

INTRO to DATA SCIENCE

LECTURE 13: REVIEW

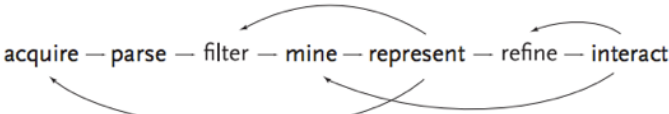


KEY OBJECTIVES

- **CONSOLIDATE KNOWLEDGE OF THE THEORY STUDIED SO FAR**
- **ACTIVATE AND DISCOVER LINKS BETWEEN THEORY AND LAB**
- **PRACTICE THE TECHNIQUES LEARNED**
- **ACQUIRE CONFIDENCE AND INDEPENDENCE IN FACING NEW**

INTRO TO DATA SCIENCE

REVIEW



	Continuous	Categorical
Supervised	regression ✓	classification ✓
Unsupervised	dimension reduction	clustering

5 Classification methods

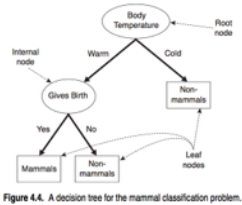
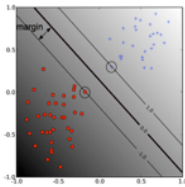
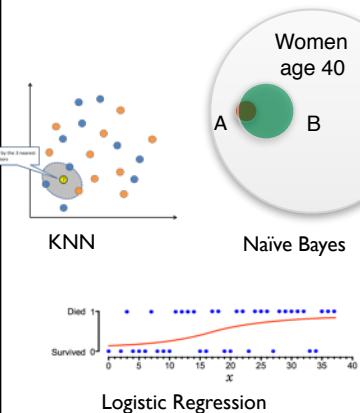


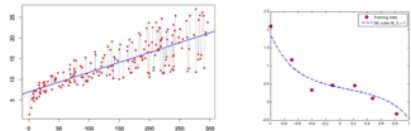
Figure 4.4. A decision tree for the mammal classification problem.

Decision Trees

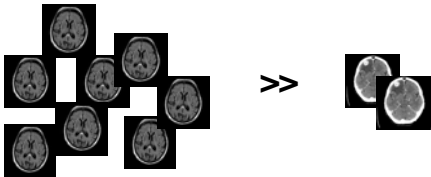
Ensembles



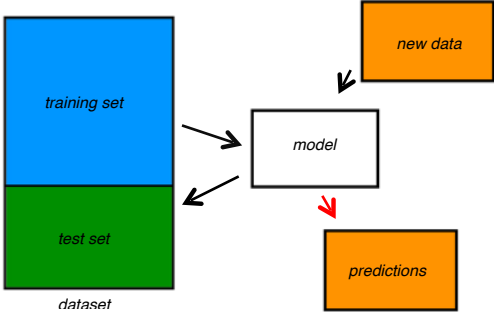
Regression & Regularization



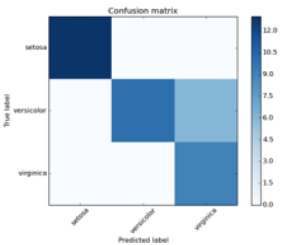
Imbalanced classes



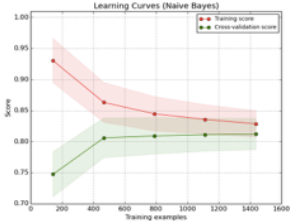
Train - Test Split



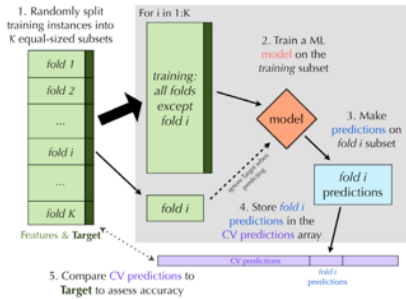
Confusion matrix

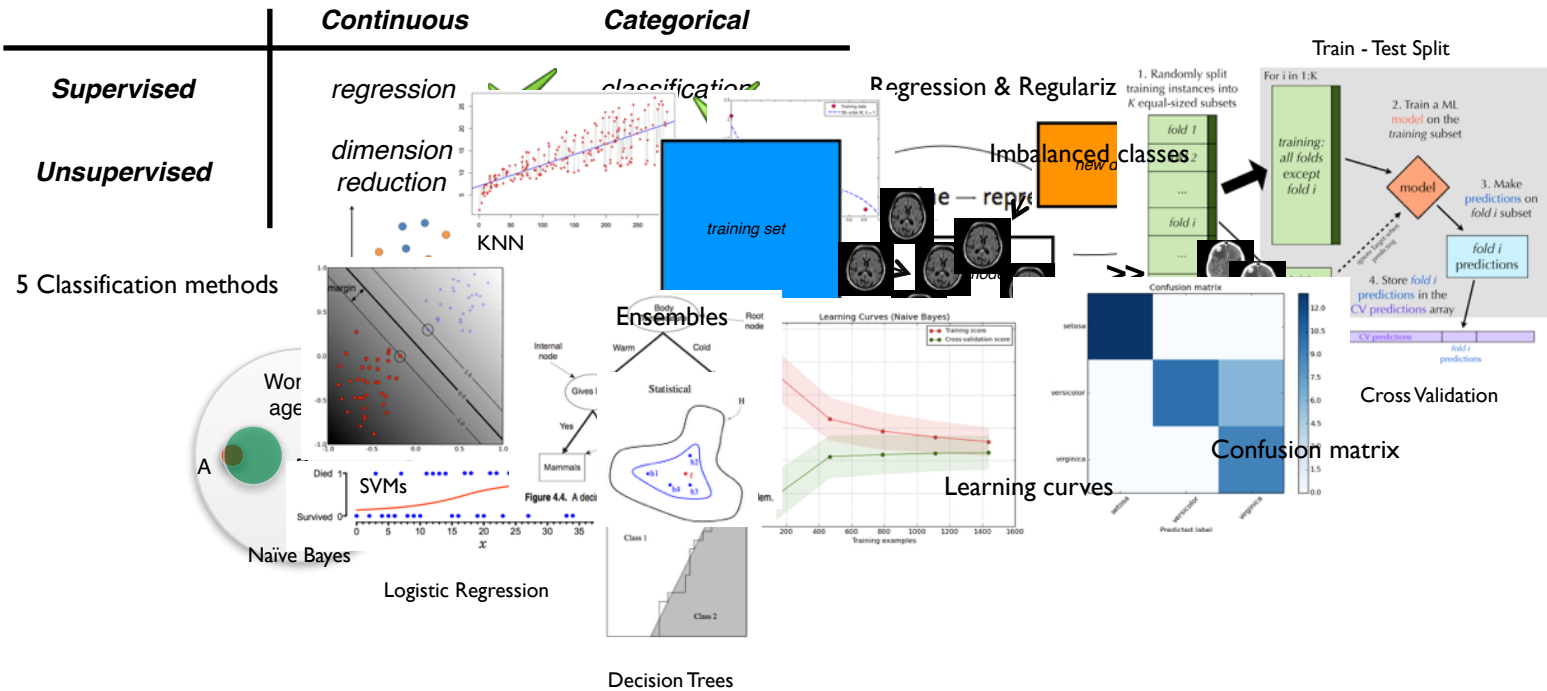


Learning curves



Cross Validation





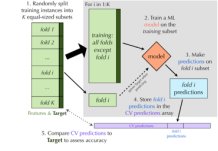
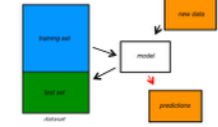
Theory

Data science process

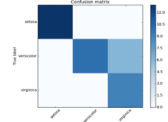


Machine Learning Problems

	Continuous	Categorical
Supervised	regression ✓ dimension reduction	classification ✓ clustering
Unsupervised		



Cross Validation



Confusion Matrix

Ensembles

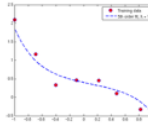
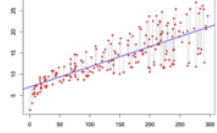
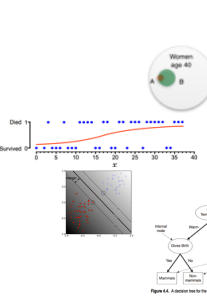
Regression

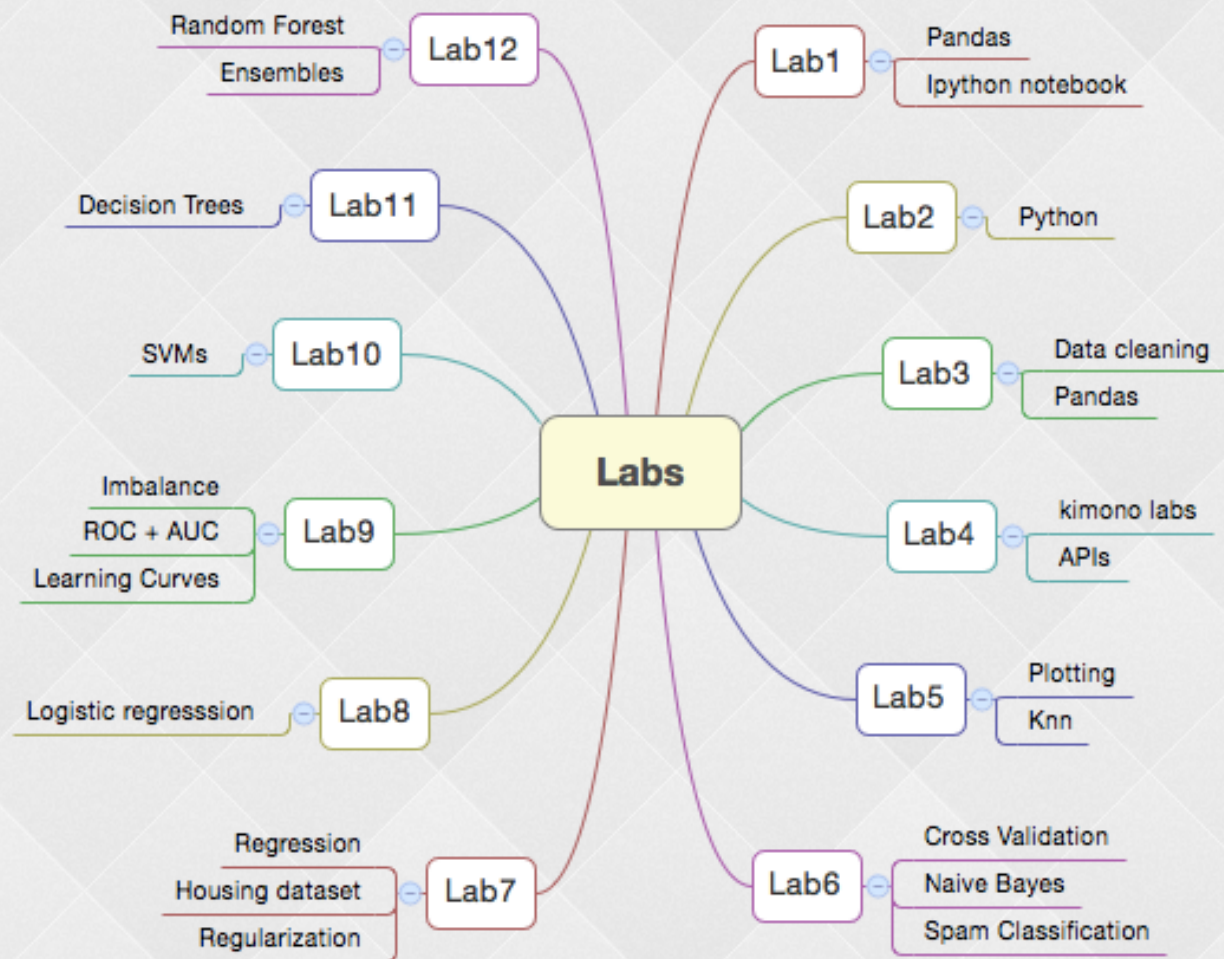
Imbalanced Classes



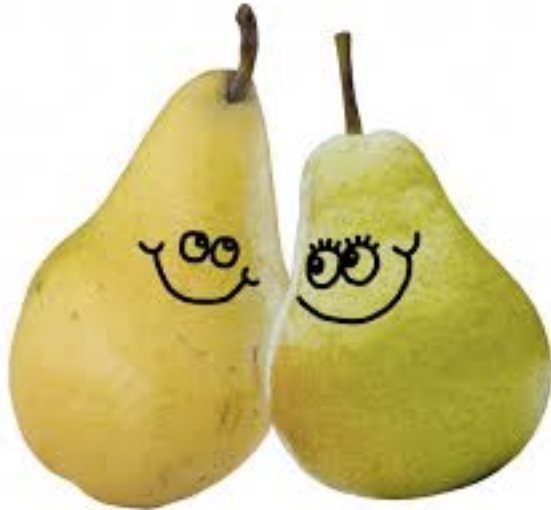
Classification

- KNN
- Naïve Bayes
- Logistic Regression
- SVMs
- Decision Trees





IN PAIRS



- Choose a lab for your pair, communicate to class
- Open the lab code and the lecture presentation
- Go through the code and discuss
 - are there any parts that are cryptic/not clear?
- Find links between lab code and theory
- Highlight missing links
 - are there any aspects of the theory you would have wanted to cover in lab or vice versa?

PAIR PROGRAMMING

PAIR PROGRAMMING



- Choose Driver and Navigator
- Open Pair programming template
- Driver writes code, Navigator thinks about the problem at hand
- After 10 minutes switch roles

ELEVATOR PITCHES