FIT1043 Introduction to Data Science

Module 5: Data Analysis Process

Lecture 9

Monash University

Discussion: Investigating Twitter data in the Shell

We have analysed a **large data file** from Twitter in the shell during the tutorial:

- Aim: understand what data the file contained, how we could reformat the data for further analysis
- Many **different types of columns**:
  - text, dates, locations, even code containing data structures
- real data: lots of missing data, errors, ...
- shell commands like *grep* and *cut* simplify the inspection and manipulation of the data

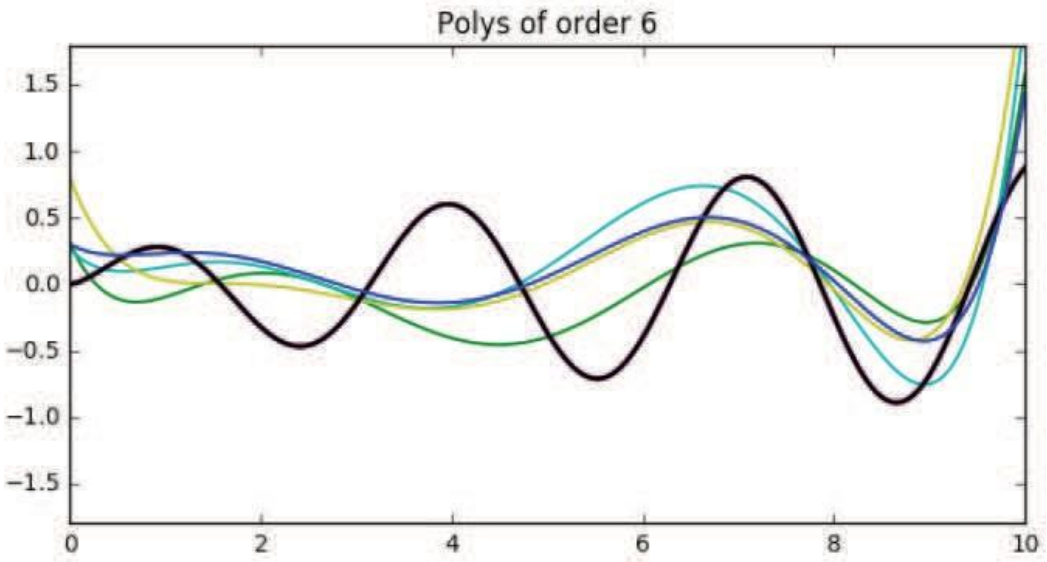
Unit Schedule: Modules

Module	Week	Content
1.	1	Overview and look at projects
	2	(Job) roles, and the impact
2.	3	Data business models / application areas
3.	4	Characterising data and "big" data
	5	Data sources and case studies
4.	6	Resources and standards
	7	Resources case studies
5.	8	**Data analysis theory**
	9	**Regression and decision trees**
	10	**Data analysis process**
6.	11	Issues in data management
	12	Data management frameworks

Theory of Data Analysis, cont.

Bias and Variance

Bias and Variance

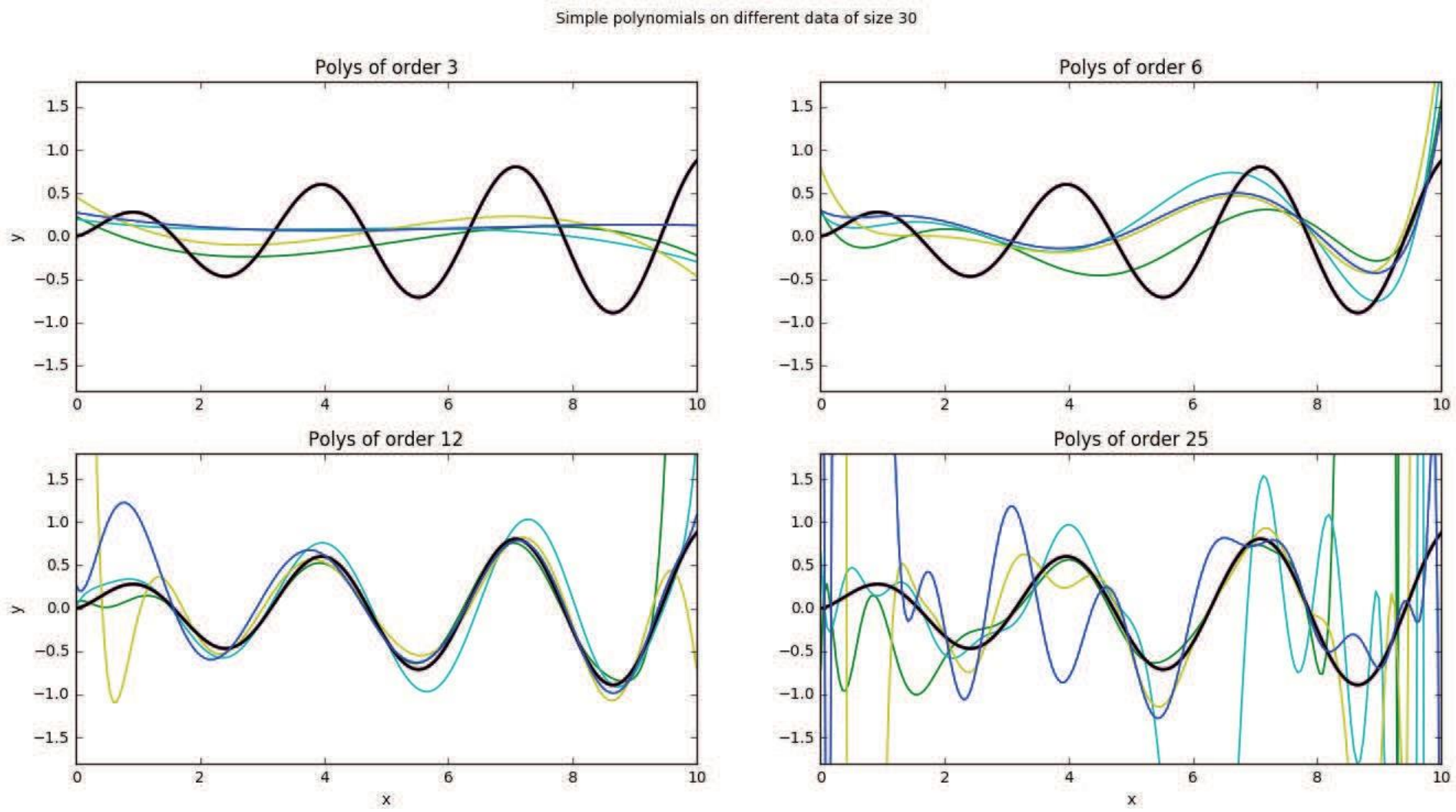


Different data sets of size 30.

Bias: measures how much the prediction differs from the desired regression function.

Variance: measures how much the predictions for individual data sets vary around their average.

Bias-Variance Examples

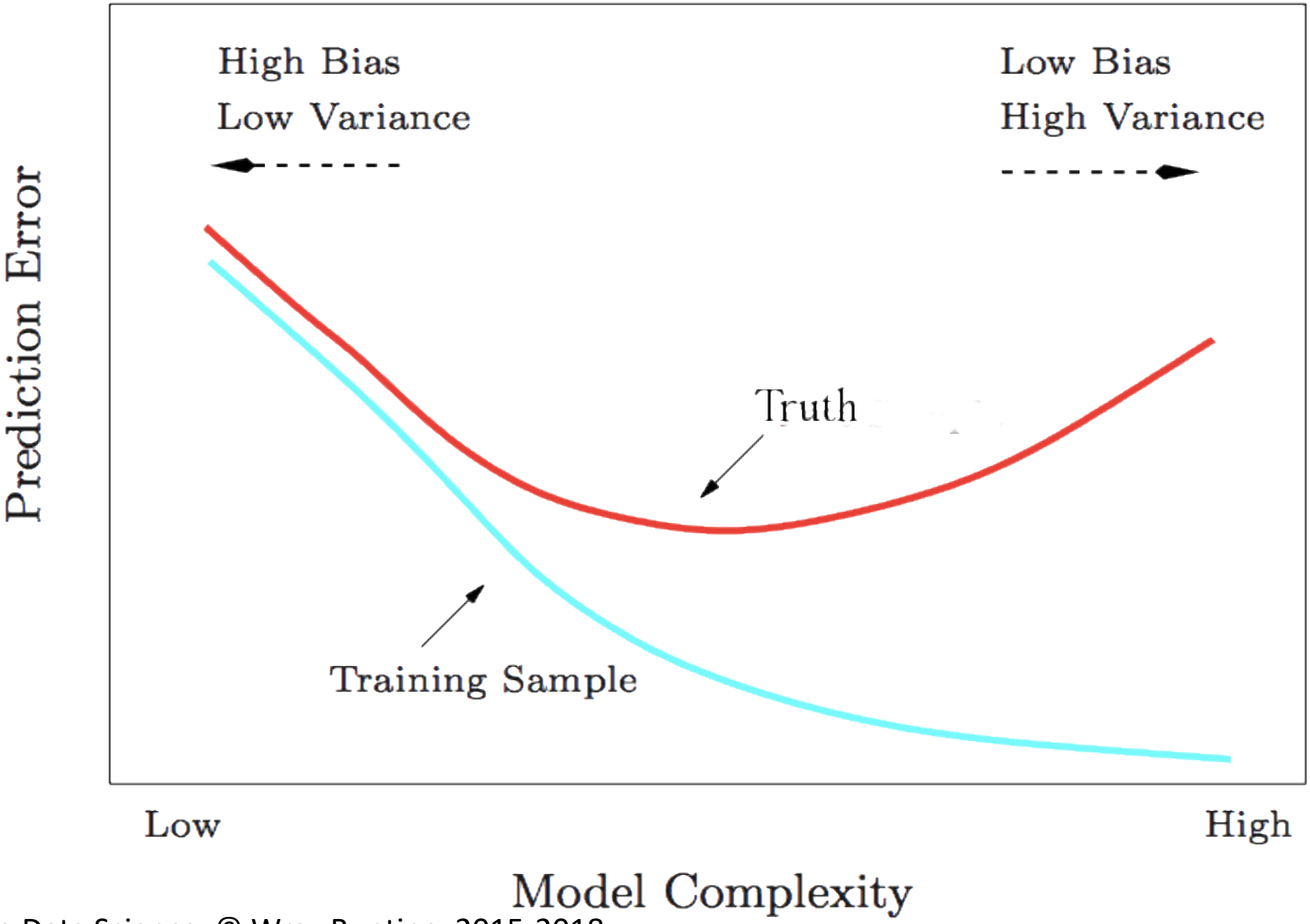


MARS Question

Which of the polynomials in the previous slide is a better model?

A. Order 3
B. Order 6
C. Order 12
D. Order 25

Bias-Variance Tradeoff



No Free Lunch Theorem

Wolpert and McCready proved:

> *if a [learning] algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems*

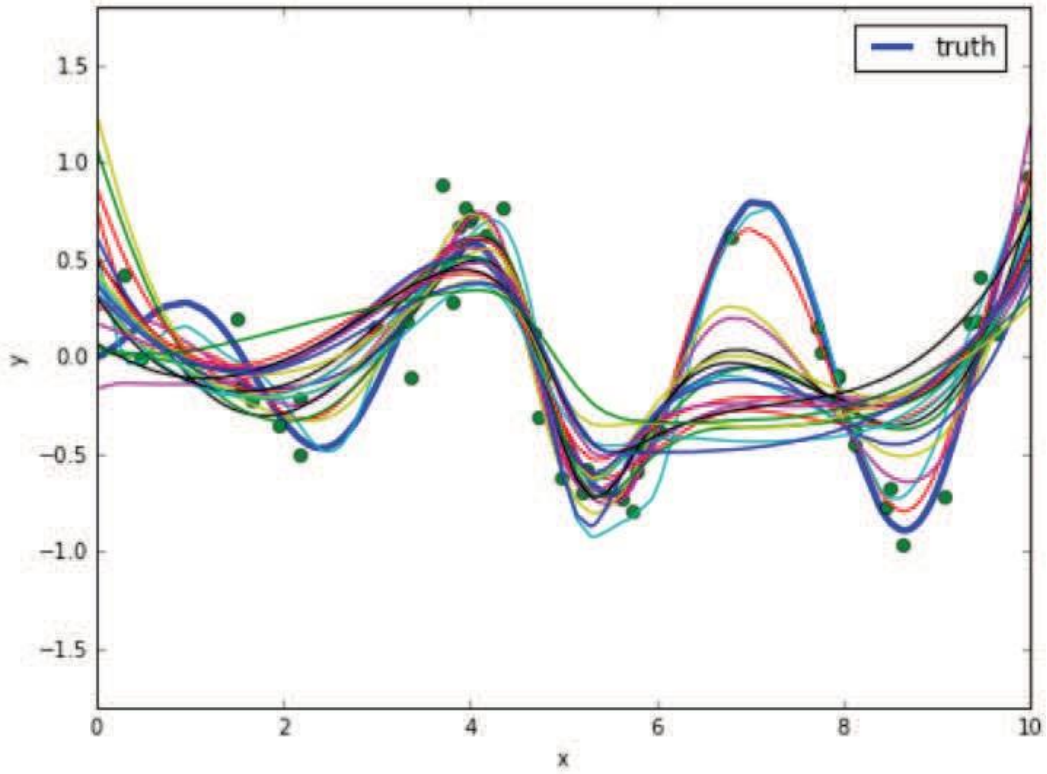
- there is no universally good machine learning algorithm (when one has finite data)

e.g. Naive Bayesian classification performs well for text classification with smaller data sets

e.g. linear Support Vector Machines perform well for text classification

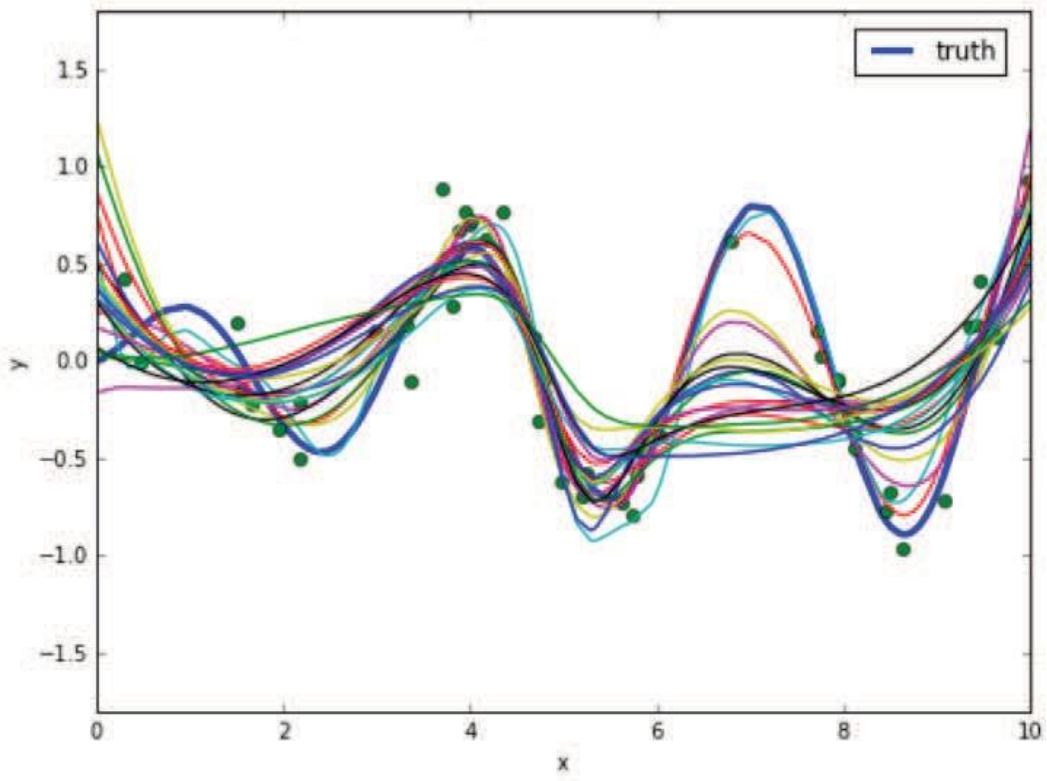
Ensembles

- given only data, we do not know the truth and can only estimate what may be the “truth”
- an ensemble is a collection of possible/reasonable models
- from this we can understand the variability and range of predictions that is realistic



Ensembles (cont.)

- generating an ensemble is a whole statistical subject in itself
- often we average the predictions over the models in an ensemble to improve performance $\hat{y}(x) = \frac{1}{M} \sum_{i=1}^{M} \hat{y}^{(i)}(x)$



Data Analysis Algorithms
Regression and decision trees

Reminder: Training Set and Test Set

- split up the data we have into two non-overlapping parts, a training set and a test set
- do your learning, run your algorithm, build your model using the training set
- run evaluation using the test set
- don't run evaluation on the training set
- how big to make the test set?

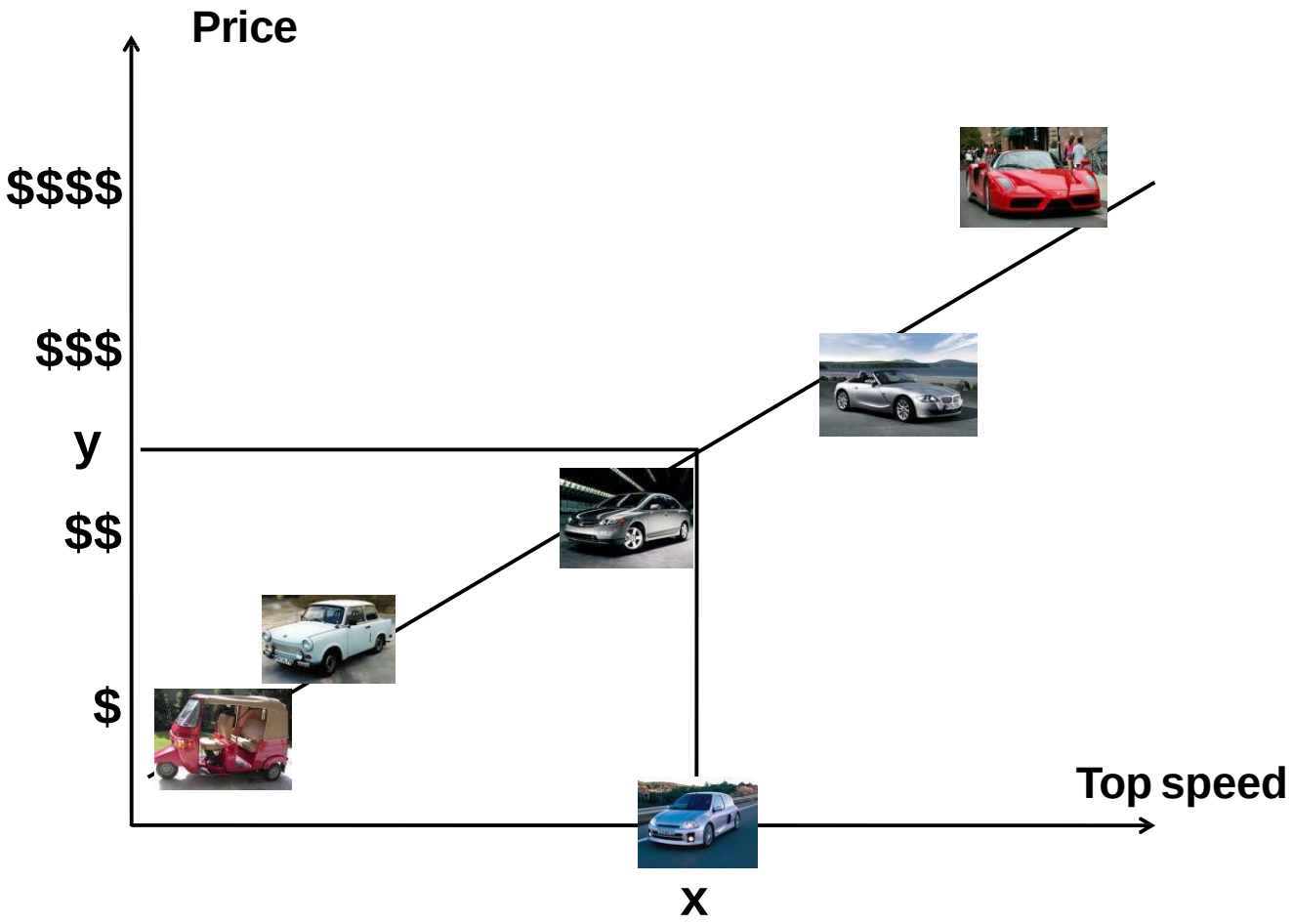
Regression

?

Regression

?

Regression (cont.)



Classification

Cat **Dog**

?

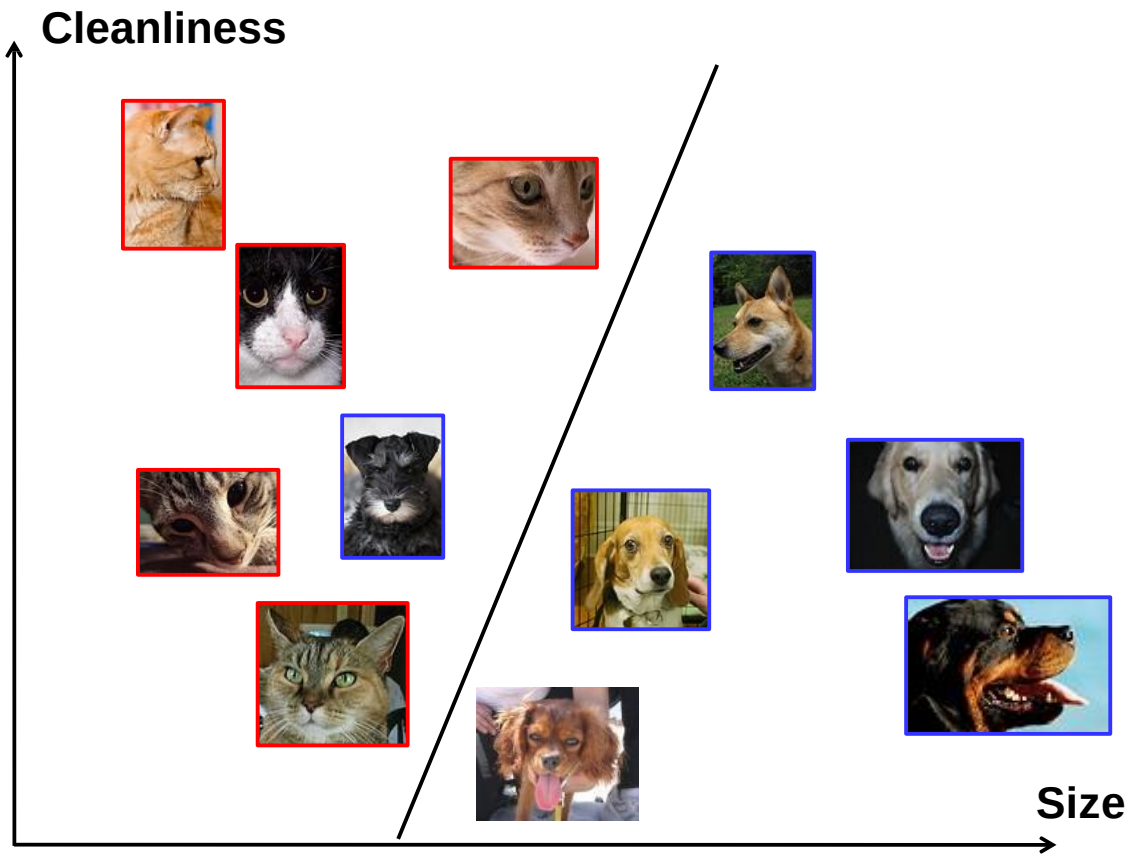
Classification

Cat

Dog

?

Classification (cont.)



What are Decision and Regression Trees?

Decision Trees:

- Predict binary (or categorical) outcomes

Regression Trees:

- Predict continuous (i.e. real) values

Tree

- Prediction model is a tree

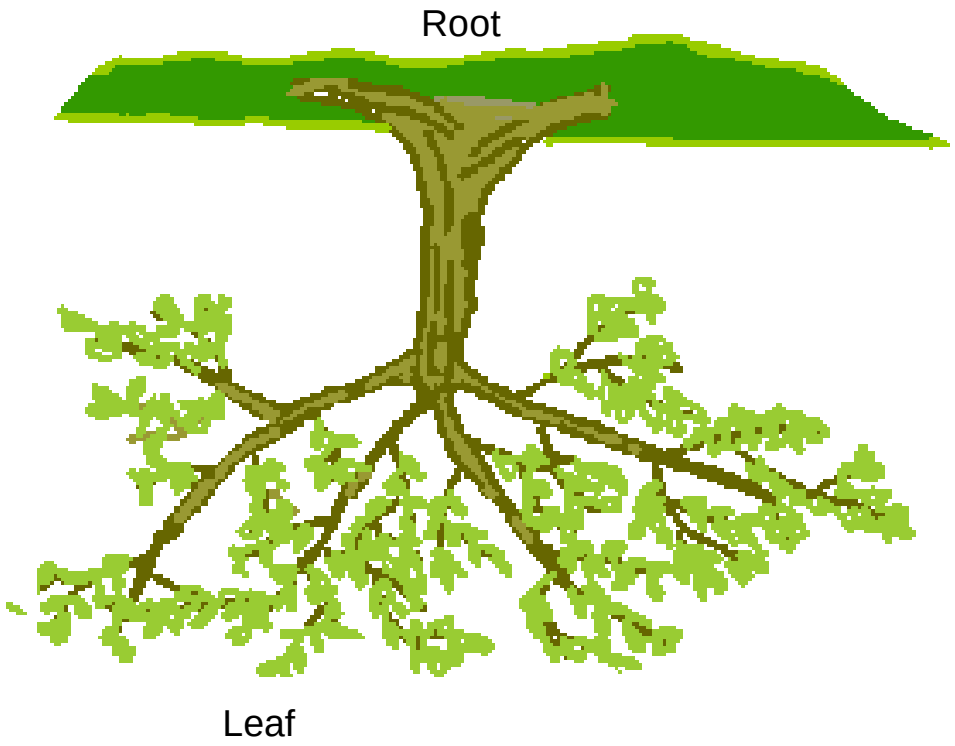
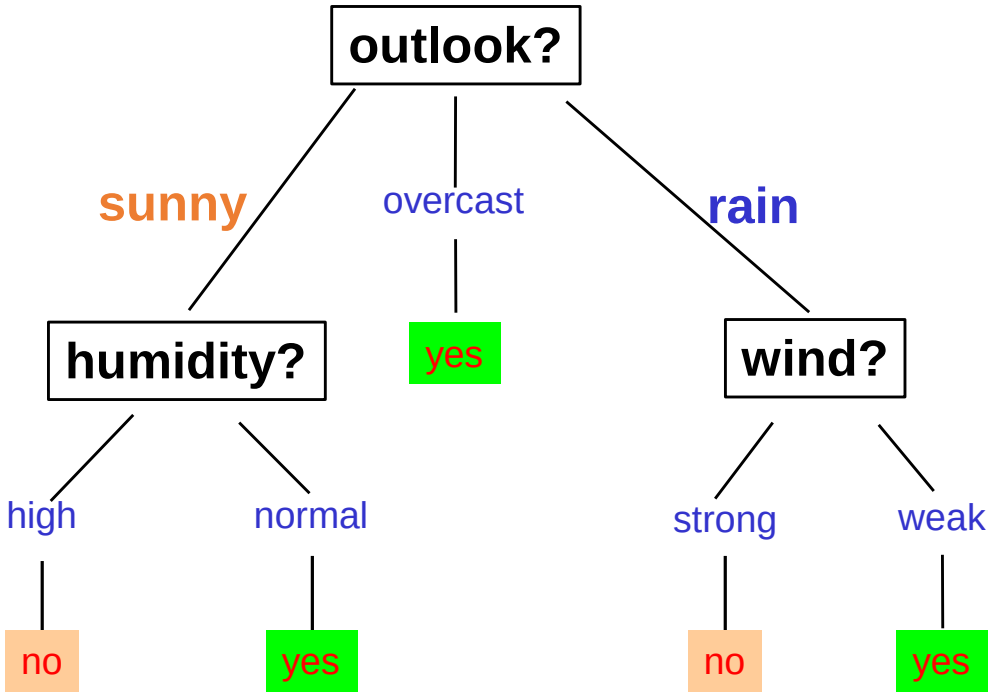
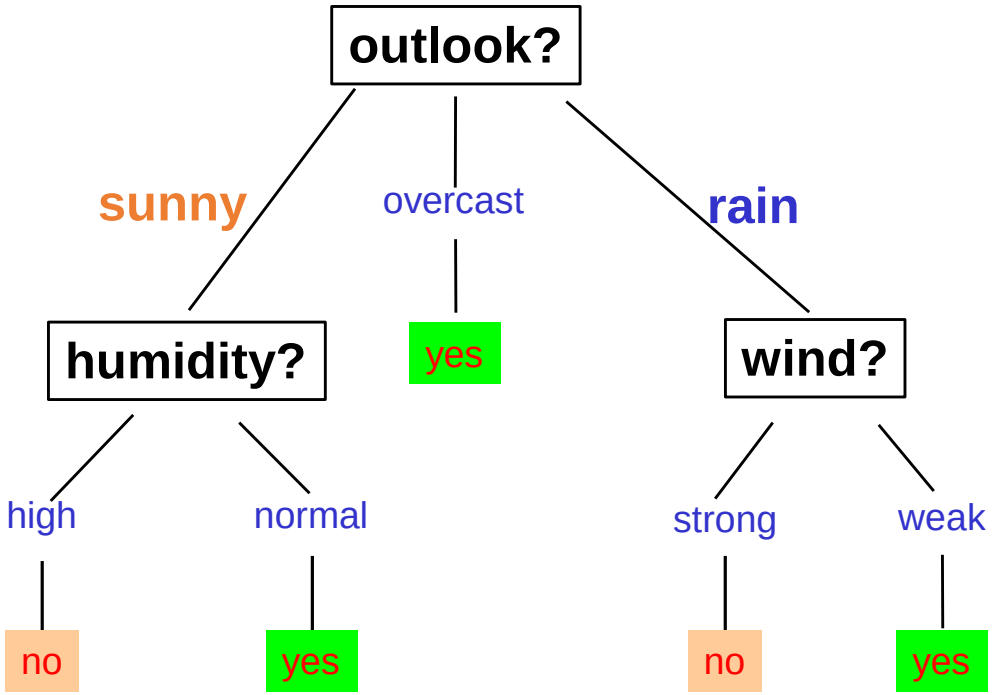


Image: webdocs.cs.ualberta.ca

Decision Tree Example



Decision Tree Example



Set of rules:

G-Day to play tennis ⇔ (Sunny and Normal) or Overcast or (Rain and Weak)

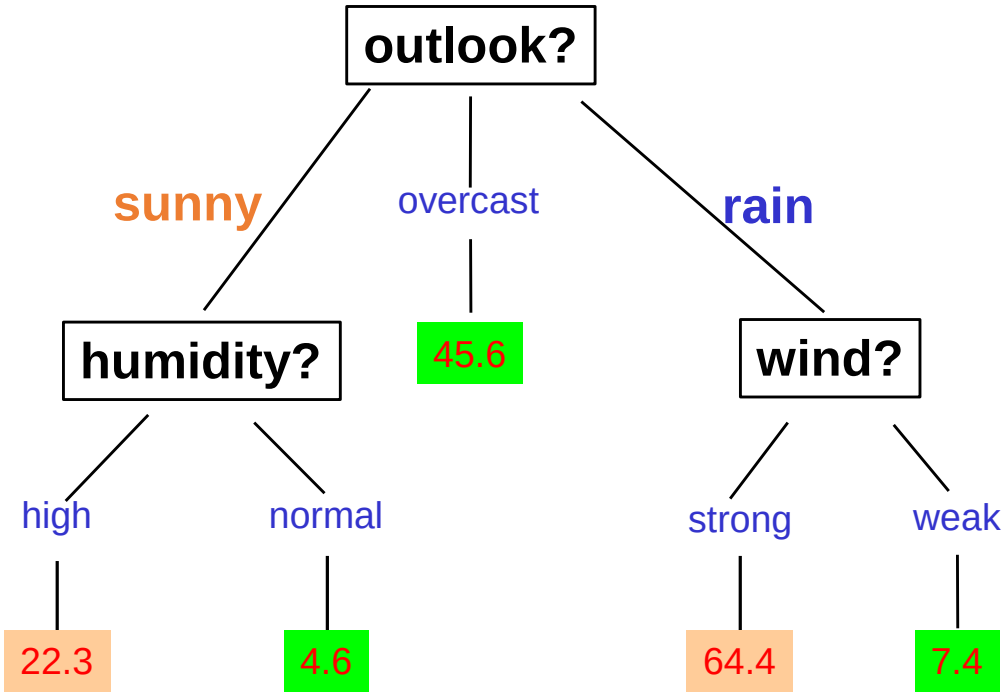
B-Day to play tennis ⇔ ?

MARS Question

According to the previous slide when is a bad day to play tennis?

A. When it's sunny and humidity is high
B. When it's rainy and wind is strong
C. Both above options

Regression Tree Example



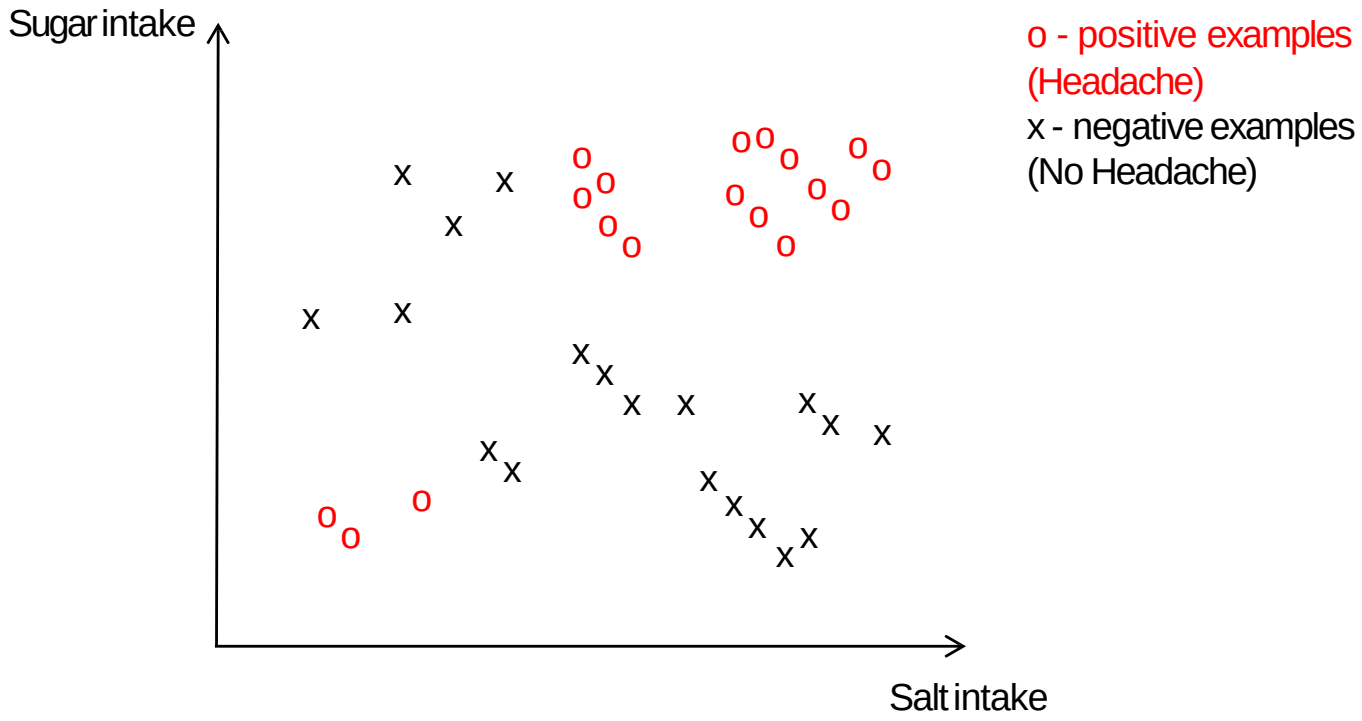
How to Build Regression and Decision Trees?

- Recursively partition (divide up) the feature space into regions

- While grouping similar instances together

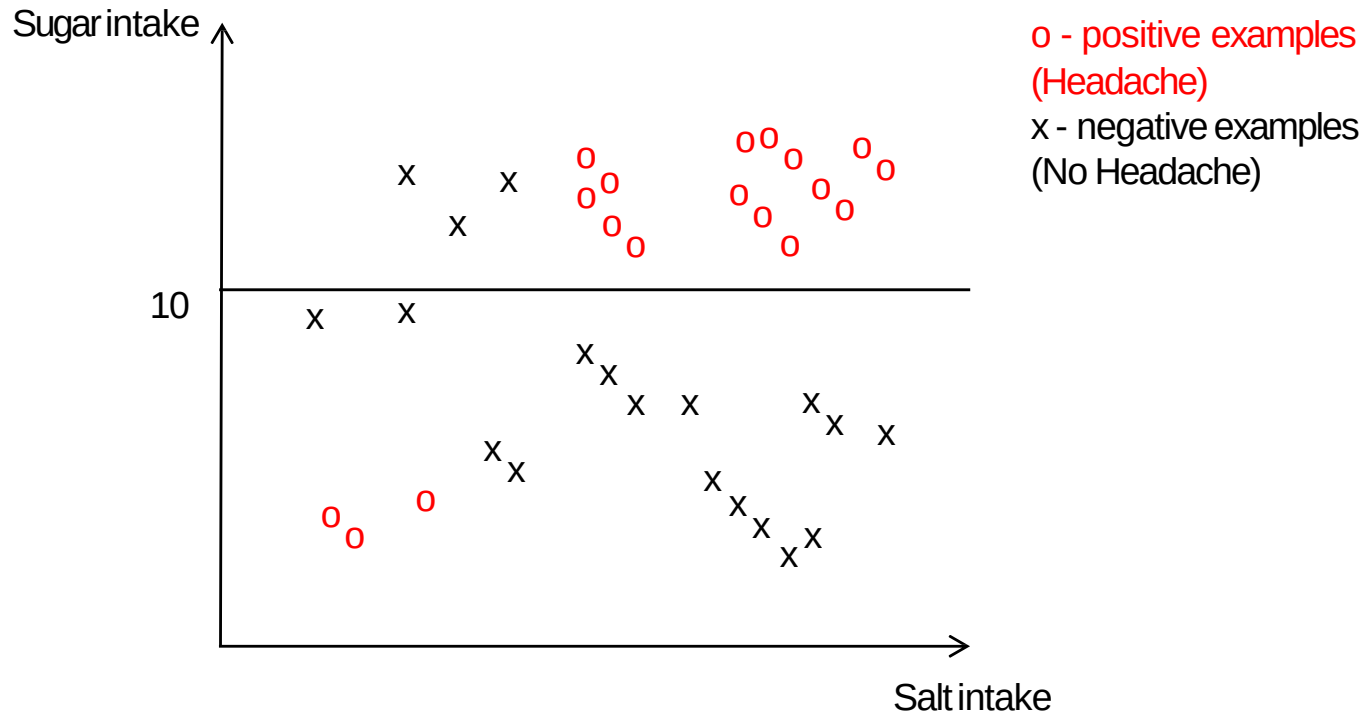
Recursive Partitioning

- At each iteration, we divide the data to group similar instances together



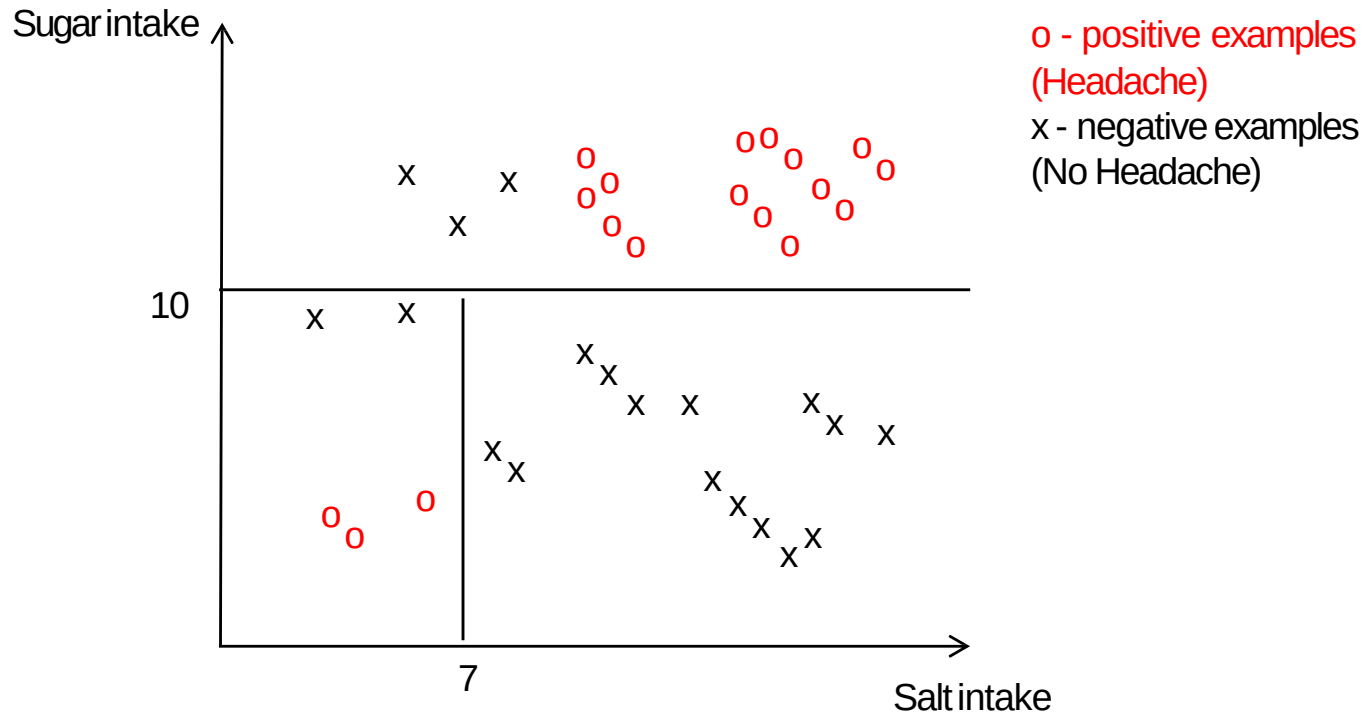
Recursive Partitioning

- At each iteration, we divide the data to group similar instances together



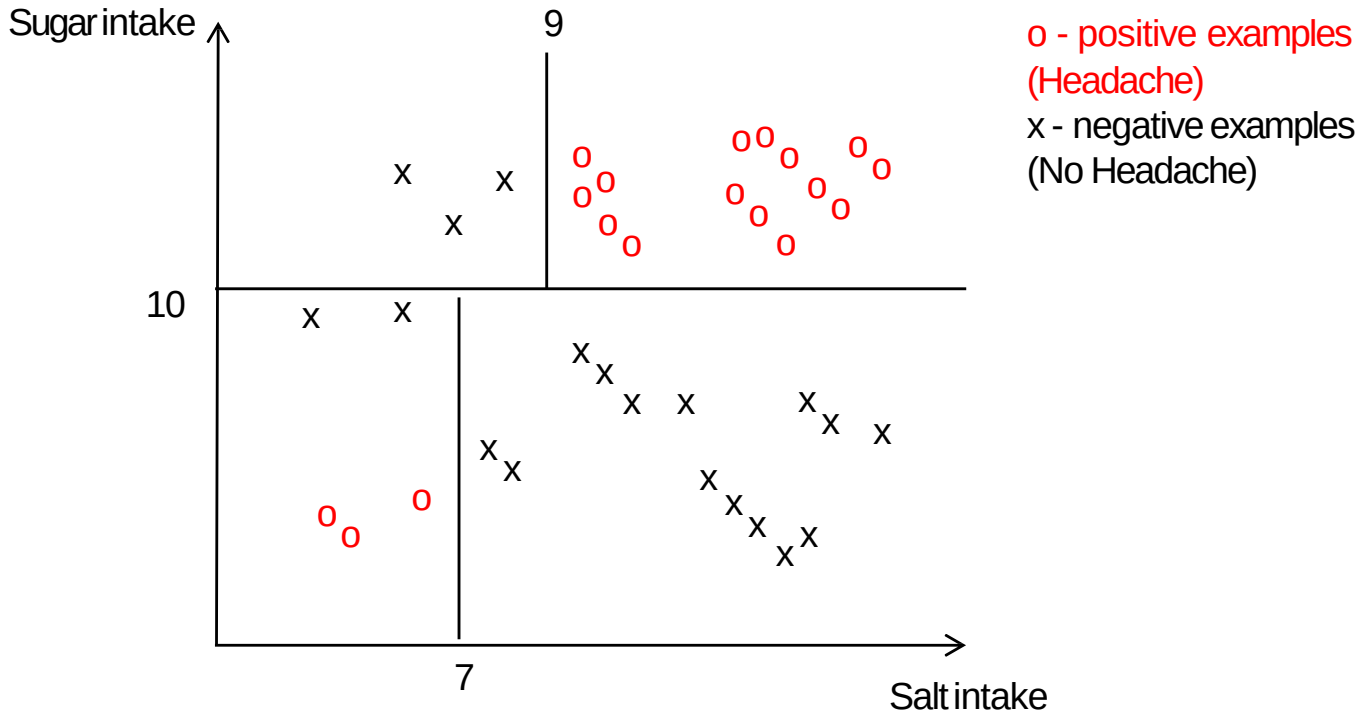
Recursive Partitioning

- At each iteration, we divide the data to group similar instances together



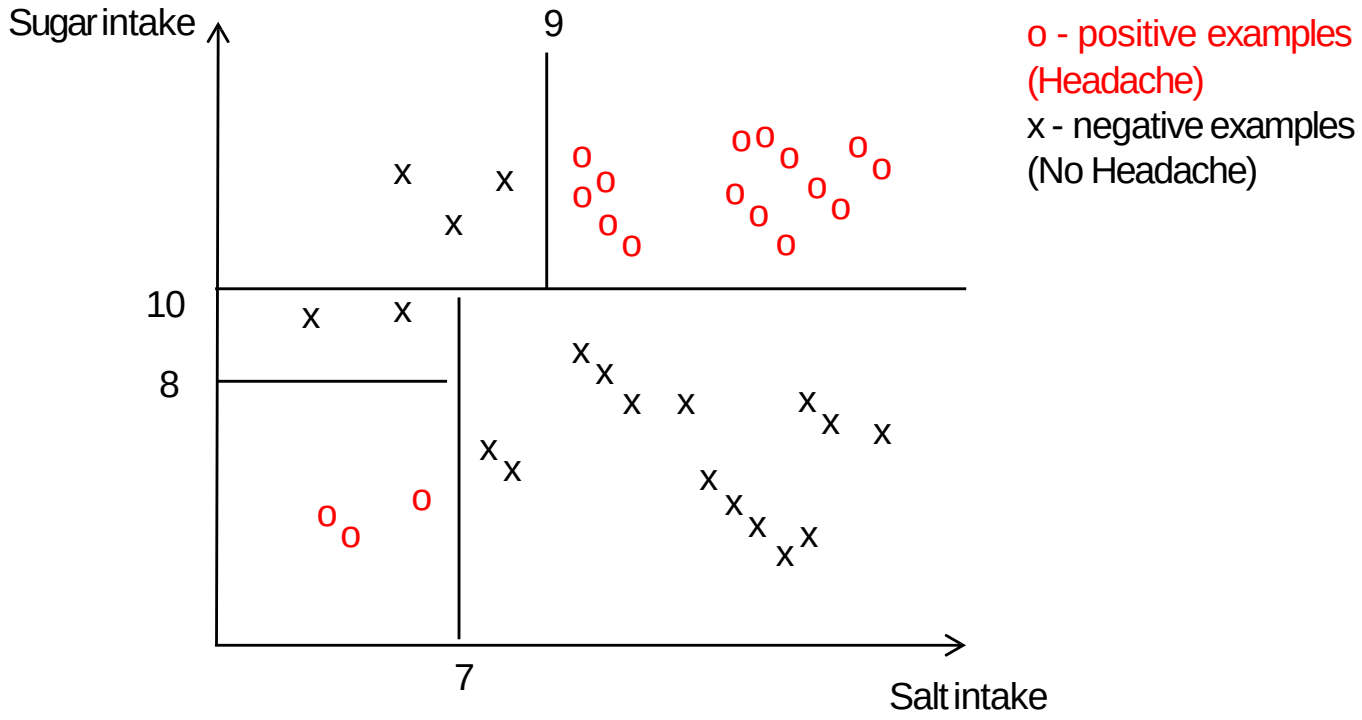
Recursive Partitioning

- At each iteration, we divide the data to group similar instances together



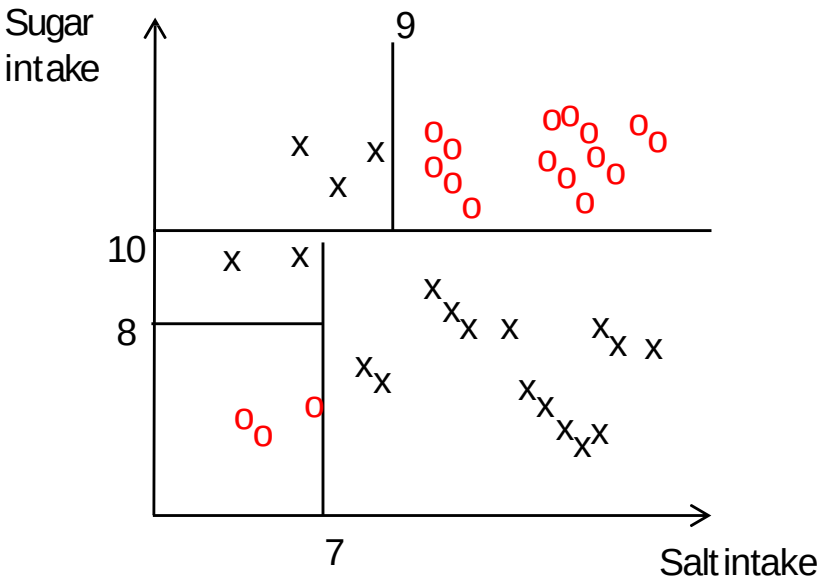
Recursive Partitioning

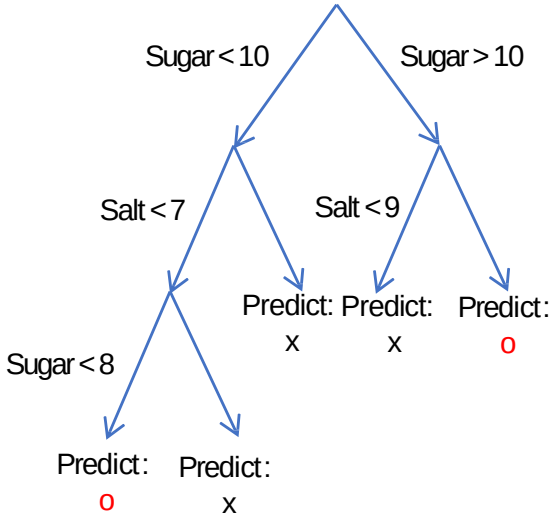
- At each iteration, we divide the data to group similar instances together



Prediction Model is a Tree

- This model learnt can be represented as a tree with predictions at the leaves:





Prediction in Decision and Regression Trees

Decision Trees:

- Prediction is the most common values in each region

Regression Trees:

- Prediction is usually the average value in each region

Using Decision/Regression Trees

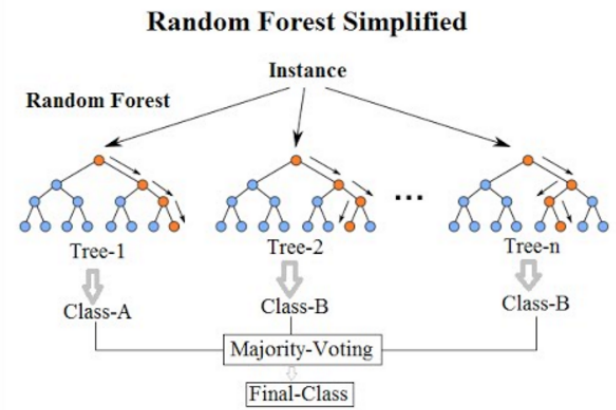
- We use BigML to train Decision and Regression Trees in the tutorial

- BigML: is a powerful Machine Learning service that offers an easy-to-use interface for you to import your data and get predictions out of it.

Decision/Regression Trees- More information

- Algorithms for building Decision & Regression trees differ on the criteria used to:
  - decide on which feature to split on in each iteration
  - decide when to stop splitting

- **Random forests**: Ensemble learning method that operates by constructing a number of decision trees



- More information on Decision & Regression trees available at:
  - https://en.wikipedia.org/wiki/Decision_tree_learning

Unit Schedule: Next Week

Module	Week	Content
1.	1	Overview and look at projects
	2	(Job) roles, and the impact
2.	3	Data business models / application areas
3.	4	Characterising data and "big" data
	5	Data sources and case studies
4.	6	Resources and standards
	7	Resources case studies
5.	8	**Data analysis theory**
	9	**Regression and decision trees**
	10	**Data analysis process**
6.	11	Issues in data management
	12	Data management frameworks

Have a Great Mid-Semester Break!