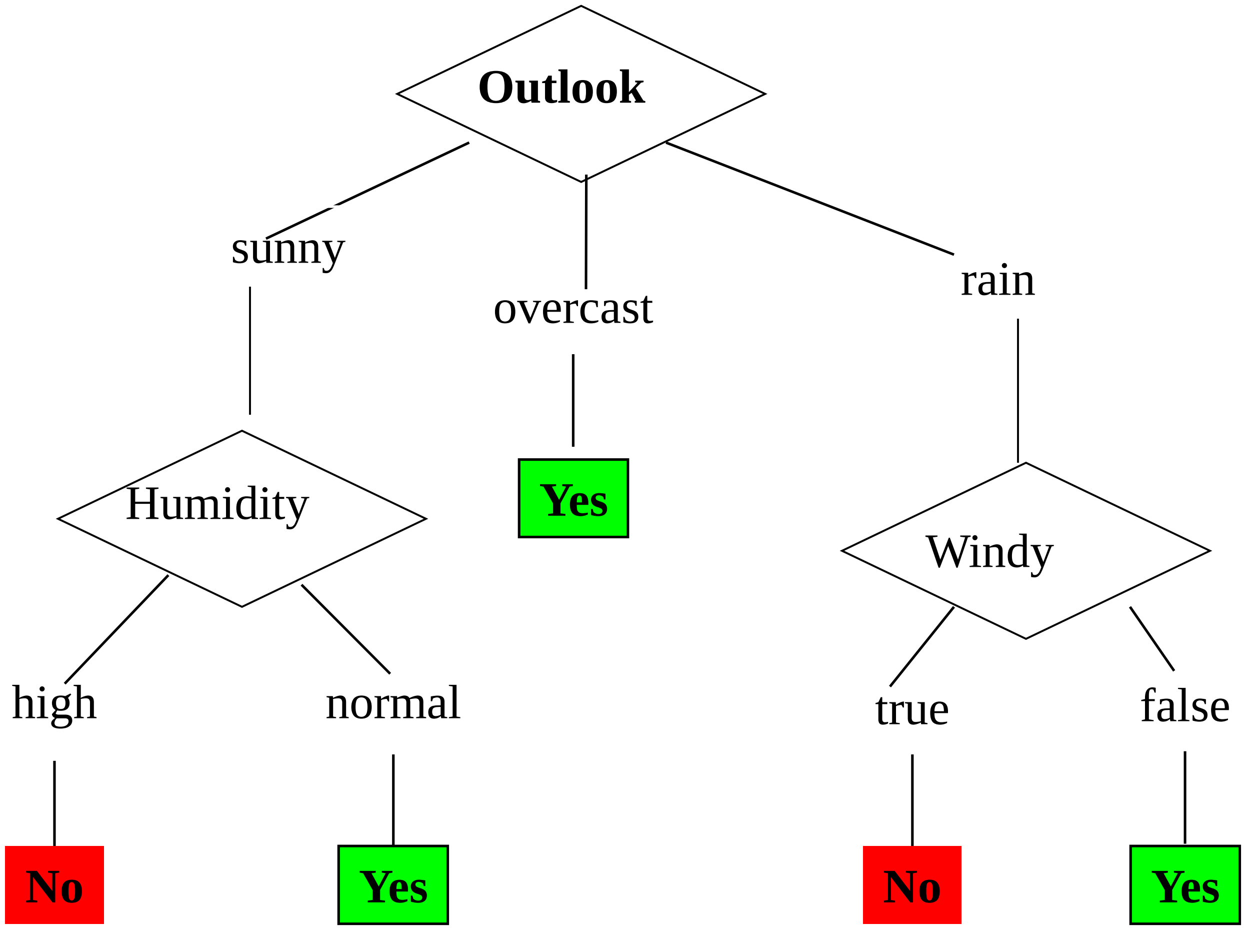
1. Guess $x^{(0)}$
2. Get $\nabla f(x)$
3. Get F(x)
4. $n = 0$
5. Calculate $\nabla f\left(x^{(n)}\right)$
6. Calculate $F\left(x^{(n)}\right)$
7. Calculate $F\left(x^{(n)}\right)^{-1}$
8. $x^{(n+1)} = x^{(n)} - F\left(x^{(n)}\right)^{-1} \cdot \nabla f\left(x^{(n)}\right)$
9. n = n + 1

Weather Data: Play or not Play?

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Note:
Outlook is the Forecast,
no relation to Microsoft email program

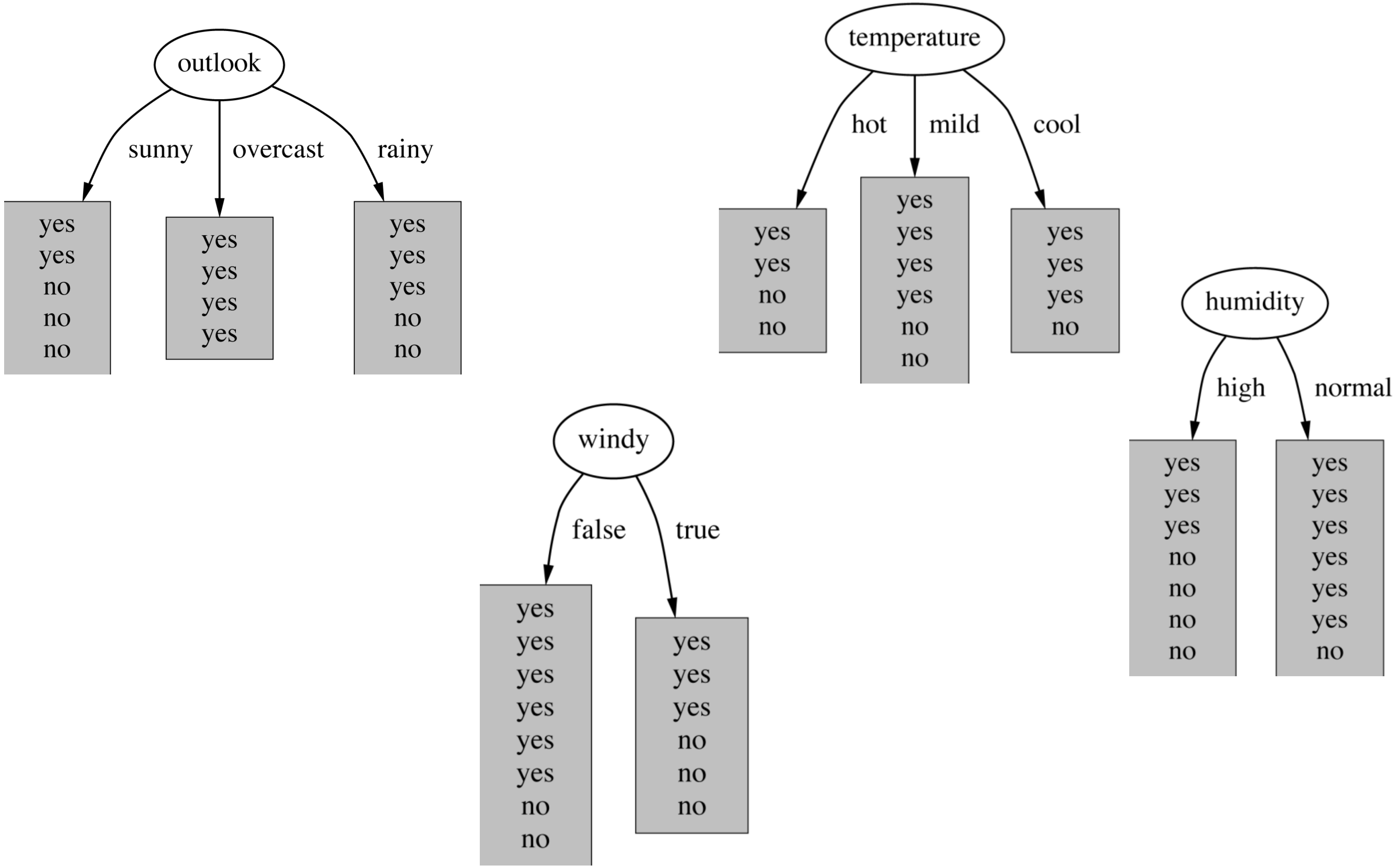
Example Tree for “Play?”



Choosing the Splitting Attribute

- At each node, available attributes are evaluated on the basis of separating the classes of the training examples. A Goodness function is used for this purpose.
- Typical goodness functions:
  - information gain (ID3/C4.5)
  - information gain ratio
  - gini index

Which attribute to select?



A criterion for attribute selection

- Which is the best attribute?
  - The one which will result in the smallest tree
  - Heuristic: choose the attribute that produces the "purest" nodes
- Popular *impurity criterion*: *information gain*
  - Information gain increases with the average purity of the subsets that an attribute produces
- Strategy: choose attribute that results in greatest information gain

Entropy of a split

- Information in a split with ***x*** items of one class, ***y*** items of the second class

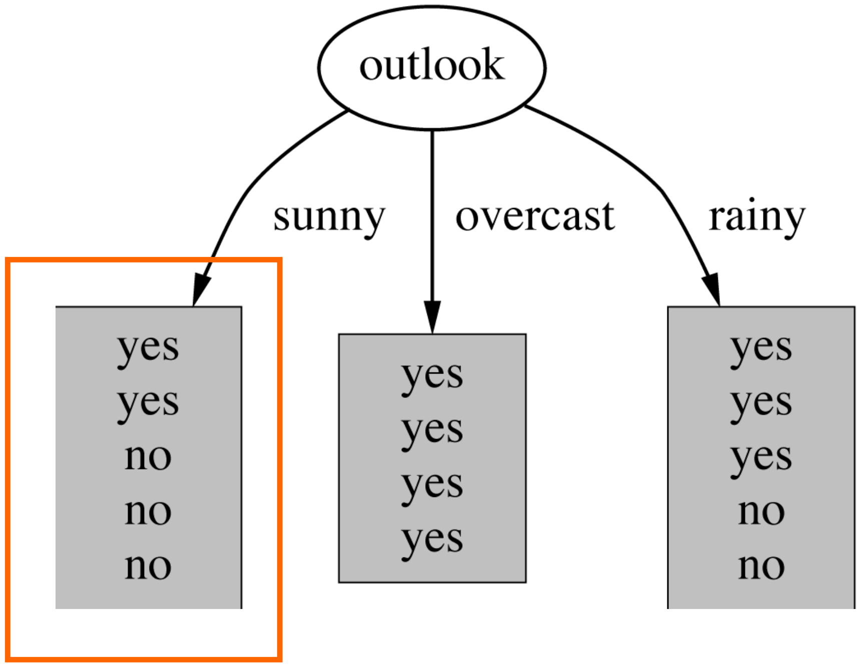
$$\mathrm{info}([\mathrm{x},\ \mathrm{y}]) = \mathrm{entropy}\left(\frac{\mathrm{x}}{\mathrm{x}+\mathrm{y}}, \frac{\mathrm{y}}{\mathrm{x}+\mathrm{y}}\right)$$

$$= -\frac{x}{x+y}\log\left(\frac{x}{x+y}\right) - \frac{y}{x+y}\log\left(\frac{y}{x+y}\right)$$

Example: attribute "Outlook"

- "Outlook" = "Sunny": 2 and 3 split

$$\text{info}([2,3]) = \text{entropy}(2/5,3/5) = -\frac{2}{5}\log(\frac{2}{5}) - \frac{3}{5}\log(\frac{3}{5}) = 0.971$$

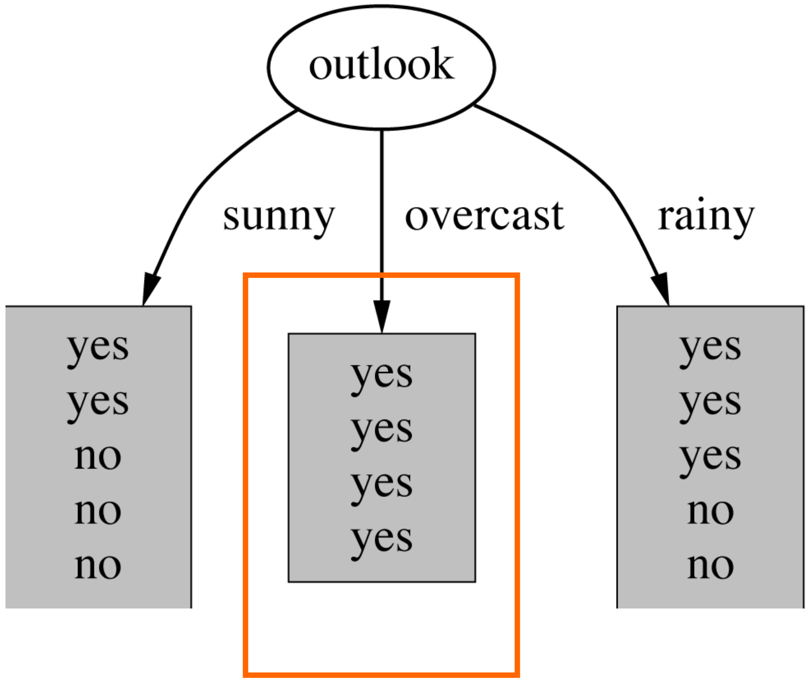


Outlook = Overcast

- "Outlook" = "Overcast": 4/0 split

$$\text{info}([4,0]) = \text{entropy}(1,0) = -1\log(1) - 0\log(0) = 0$$

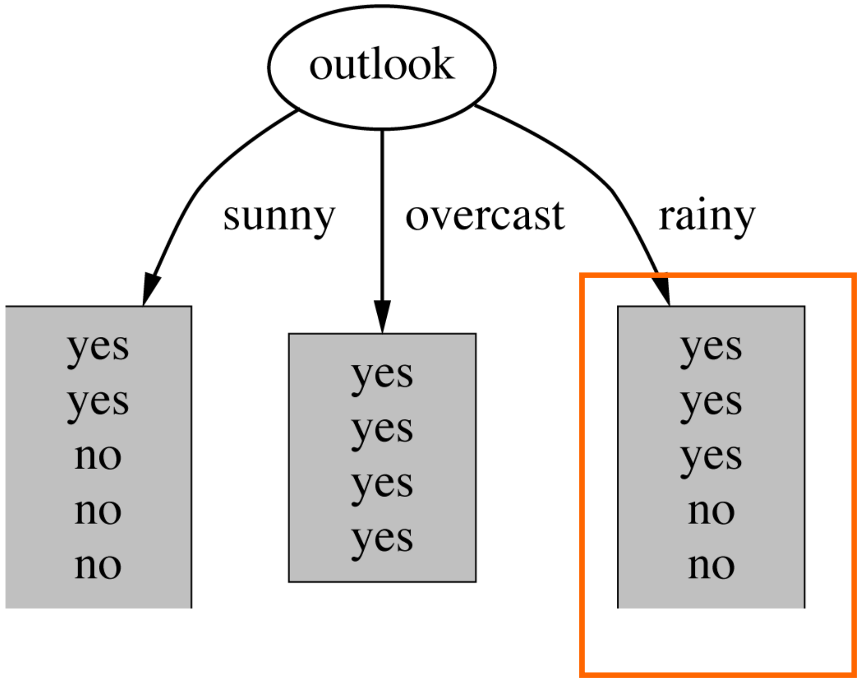
Note: log(0) is not defined, but we evaluate 0*log(0) as zero



Outlook = Rainy

- “Outlook” = “Rainy”:

$$\text{info}([3,2]) = \text{entropy}(3/5, 2/5) = -\frac{3}{5}\log(\frac{3}{5}) - \frac{2}{5}\log(\frac{2}{5}) = 0.971$$



Expected Information

Expected information for attribute:

$$\text{info}([3,2],[4,0],[3,2]) = (5/14)\times 0.971 + (4/14)\times 0 + (5/14)\times 0.971$$

$$= 0.693$$

Computing the information gain

- Information gain:

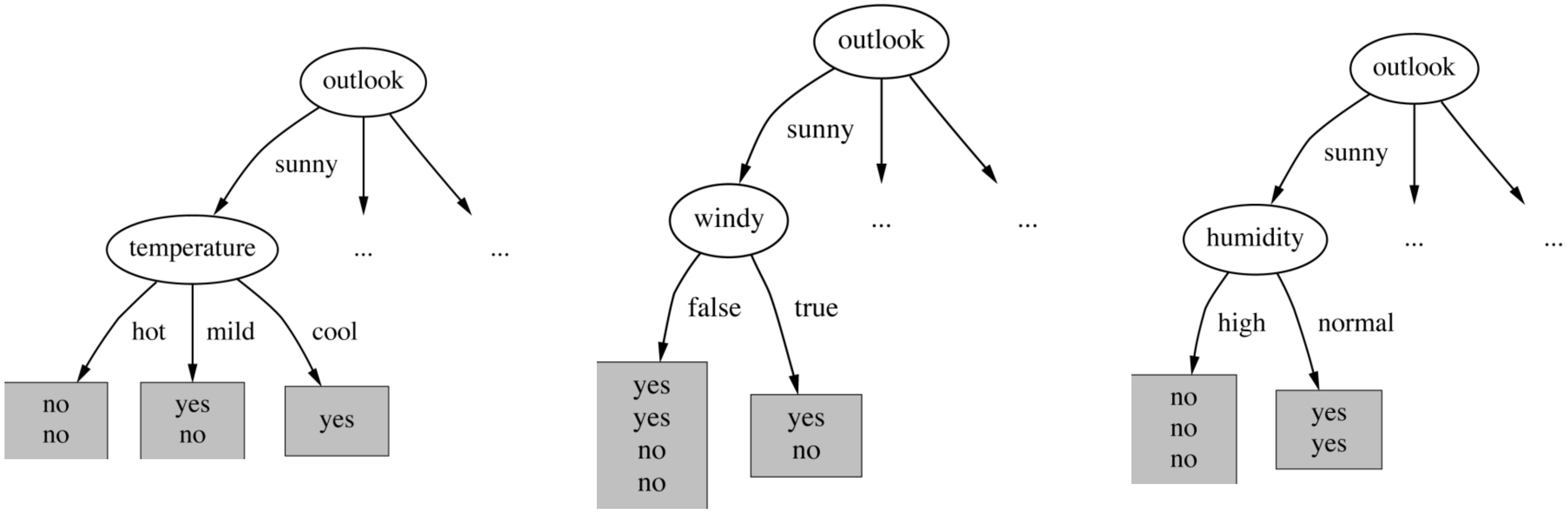
(information before split) – (information after split)

$$\text{gain}(\text{"Outlook"}) = \text{info}([9,5]) - \text{info}([2,3],[4,0],[3,2]) = 0.940 - 0.693$$
$$= 0.247$$

- Information gain for attributes from weather data:

$$\text{gain}(\text{"Outlook"}) = 0.247$$
$$\text{gain}(\text{"Temperature"}) = 0.029$$
$$\text{gain}(\text{"Humidity"}) = 0.152$$
$$\text{gain}(\text{"Windy"}) = 0.048$$

Continuing to split

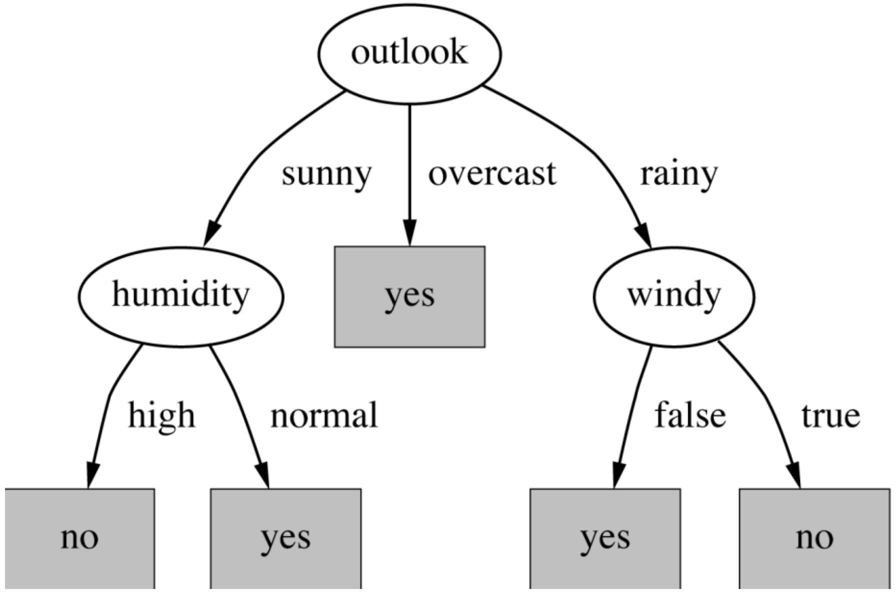


$$\text{gain}(\text{"Humidity"}) = 0.971$$

$$\text{gain}(\text{"Temperature"}) = 0.571$$

$$\text{gain}(\text{"Windy"}) = 0.020$$

The final decision tree



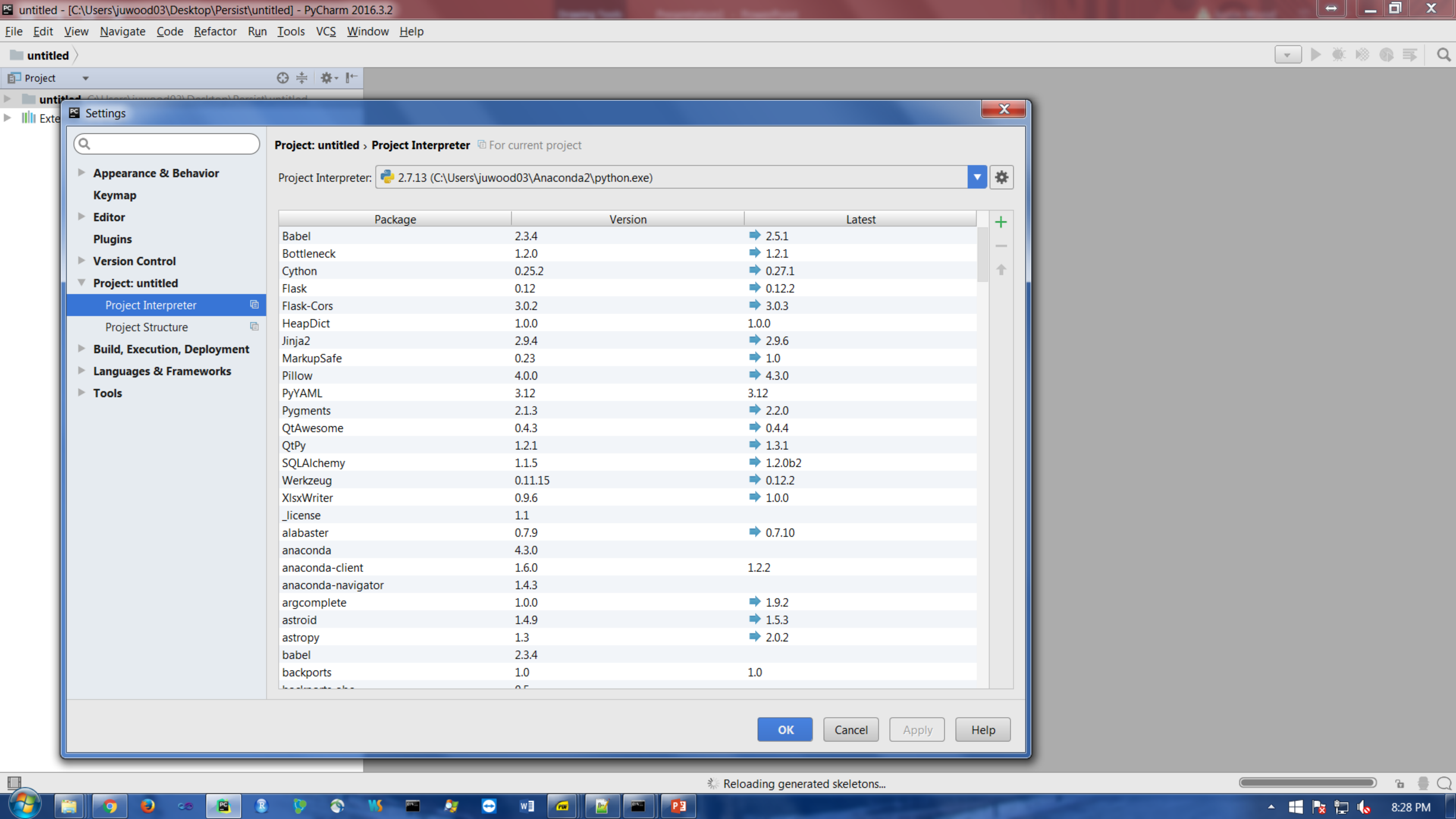
- Note: not all leaves need to be pure; sometimes identical instances have different classes

  $\Rightarrow$ Splitting stops when data can't be split any further

Twitter API

python

- Get python (Anaconda recommended)
  - https://www.anaconda.com/
- Get an IDE (PyCharm)
  - https://www.jetbrains.com/pycharm/
- Set PyCharm interpreter to Anaconda
  - File → Settings → Project: <name> → Python Interpreter



untitled - [C:\Users\juwood03\Desktop\Persist\untitled] - PyCharm 2016.3.2
File Edit View Navigate Code Refactor Run Tools VCS Window Help
untitled
Project
Settings
Appearance & Behavior
Keymap
Editor
Plugins
Version Control
Project: untitled
Project Interpreter
Project Structure
Build, Execution, Deployment
Languages & Frameworks
Tools
Project: untitled › Project Interpreter
For current project
Project Interpreter: 2.7.13 (C:\Users\juwood03\Anaconda2\python.exe)
Package Version Latest
Babel 2.3.4 2.5.1
Bottleneck 1.2.0 1.2.1
Cython 0.25.2 0.27.1
Flask 0.12 0.12.2
Flask-Cors 3.0.2 3.0.3
HeapDict 1.0.0 1.0.0
Jinja2 2.9.4 2.9.6
MarkupSafe 0.23 1.0
Pillow 4.0.0 4.3.0
PyYAML 3.12 3.12
Pygments 2.1.3 2.2.0
QtAwesome 0.4.3 0.4.4
QtPy 1.2.1 1.3.1
SQLAlchemy 1.1.5 1.2.0b2
Werkzeug 0.11.15 0.12.2
XlsxWriter 0.9.6 1.0.0
_license 1.1
alabaster 0.7.9 0.7.10
anaconda 4.3.0
anaconda-client 1.6.0 1.2.2
anaconda-navigator 1.4.3
argcomplete 1.0.0 1.9.2
astroid 1.4.9 1.5.3
astropy 1.3 2.0.2
babel 2.3.4
backports 1.0 1.0
OK
Cancel
Apply
Help
Reloading generated skeletons...
8:28 PM

python

- Get python (Anaconda recommended)
  - https://www.anaconda.com/
- Get an IDE (PyCharm)
  - https://www.jetbrains.com/pycharm/
- Set PyCharm interpreter to Anaconda
  - File → Settings → Project: <name> → Python Interpreter
- Make sure command line python and pip are pointing to Anaconda

C:\Windows\system32\cmd.exe

```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\juwood03>where python
C:\Users\juwood03\Anaconda2\python.exe

C:\Users\juwood03>where pip
C:\Users\juwood03\Anaconda2\Scripts\pip.exe

C:\Users\juwood03>
```

Slide 4 of 4 Notes 88% 8:33 PM

python-twitter

- pip install tweepy

Twitter

- Sign-up
  - https://twitter.com/signup (must add phone number)
- Register an app
  - https://apps.twitter.com/ → Create New App

https://apps.twitter.com/app/new

Application Management

Create an application

Application Details

Name *

stock-prediction

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

Predicting stocks

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

https://github.com/juwood03/stock_prediction

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

http://127.0.0.1:8080

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

Twitter

- Sign-up
  - https://twitter.com/signup
- Register an app
  - https://apps.twitter.com/ → Create New App
- Get the keys and access tokens
  - Keys and Access Tokens tab → Create my access token

Piazza | Course | Re: Qu | Re: 14 | CS145 | what is | piazza | Week | Draft - | Assign | Mining | Gettin | ucla-st | Twitte

Secure | https://apps.twitter.com/app/14326148/keys

PARSE | (43) UCLAIDEALab - | UFLDL Tutorial - Ufldl | juwood03 | command line - How | How to Write a Thesis | web.cs.ucla.edu/~yzs | probability-rules.dvi | CS145 Fall 2017 - Go | CS 145

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	a8cP5hihOsL74p0K3yWTW1Wtn
Consumer Secret (API Secret)	yhNYrOPHQ86XP0e7lG8e8kYbo8KCwSLveo6e68slxp2GXlsx1H
Access Level	Read and write (modify app permissions)
Owner	juwood04
Owner ID	917234971072270337

Application Actions

Regenerate Consumer Key and Secret | Change App Permissions

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	917234971072270337-BdTFplE5lPpLSpQZJm3X1sQ6ylIT8CT
Access Token Secret	S9yQCQrs2q06gSeikhGQZEaC4lKS1Vd4K8biYhKPCvhvj
Access Level	Read and write
Owner	juwood04
Owner ID	917234971072270337

Homework_2016.zip | Show all

untitled - [C:\Users\juwood03\Desktop\Persist\untitled] - ...\streaming.py - PyCharm 2016.3.2

File Edit View Navigate Code Refactor Run Tools VCS Window Help

untitled 〉 streaming.py 〉 streaming

streaming.py ×

```python
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

consumer_key="a8cP5hihOsL74p0K3yWTW1Wtn"
consumer_secret="yhNYrOPHQ86XP0e71G8e8kYbo8KCwSLveo6e68sIxp2GXIsx1H"
access_token="917234971072270337-BdTFplE5lPpLSpQZJm3X1sQ6yllT8CT"
access_token_secret="S9yQCQrs2q06gSeikhGQZEaC4lKS1Vd4K8biYhKPCvhvj"

class StdOutListener(StreamListener):
    def on_data(self, data):
        print(data)
        return True

    def on_error(self, status):
        print(status)

if __name__ == '__main__':
    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
```

Run streaming

```
C:\Users\juwood03\Anaconda2\python.exe C:/Users/juwood03/Desktop/Persist/untitled/streaming.py
```

{"created_at":"Fri Oct 13 07:16:50 +0000 2017","id":918737277185953792,"id_str":"918737277185953792","text":"RT @sinndarDerby: \u300e\u30e9\u30a4\u30aa\u30f3\u30ba\u3067\u6b7b\u306e\u3

{"created_at":"Fri Oct 13 07:16:53 +0000 2017","id":918737286535221248,"id_str":"918737286535221248","text":"@KNBSB bij @WBSC extraordinary congres. Belangrijkste punt de wijziging var

```
Process finished with exit code 1
```

Platform and Plugin Updates: PyCharm is ready to update. (yesterday 9:57 AM)   8:1 CRLF UTF-8

untitled - [C:\Users\juwood03\Desktop\Persist\untitled] - ...\search.py - PyCharm 2016.3.2

File Edit View Navigate Code Refactor Run Tools VCS Window Help

untitled 〉 search.py 〉

search.py ×

```python
import tweepy

consumer_key="a8cP5hihOsL74p0K3yWTW1Wtn"
consumer_secret="yhNYrOPHQ86XP0e71G8e8kYbo8KCwSLveo6e68sIxp2GXIsx1H"
access_token="917234971072270337-BdTFplE51PpLSpQZJm3X1sQ6yllT8CT"
access_token_secret="S9yQCQrs2q06gSeikhGQZEaC41KS1Vd4K8biYhKPCvhvj"

if __name__ == '__main__':
    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    api = tweepy.API(auth)
    results = api.search(q="cubs")
    for r in results:
        print("----------------------------------------------------")
        print(r.text)
    print("----------------------------------------------------")
```

Run search

```
----------------------------------------------------
RT @Cubs: What a game.

What a series. #FlyTheW https://t.co/xvo8IzpLpJ
----------------------------------------------------
RT @mas_yusak: @JackOHara07 @Cubs No hair flip no bat flip and certainly no LA trip
----------------------------------------------------
RT @Nationals: Congrats on a hard-fought series, @Cubs. https://t.co/YhZDkMI5v2
----------------------------------------------------
RT @DodgersNation: What should the Cubbies expect in the Ravine this Weekend? https://t.co/1XQXiiuvMu cc: @Andre_CA7 https://t.co/rUCMoRCzGv
----------------------------------------------------
RT @Cubs: In the face! #FlyTheW https://t.co/O54SWe6PT7
----------------------------------------------------
Big's Updates: Did You See This? Cubs' challenge of overturned tag play pays off https://t.co/nSuzq2KcMF
----------------------------------------------------

Process finished with exit code 0
```

Platform and Plugin Updates: PyCharm is ready to update. (yesterday 9:57 AM) 37:1 CRLF UTF-8

untitled - [C:\Users\juwood03\Desktop\Persist\untitled] - ...\location.py - PyCharm 2016.3.2

File Edit View Navigate Code Refactor Run Tools VCS Window Help

untitled 〉 location.py 〉

location.py ×

```python
import tweepy

consumer_key="a8cP5hihOsL74p0K3yWTW1Wtn"
consumer_secret="yhNYrOPHQ86XP0e7lG8e8kYbo8KCwSLveo6e68sIxp2GXIsx1H"
access_token="917234971072270337-BdTFplE5lPpLSpQZJm3X1sQ6yllT8CT"
access_token_secret="S9yQCQrs2q06gSeikhGQZEaC4lKS1Vd4K8biYhKPCvhvj"

if __name__ == '__main__':
    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    api = tweepy.API(auth)
    results = api.search(q="james",geocode="41.489381,-81.667486,10mi")
    for r in results:
        print("-----------------------------------------------------")
        print(r.text)
    print("-----------------------------------------------------")
    results = api.search(q="james")
    for r in results:
        print("*************************************************")
        print(r.text)
    print("*************************************************")
```

Run location

```
************************************************
RT @Iennonmccartney: "Cause we are a family and they're just a team." - James McAvoy on x-men vs the avengers https://t.co/TTE6zPRIXB
************************************************
RT @alexandermcdizz: Is there an echo in here? Mind your own business..again! &amp; to the ppl wishing cancer on James, hope yall dont choke on…
************************************************
RT @meghan_raee: The cancer won in the last round. I'm sorry this happened to you momma. I'm so happy you're with the Lord. I love you w al…
************************************************
RT @jamesaiken09: Skewed Cubes Abstract - photograph by James Aiken
https://t.co/Z1jieYzkVi
@jamesaiken09 #urbanmyopia #minimalism #abstr…
************************************************
RT @Mafe_Navarro: El misterioso poder de una madre. - James VDC https://t.co/CvcKwibQJk
************************************************
RT @alexandermcdizz: Is there an echo in here? Mind your own business..again! &amp; to the ppl wishing cancer on James, hope yall dont choke on…
************************************************

Process finished with exit code 0
```

Platform and Plugin Updates: PyCharm is ready to update. (yesterday 9:57 AM)   70:1 CRLF UTF-8

Twitter

- Rate limits
  - Searching
    - 24 hours x 4 15-minute increments x 450 requests per 15-minute increments = 43,200 requests per day
  - Streaming
    - $\approx$ 1%?

Likelihood

- Is likelihood a density or probability?

Likelihood

- Is likelihood a density or probability?
  - No, it is the multiplication of densities

Likelihood

- Is likelihood a density or probability?
  - No, it is the multiplication of densities
- Densities often < 1
  - Multiplication approaches epsilon (smallest non-zero positive value any language can handle) exponentially
- Likelihood used in gradient ascent
  - if complex function partial derivative can get messy

Likelihood

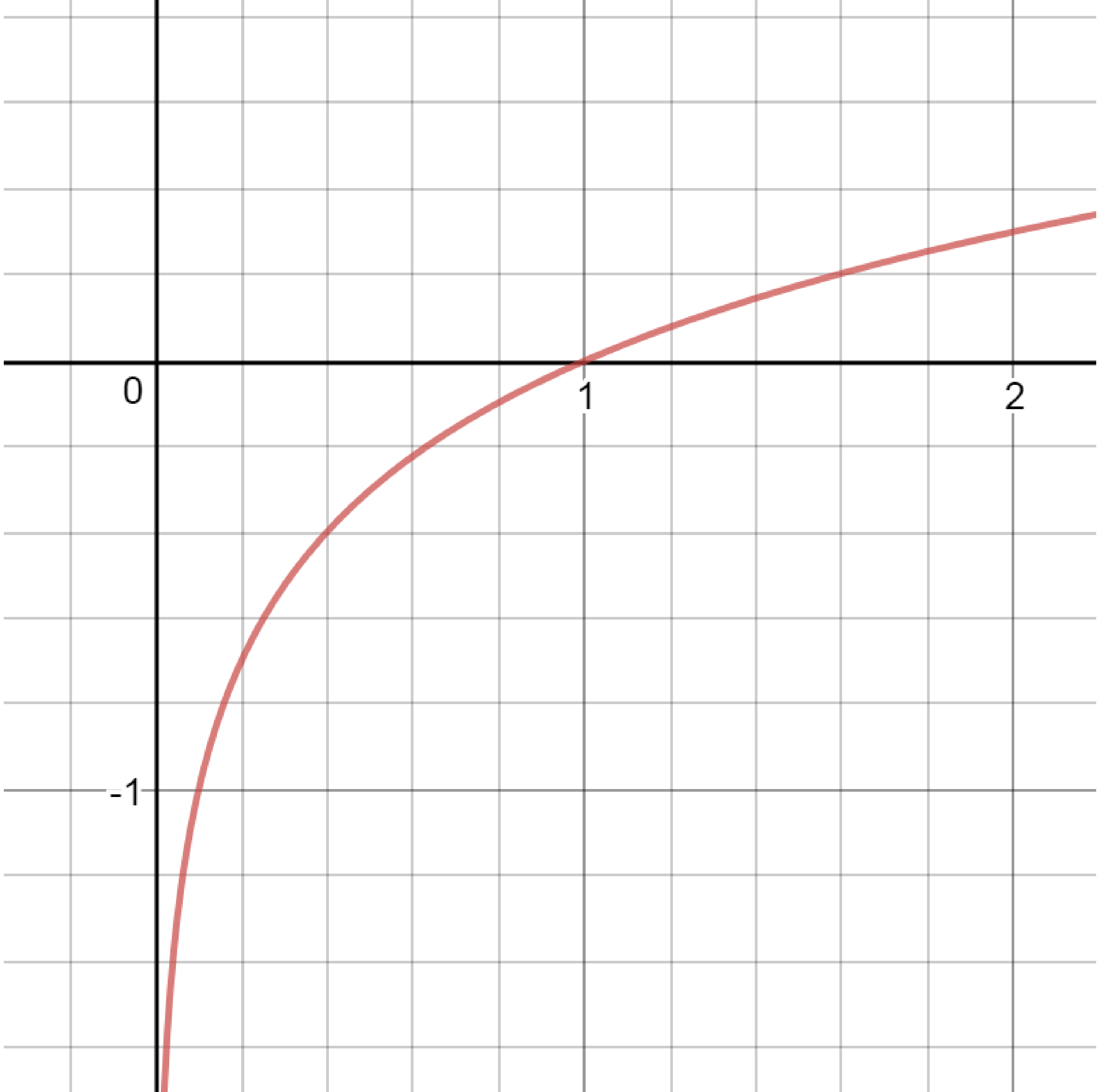
- Solution?

Likelihood

- Solution?
  - Take the log

Likelihood

- Solution?
  - Take the log
- $\log(x \cdot y) = \log(x) + \log(y)$
  - Approaches $\pm\infty$ linearly
  - Easier to take derivative
- Density > 1
  - Log-likelihood > 1
- Density < 1
  - Log-likelihood < 0


0
1
2
-1