Machine Learning, Part II

INFO 370

Learning Objectives

Discuss final projects

Understand the use-cases for **machine learning**

Distinguish between **supervised** and **unsupervised** tasks

Understand the algorithm behind **decision trees**

Understand the importance of and syntax for creating **training and testing** data

Understand the algorithm behind **K Nearest Neighbors**

Be able to create and use a **validation** data set

Search for the best parameters for you models using **grid search**

Articulate the importance of (and process for) **normalizing** (scaling) your data

Today's Activities

Discuss Project Proposals

Discuss Assignment 3

Review machine learning concepts

Complete notebook 6

Work on project proposals (due **Tuesday**)

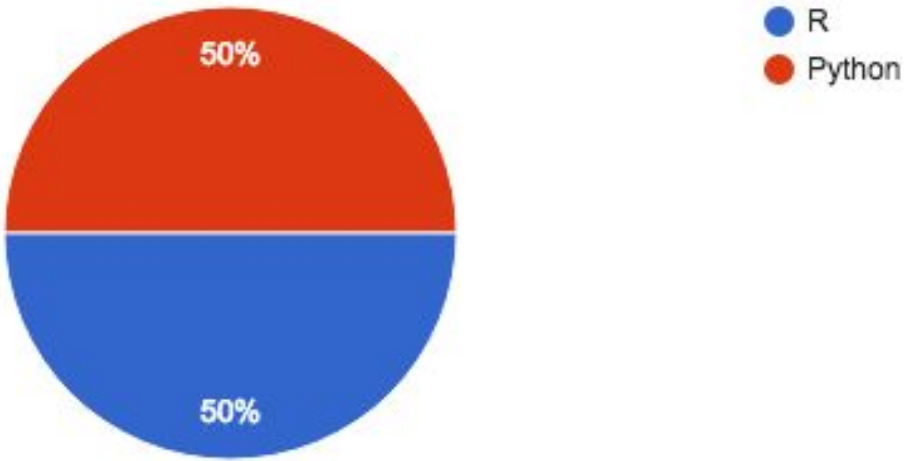
Project Proposals

You should have a **scientific question** that you're attempting to answer with stats/ml.

Assignment 3 Discussion

Which analytical software program did you use for the majority of your analysis

12 responses



Assignment 3 Discussion

What did you find **challenging** about this assignment?

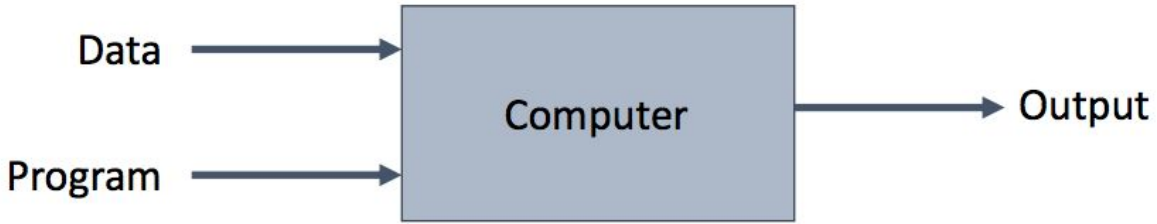
What results did you find **surprising** in this assignment?

What **advice** do you have for future students?

Machine Learning Review

What is ML?

Traditional Programming



Machine Learning

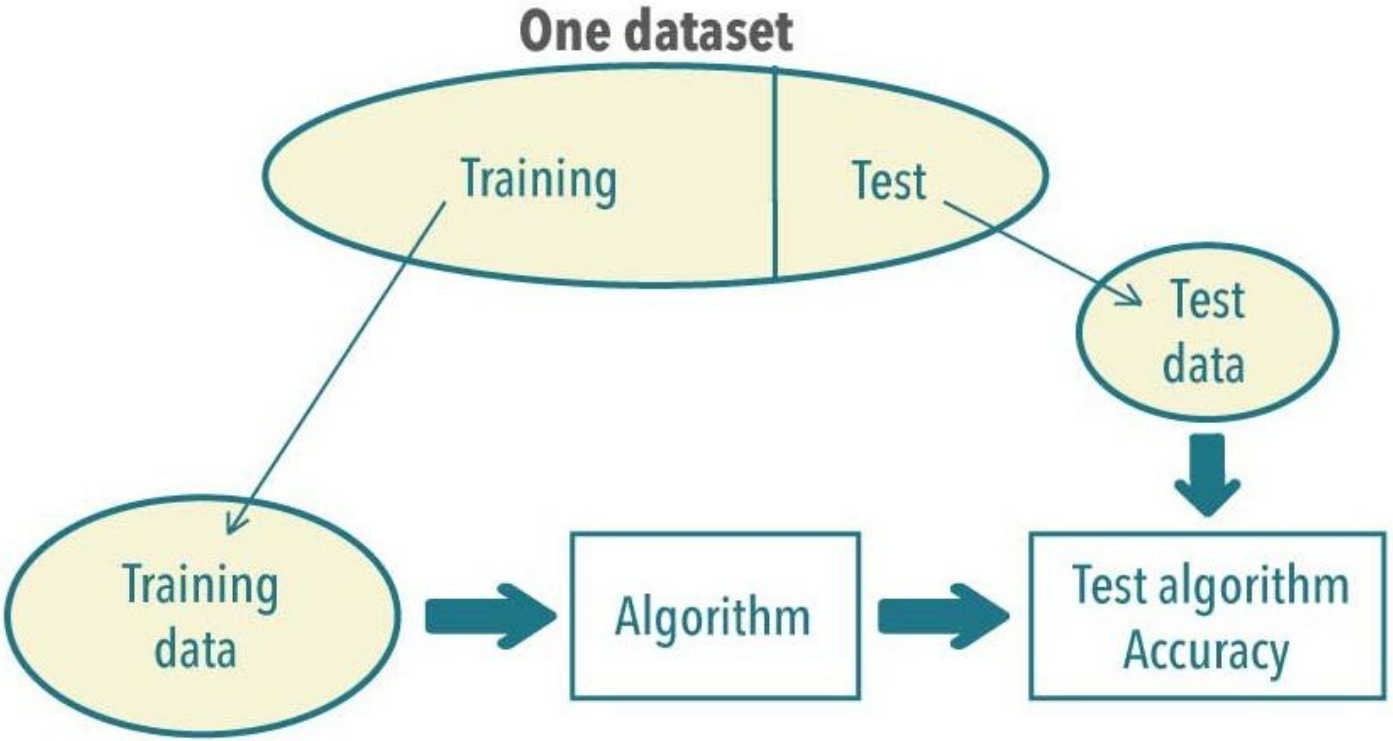
Diagram from Lavi Aulck

Types of Machine Learning

Generally two types of machine learning: **supervised** and **unsupervised**

Key distinction is whether you know the correct answer (**outcome**)

Training data vs. test data



Training and Testing data (source)

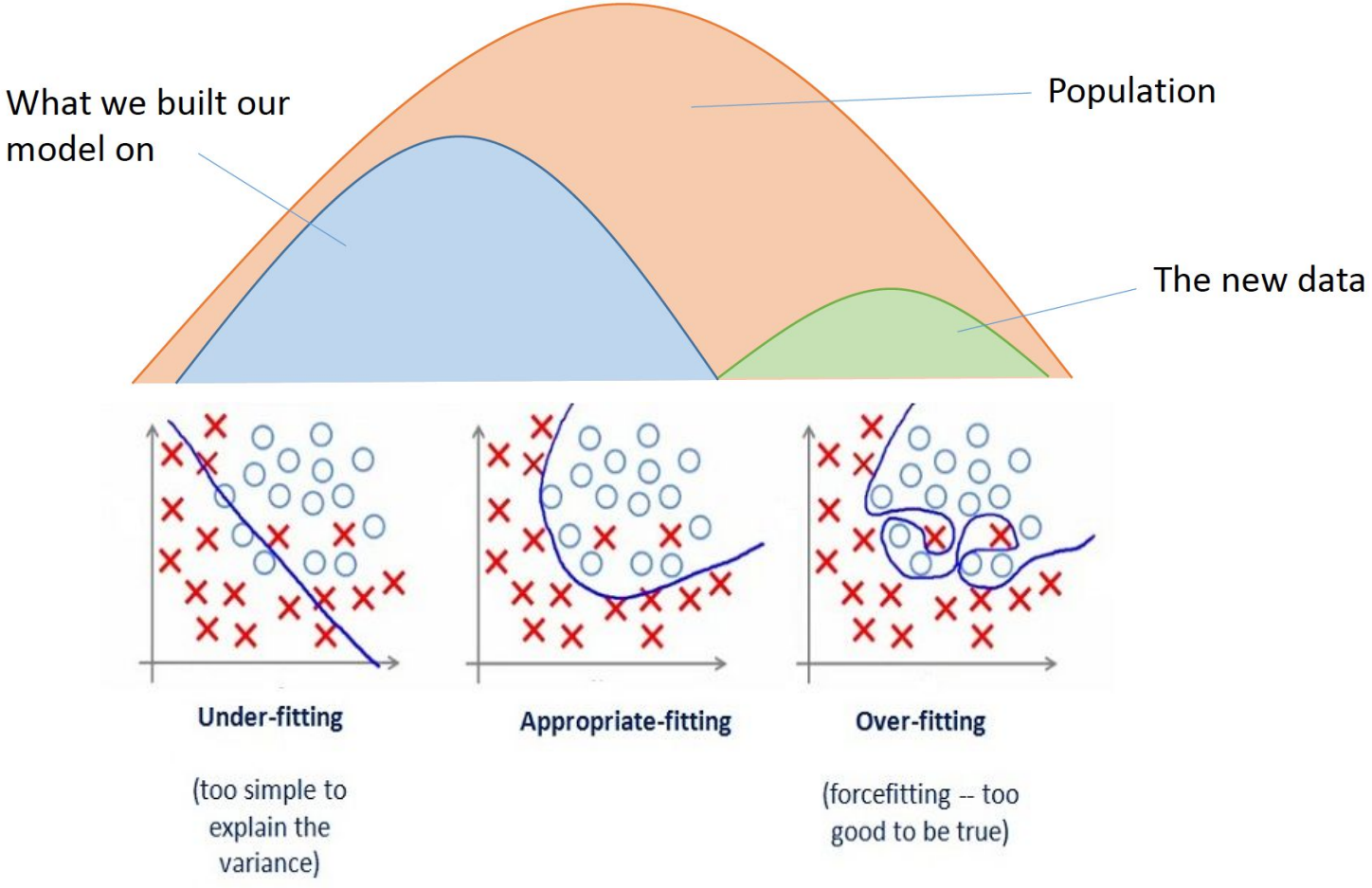
Algorithm Review

How is a decision tree built?

What parameters might we want to control about building the tree?

How does the K Nearest Neighbors algorithm classify an observation?

Can decision trees and KNN be used to predict both continuous and categorical outcomes?



Overfitting (source)

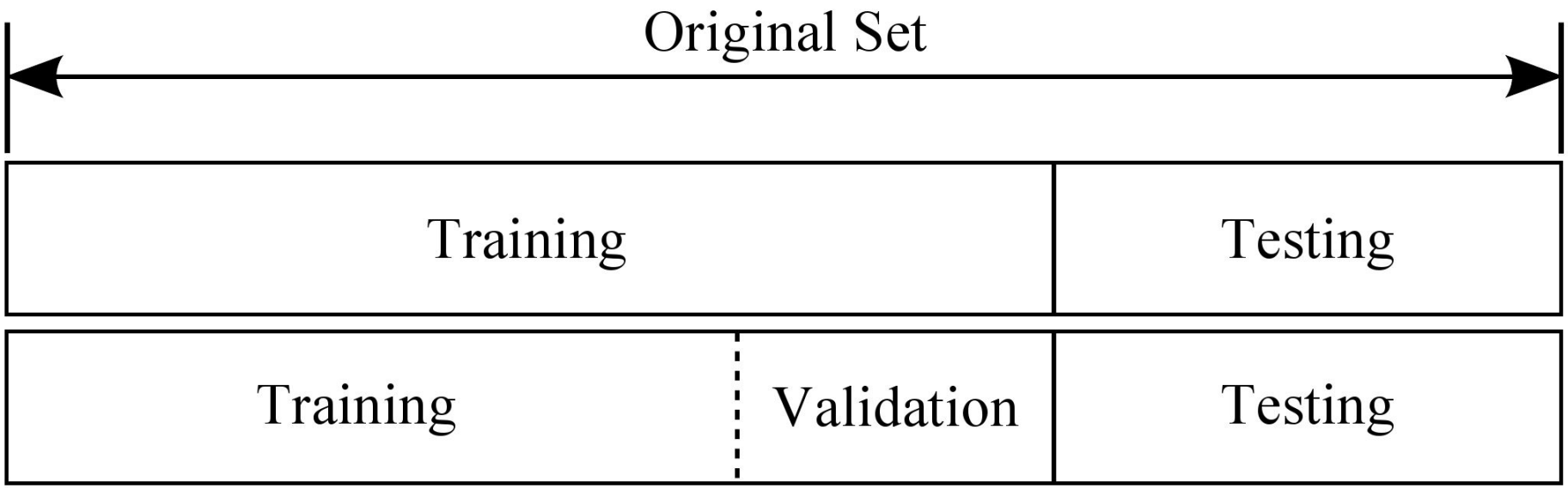
Validating results

We **never look at the test data** until our model is complete

We may want to compare multiple models to one another

We can use some of the training data to assess (**validate**) our models

This is something we may want to repeat to avoid errors due to randomness



Creating validation data (source)

Cross Validation

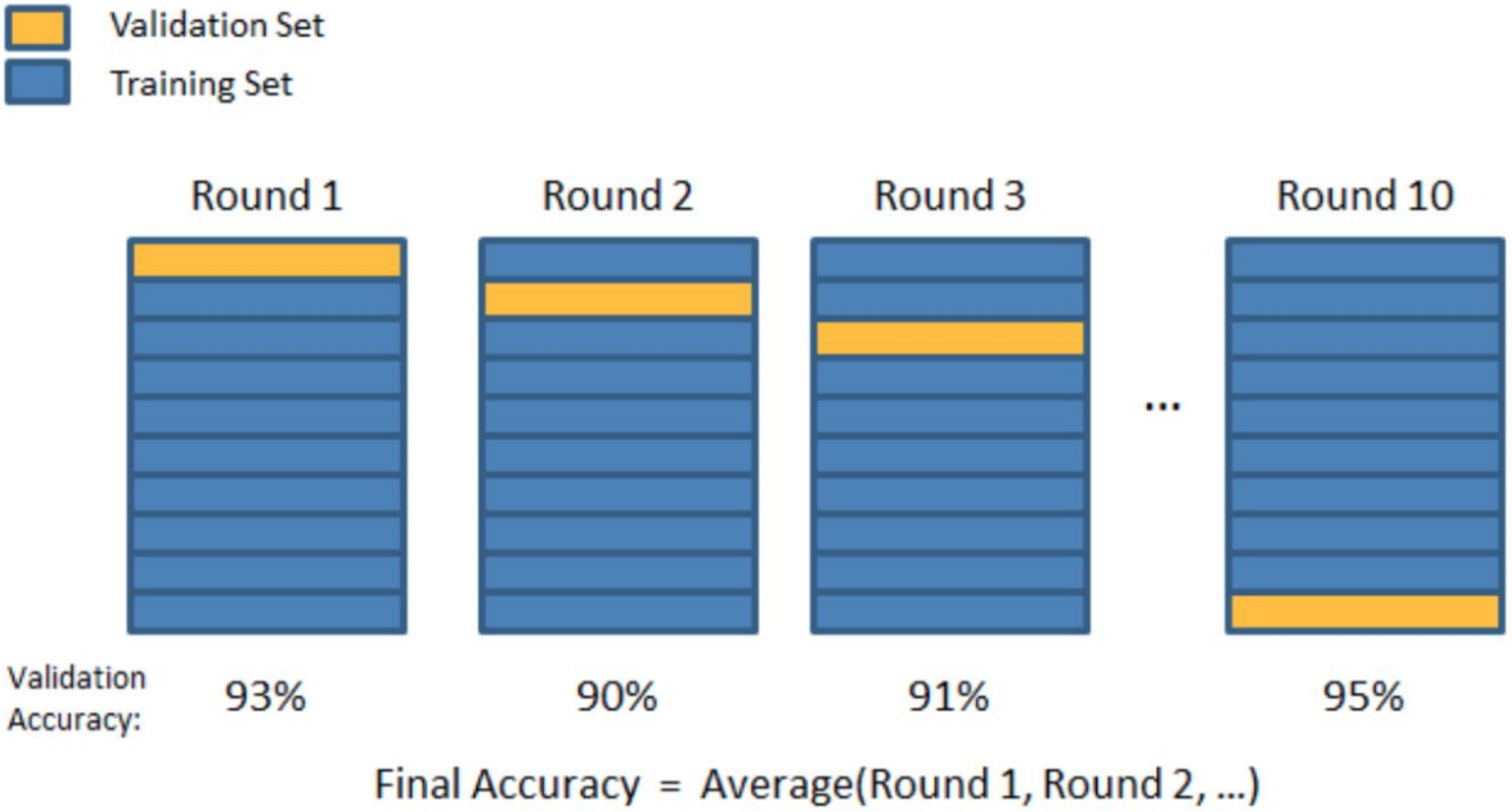
Cross Validation

We currently only use one validation set (a subset of our training data)

- This is subject to variation due to randomness
- Doesn't harness the full potential of our dataset

We can repeat our validation process on different subsets of our training data

This is called **cross validation**



KFold Cross Validation, done for **each model** (source)

Grid Search

Searching for parameters

So far:

- Create testing and training data
- Use cross validation to assess model performance
- Predict on our dataset

We'll want to find the **optimal set** of parameters for creating our models

This is call **tuning** your model (or, to be really fancy, *hyperparameter tuning*)

To tune our model, we'll perform the above steps separately for each parameter set

Normalizing // Scaling Data

Normalizing // Scaling Data

Many algorithms are distance based (KNN)

You'll need to normalize (scale) your data to weight features consistently

Various ways to normalize your data

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$$x_{new} = \frac{x - \mu}{\sigma}$$

Upcoming...

r4-ethics due ***next Tuesday***

Project proposals due ***next Tuesday***

Notebook 6 due **Friday night**

Next Week

- Applied Machine Learning