HUDK 4051: LEARNING ANALYTICS: PROCESS & THEORY

1/25/17 1:49 PM

Please Sign Up for Podcast

In the news

EDUCATION WEEK

Trump Now Leads the Education Dept.; What's the Impact on Data and ESSA?

Which Students Are Arrested the Most?

Education DIVE

Why are schools so behind on using data to make decisions?

LAUSD board: If Trump administration asks for student data, district will resist

89.3 KPCC

eCAMPUS NEWS

5 ways data analytics education is advancing

SmartDataCollective

The World's Best Thinkers on Data

Challenges of Big Data in Education

THE WORLD UNIVERSITY RANKINGS

UK government refuses to release 'uncomfortable' visa figures

EdSurge News

To Re-Capture the Education Market, Microsoft Aims to Offer a Compelling Alternative to Google's Chromebook

Arntzen: Montana in noncompliance with Department of Education

Great Falls Tribune

EdTech Focus On K-12

Visualizing Data Can Help Stop Cyberattacks, Identify Trends

Department of Education Creates New Evaluation Tool for K–12 Administrators

A simplified ed tech evaluation process is on the horizon.

The next seminar in the Learning Analytics Series:

Dan Davis - Delft University of Technology
Friday, January 27th
2:30 PM - 4:30 PM EST
Grace Dodge Hall | GDH 457
Teachers College – Columbia University

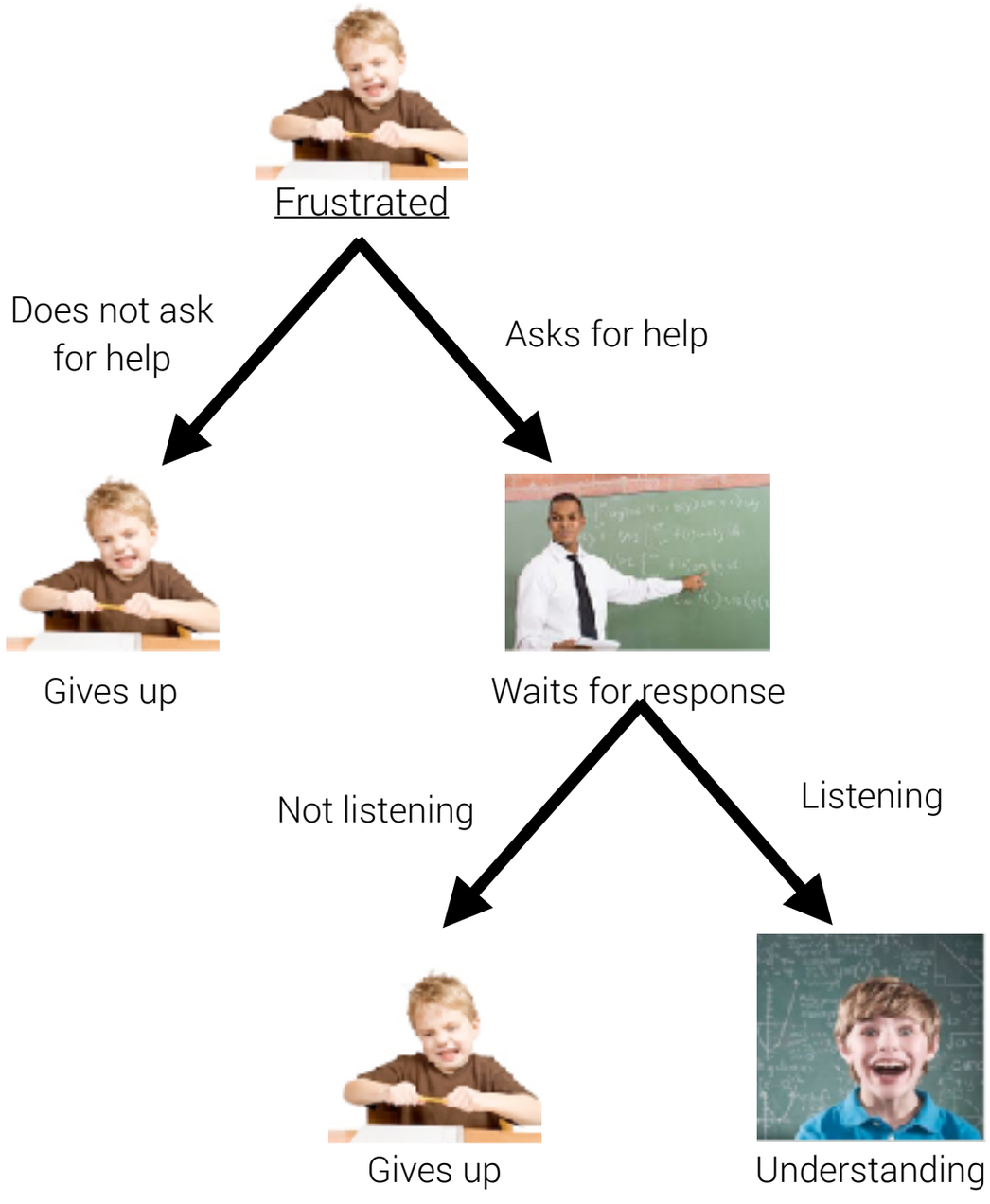
Social Comparison as a Means to Improve MOOC Completion Rates

R Documentation

Review CART Trees

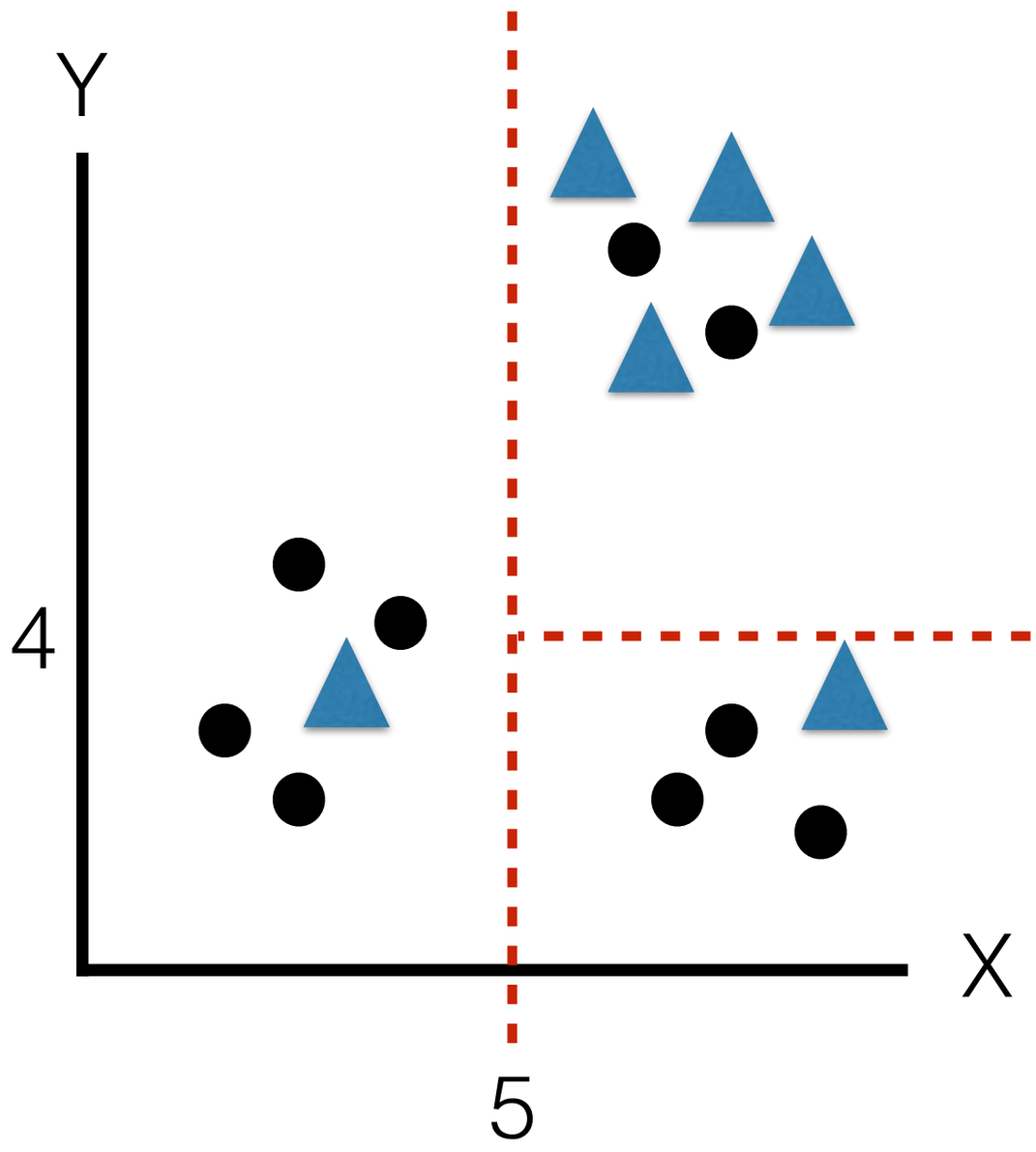
Classification Tree

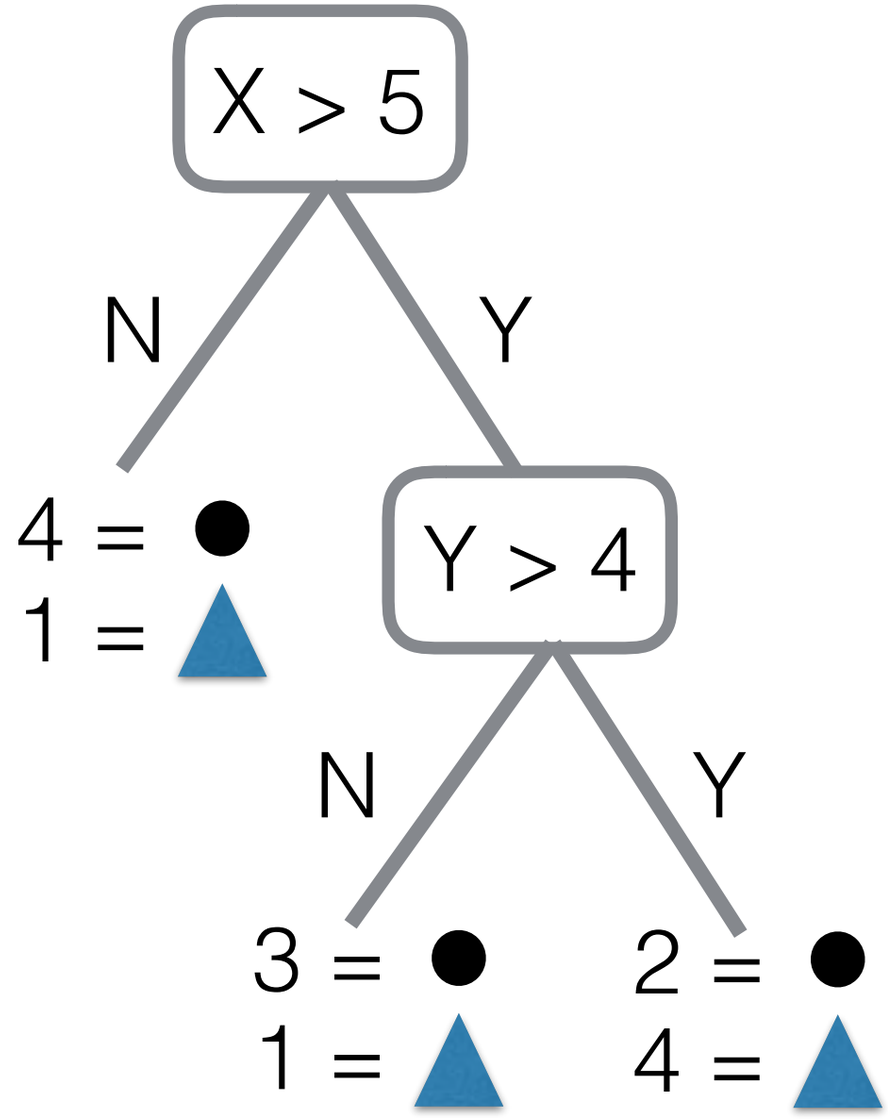
- Decision tree
- Map observations (branches) onto classes (leaves)
- Tree describes the data but can be used for classification
- EG: student states = leaves, student actions = branches



Binary Classification Tree

* Minimize the error

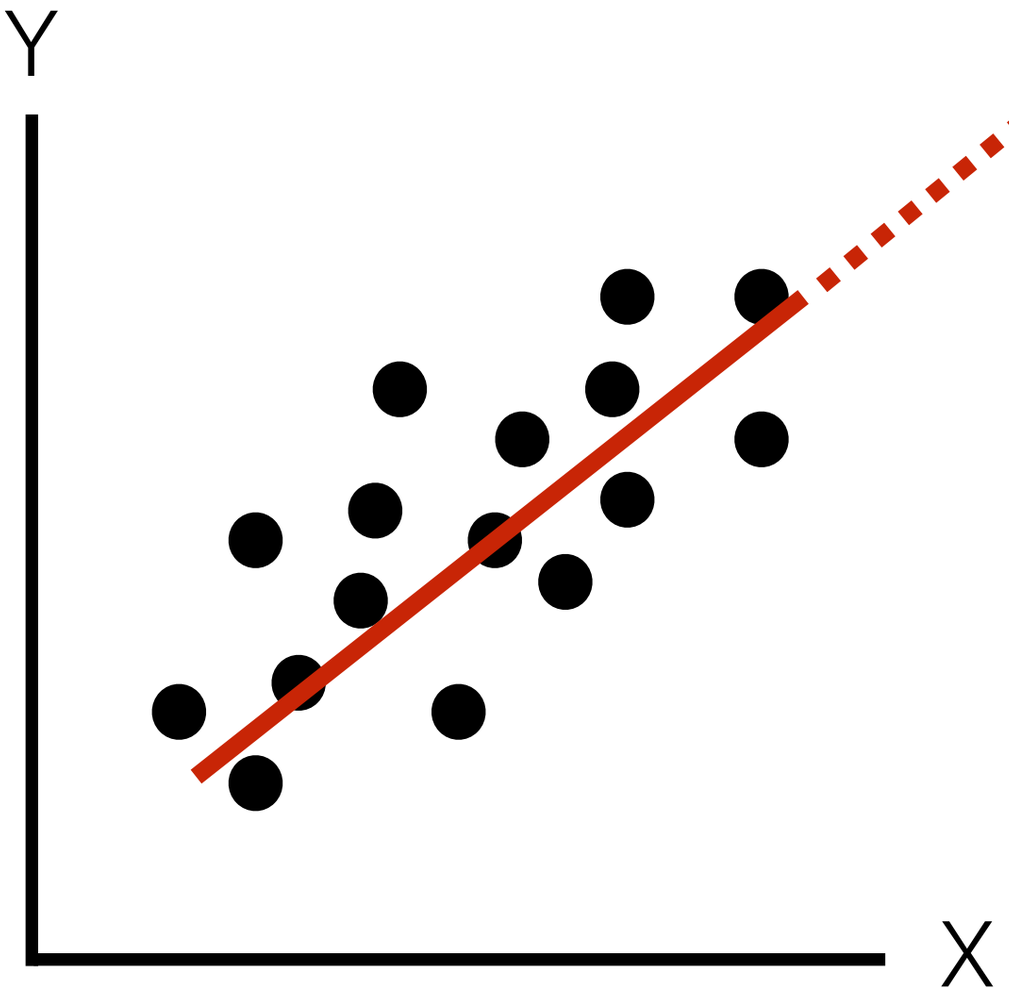




Regression

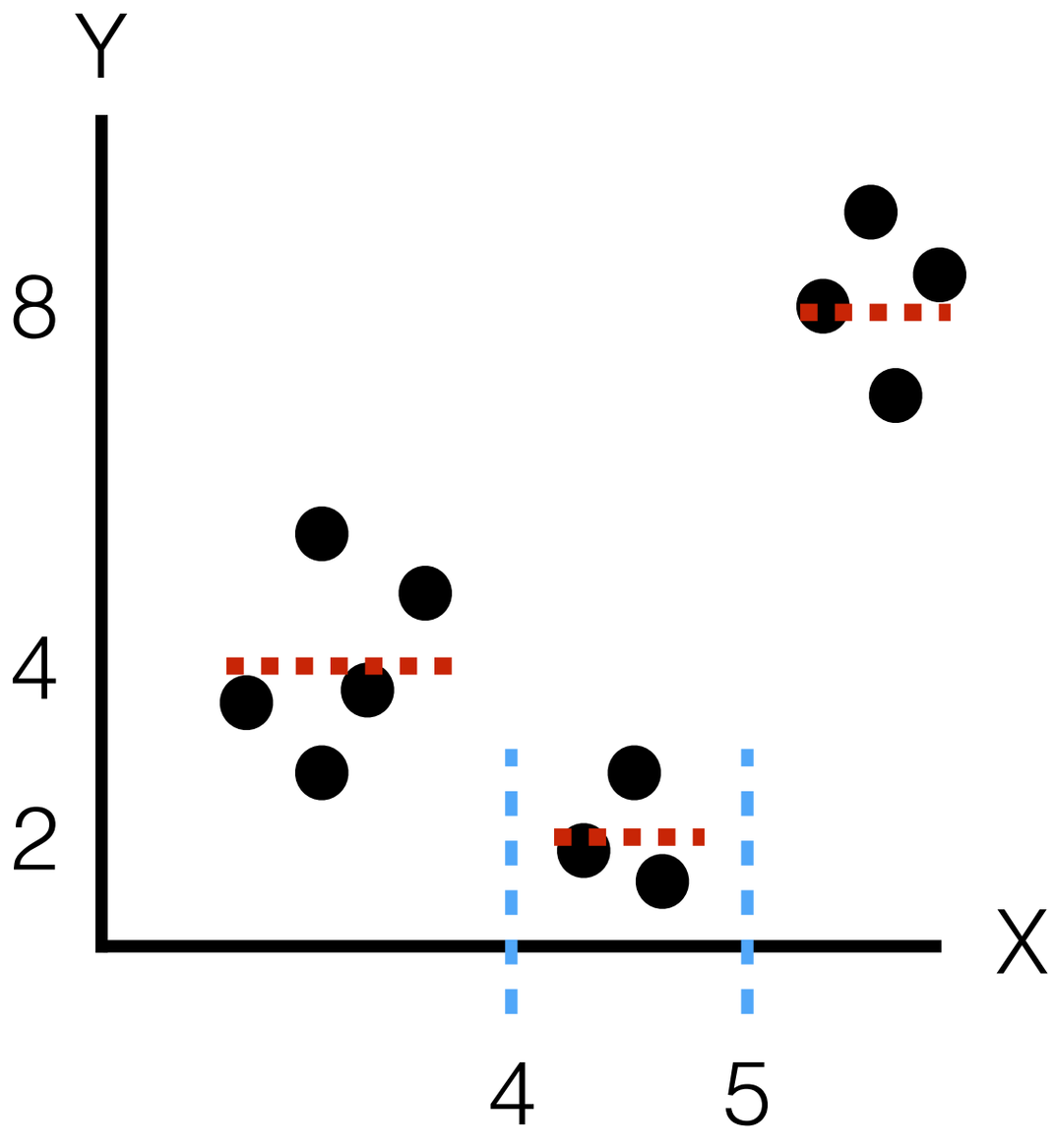
- <u>In Ed Stat</u> = OLS Regression/Logistic Regression (characterize)

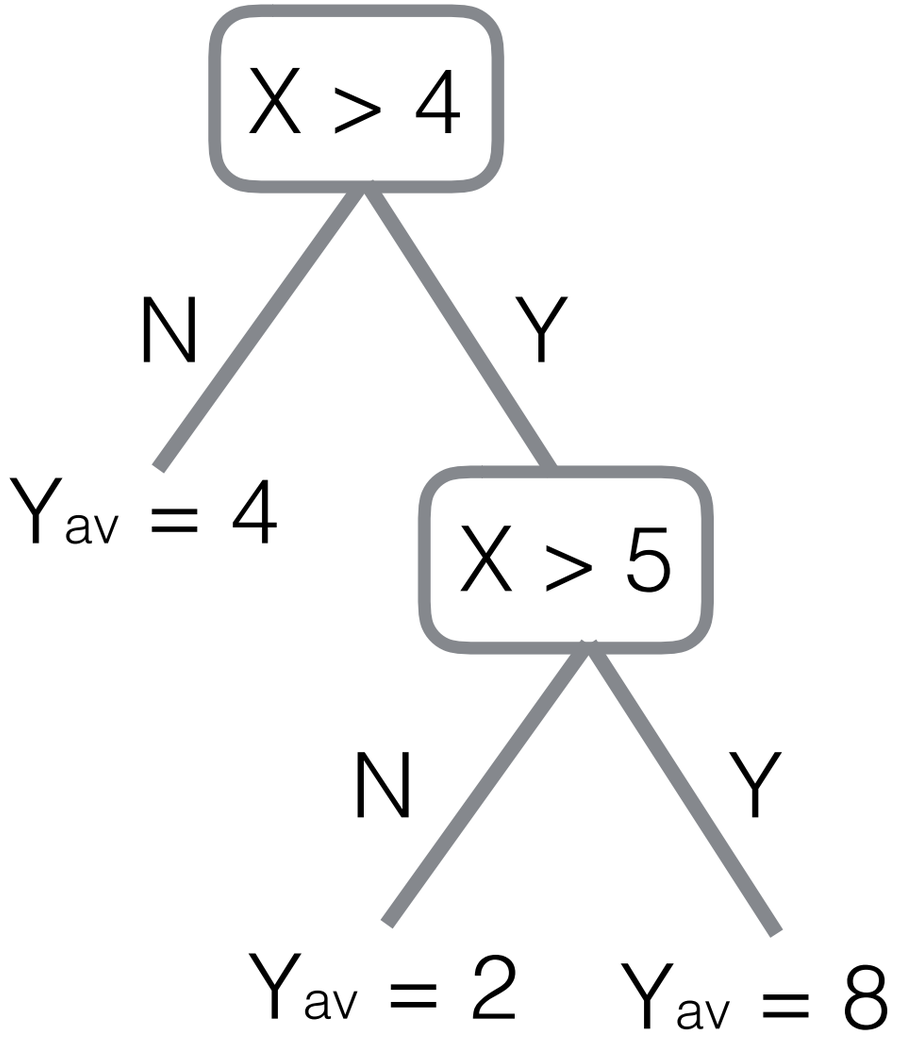
- <u>In ML</u> = Mapping from unlabeled instances to a value within a continuous range (future)



Binary Regression Tree

* Minimize the error



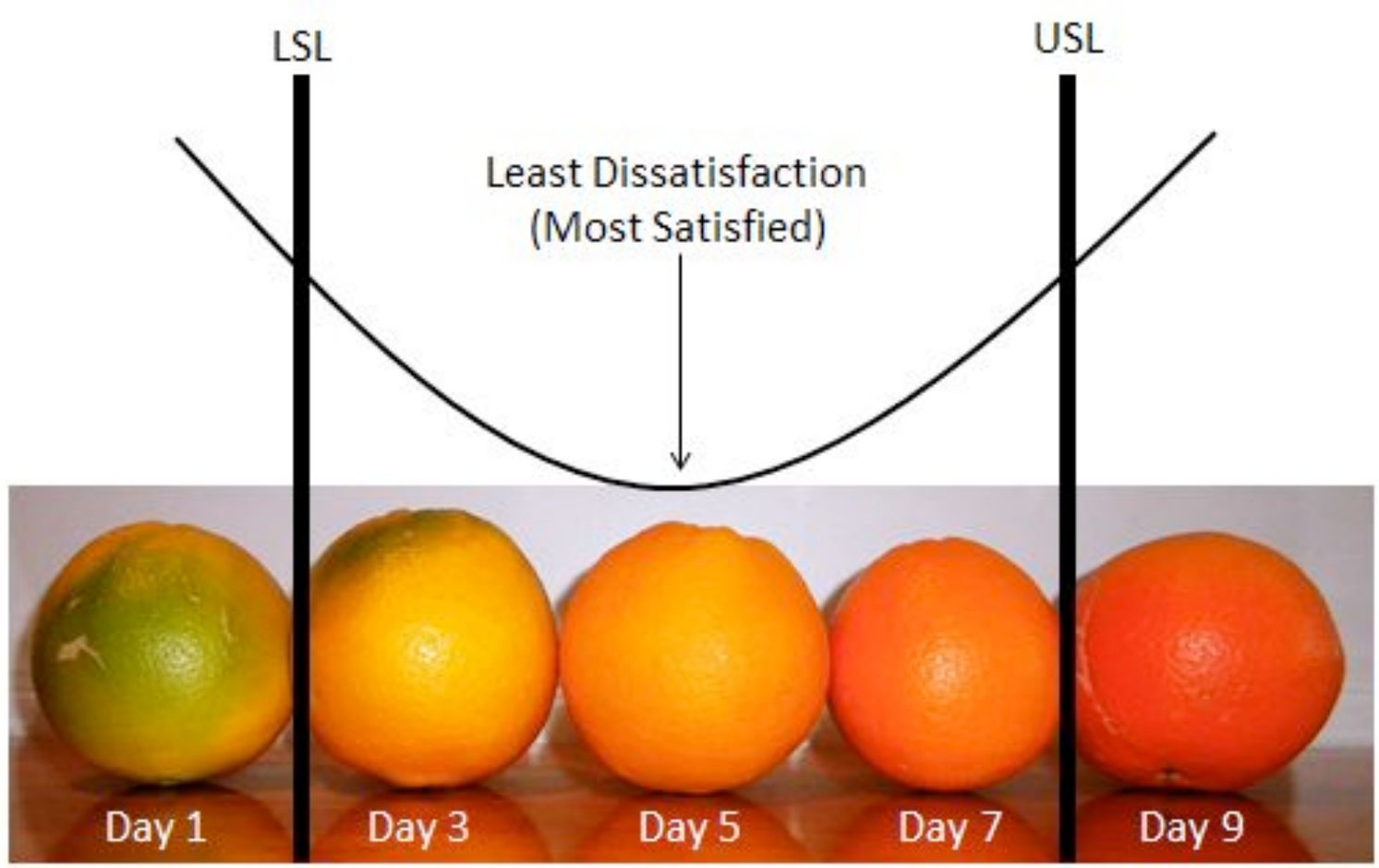


The thing is…

- I left some stuff out last time 🙄
- Namely, how the algorithm determines how to construct the tree (**splitting criterion**)
- Machine Learning algorithms often try to minimize a **loss function**
- Generate splitting criterion using **recursion**

Loss/Cost Functions

Maps an event (variables) onto a real number intuitively representing some "cost" associated with the event



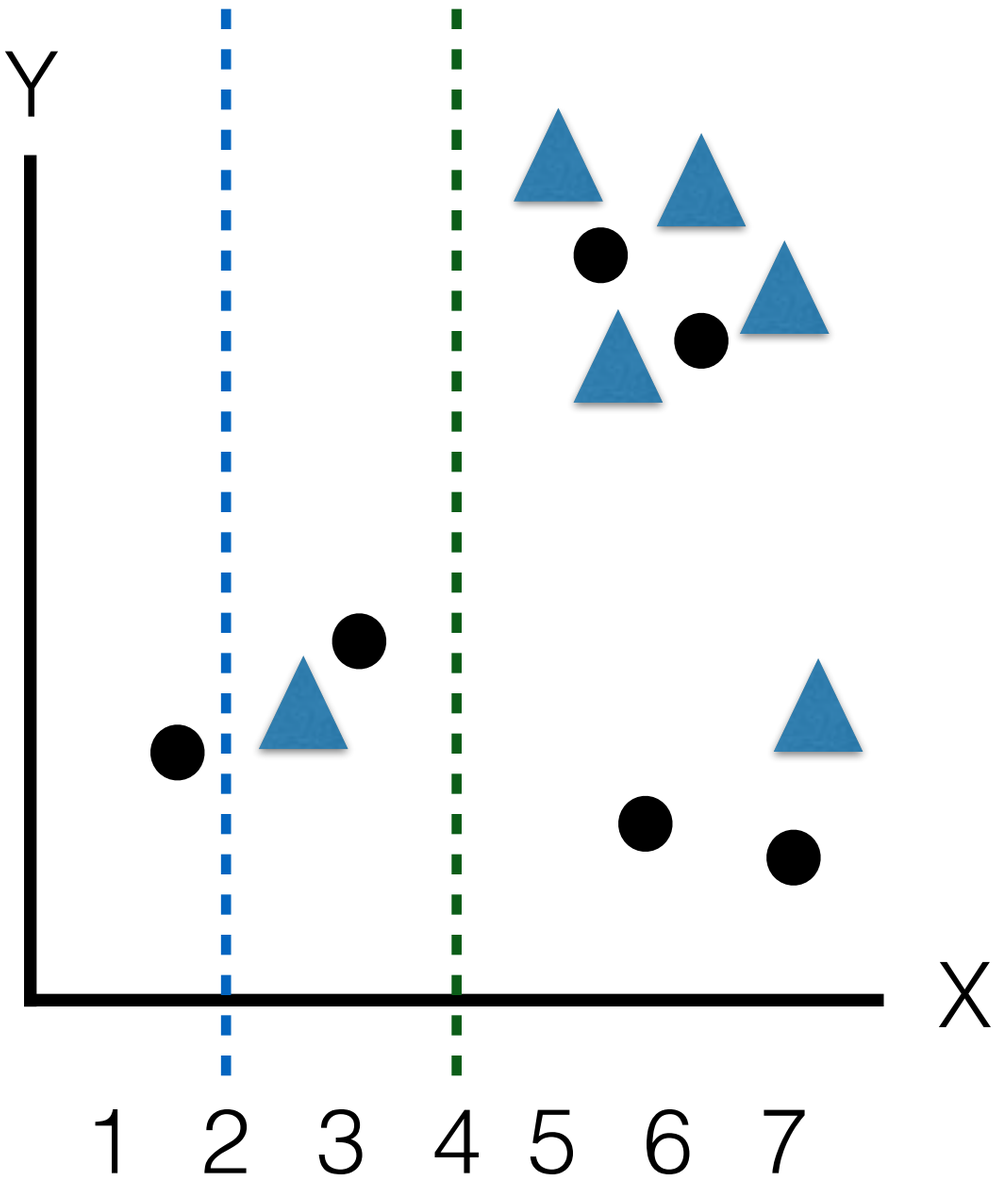
CART

- Gini Impurity (NOT index/coefficient)
- Sum of probability of an item being labelled *i*, multiplied by the probability of a mistake

$$Gini(E) = 1 - \sum_{j=1}^{c} p_j^2$$

- Zero when there are no mistakes

Gini Impurity



▲ = $1 - (6/6)^2 - (0/6)^2 = 0$

● = $1 - (1/6)^2 - (5/6)^2 = 0.28$

$G = 0 \times 6/12 + 0.28 \times 9/12$

$= 0.21$

▲ = $1 - (5/6)^2 - (1/6)^2 = 0.28$

● = $1 - (2/6)^2 - (4/6)^2 = 0.44$

$G = 0.28 \times 6/15 + 0.44 \times 9/15$

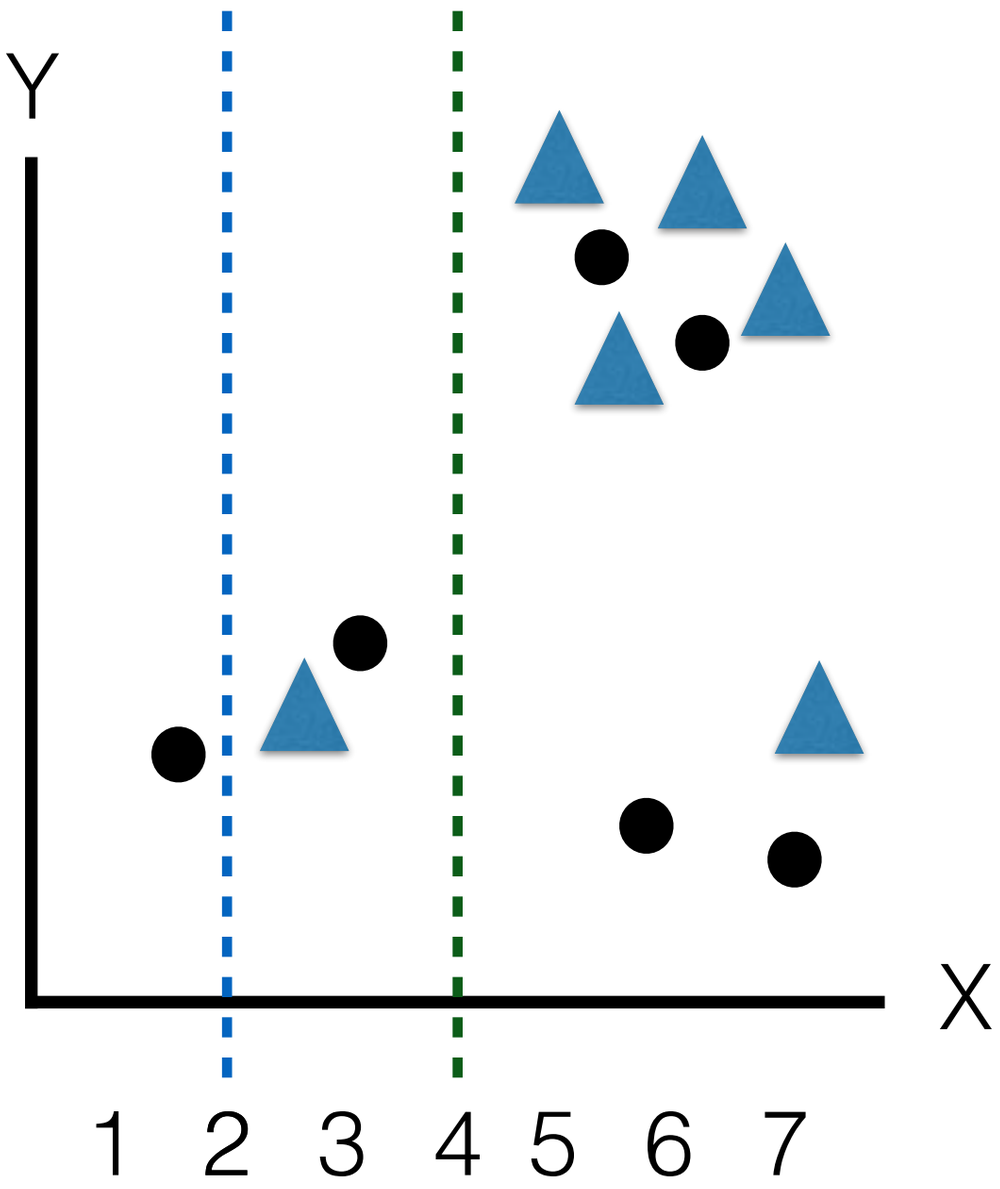
$= 0.37$

Entropy

- The amount of disorder or information loss in a system

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Entropy/Information Gain



▲ = -(11/12) x log2(11/12) = 0.11

● = -(1/12) x log2(1/12) = 0.3

E = 0.11 + 0.3 = 0.41

IG = 1 - 0.41 = 0.59

▲ = -(5/12) x log2(5/12) = 0.53

● = -(2/12) x log2(2/12) = 0.43

E = 0.96

IG = 1 - 0.96 = 0.04

Recursion

- Each sub-population may in turn be split an indefinite number of times
- Splitting process terminates after a particular **stopping criterion** to avoid overfitting
  - All samples in a leaf are being labelled the same
  - Have a pre-set number of nodes
  - Have a pre-set tree depth
  - Impurity or entropy no longer decrease/significant

Algorithms

- CART - Gini Impurity
- ID3 (Iternative Dichotomizer) - Entropy (Quinlan, 1980s)
- C4.5/C5 - Entropy (Quinlan, 1990s)
- CHAID - Chi-square Automatic Interaction Detector
- MARS - Proprietary
- Conditional Inference Trees

Pruning

- Reduce complexity of tree to prevent **overfitting**

- **Top down:** When you prune at the creation of the root

- **Bottom up:** When you compare leaves and remove