

Introduction of data mining & machine learning algorithms



Module One

Agenda

- Introduction to machine learning
- Overview of data mining process
- Introduction and setup of the data mining tools (Python)

Review of machine learning

- A machine-learning system is *trained* rather than explicitly programmed.
- It's presented with many examples relevant to a task, and it finds statistical structure in these examples that eventually allows the system to come up with rules for automating the task.
- For instance, if you wished to automate the task of tagging your vacation pictures, you could present a machine-learning system with many examples of pictures already tagged by humans, and the system would learn statistical rules for associating specific pictures to specific tags.

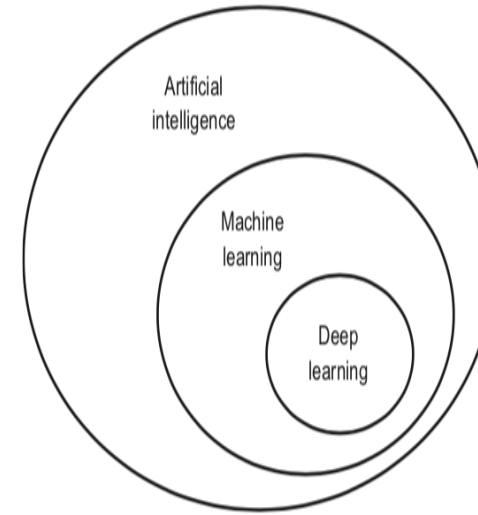


Figure 1.1 Artificial intelligence, machine learning, and deep learning

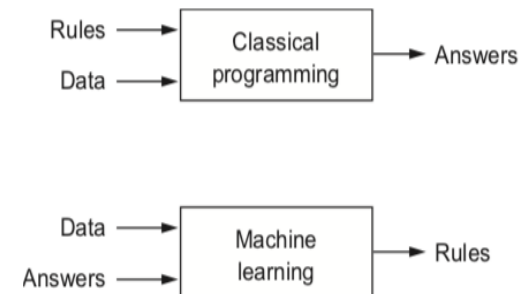


Figure 1.2 Machine learning: a new programming paradigm

Review of machine learning

- machine learning discovers rules to execute a data-processing task, given examples of what's expected.
- To perform machine learning, we need three things:
- ***Input data points***—For instance, if the task is speech recognition, these data points could be sound files of people speaking. If the task is image tagging, they could be pictures.
- ***Examples of the expected output***—In a speech-recognition task, these could be human-generated transcripts of sound files. In an image task, expected outputs could be tags such as “dog,” “cat,” and so on.
- ***A way to measure whether the algorithm is doing a good job***—This is necessary in order to determine the distance between the algorithm's current output and its expected output. **The measurement is used as a feedback signal to adjust the way the algorithm works. This adjustment step is what we call *learning*.**

Review of machine learning

- Let's make this concrete. Consider an x-axis, a y-axis, and some points represented by their coordinates in the (x, y) system, as shown in the figure .
- As you can see, we have a few white points and a few black points. Let's say we want to develop an algorithm that can take the coordinates (x, y) of a point and output whether that point is likely to be black or to be white. In this case:
- The inputs are the coordinates of our points.
- The expected outputs are the colours of our points.
- A way to measure whether our algorithm is doing a good job could be, for instance, the percentage of points that are being correctly classified.

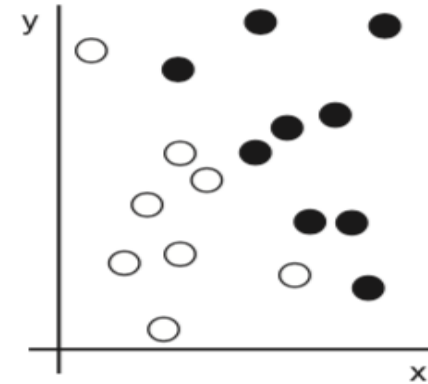
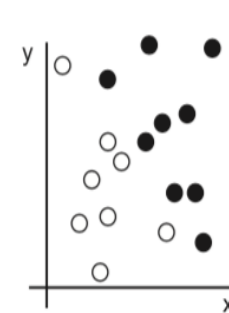


Figure 1.3
Some sample data

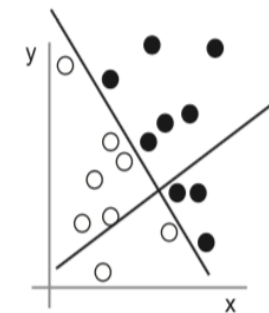
Review of machine learning

- What we need here is a new representation of our data that cleanly separates the white points from the black points.
- One transformation we could use, among many other possibilities, would be a coordinate change, illustrated in the figure.
- All machine-learning algorithms consist of automatically finding such transformations that turn data into more useful representations for a given task.
- These operations can be coordinate changes, as you just saw, or linear projections, translations, nonlinear operations and so on

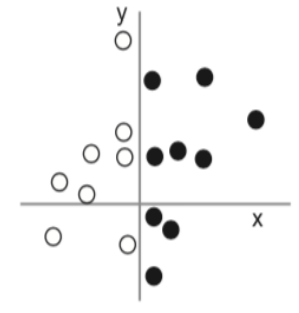
1: Raw data



2: Coordinate change



3: Better representation

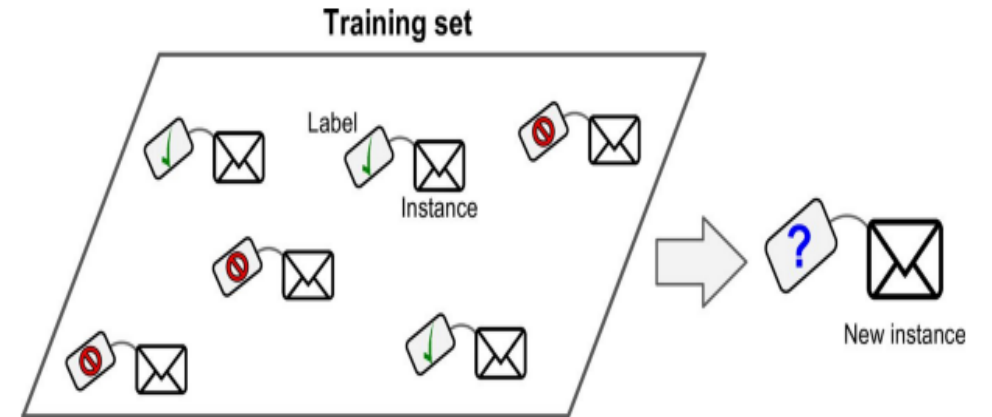


Major types of machine learning algorithms

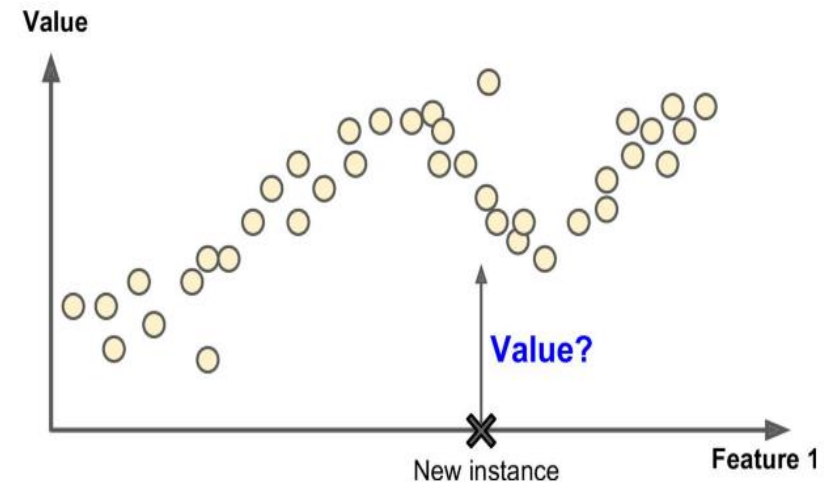
- Supervised learning
- Unsupervised learning

Supervised learning

- There is label field in the training dataset
- Examples are:
 - KNN
 - Decision tree
 - Random forest
 - Regression
 - Logistic regression
 - ANN
 - Deep Neural network
 - SVM



Classification techniques



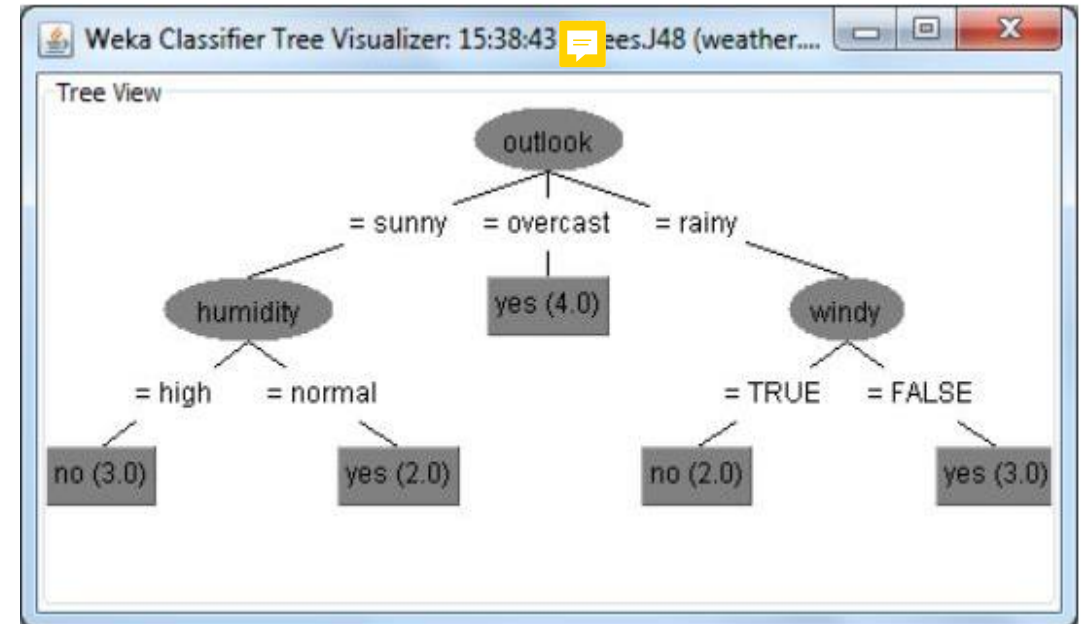
Regression techniques

Example of classification: Decision tree algorithm

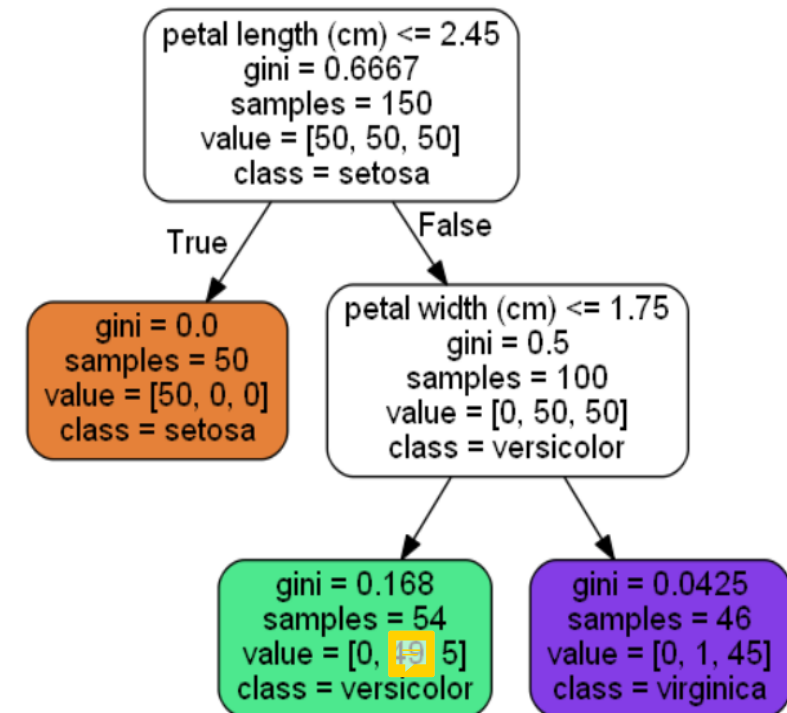
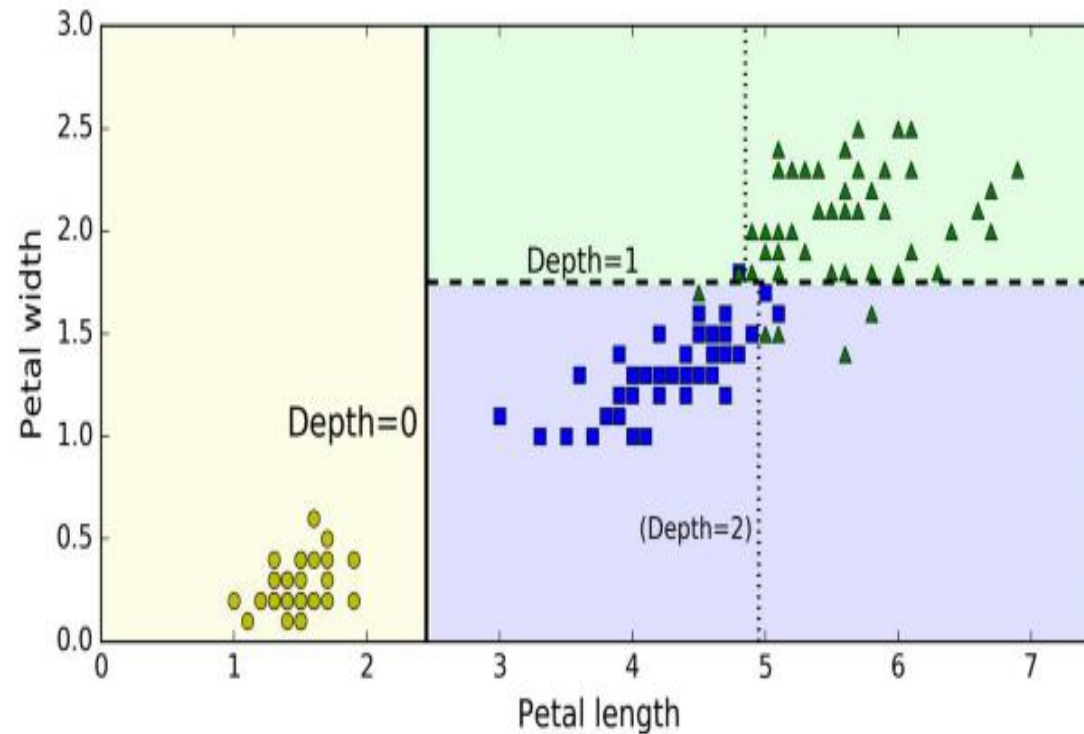
Training dataset

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



Decision tree

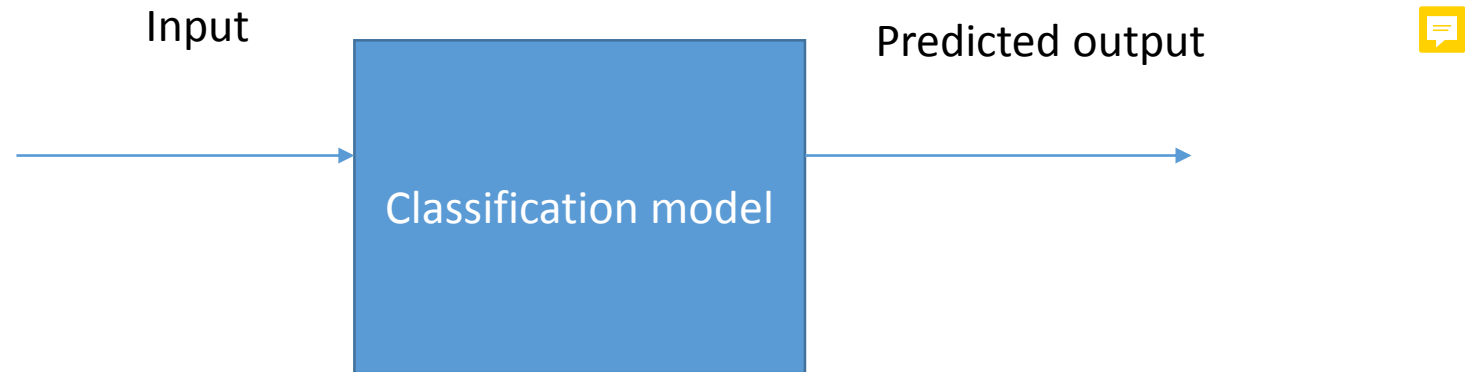


Decision tree decision boundaries

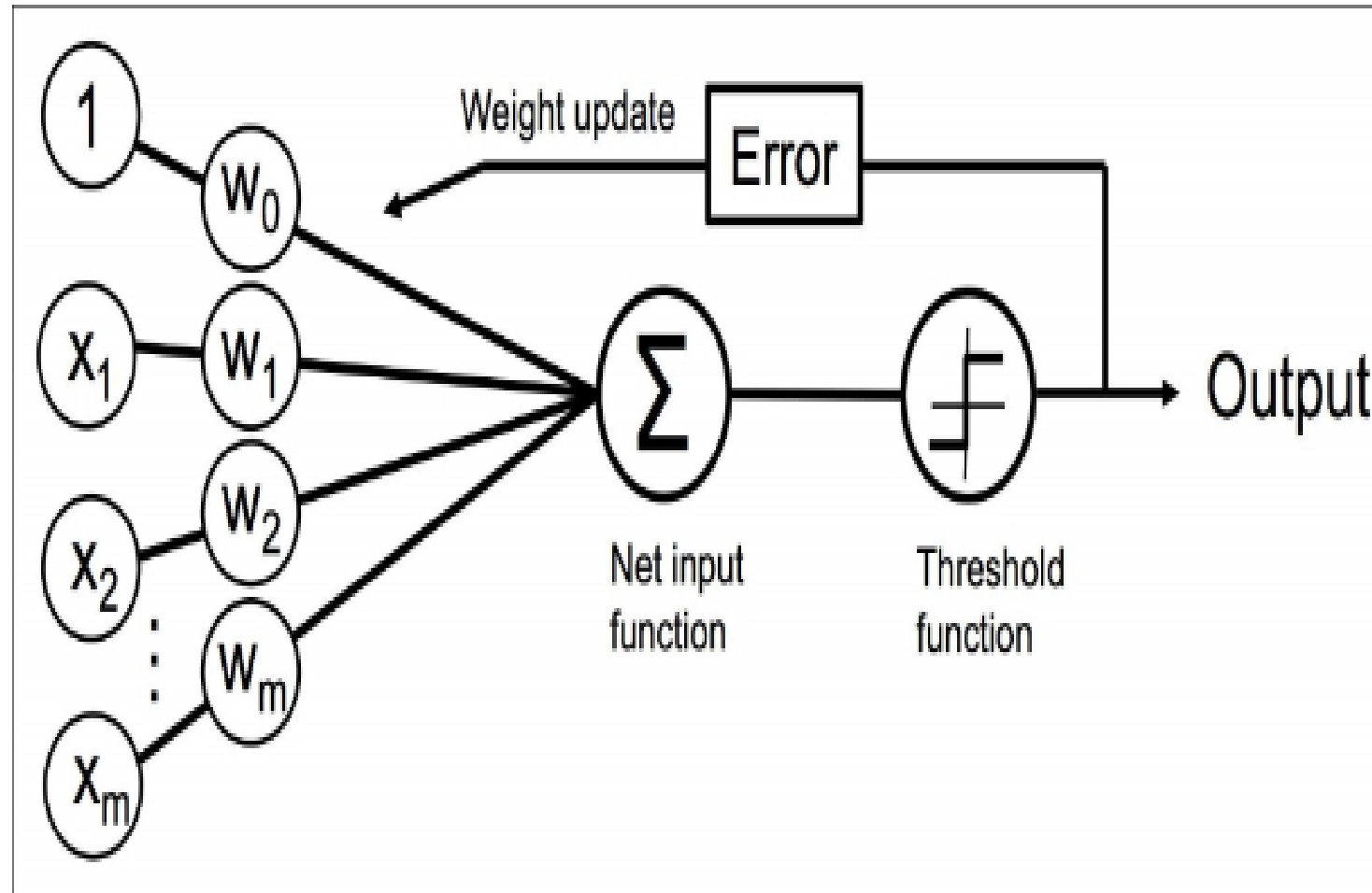


Example of classification modeling

Default 	Income 	Age	XXXX
Y	20000	44	XXXX
N	30000	30	XXX
N	40000	38	XXXX
XXXX	XXXX		

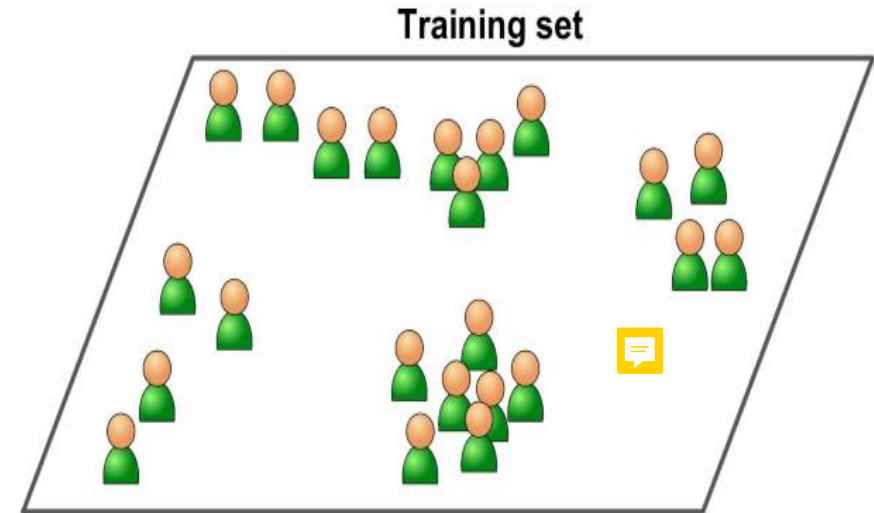


Neural network



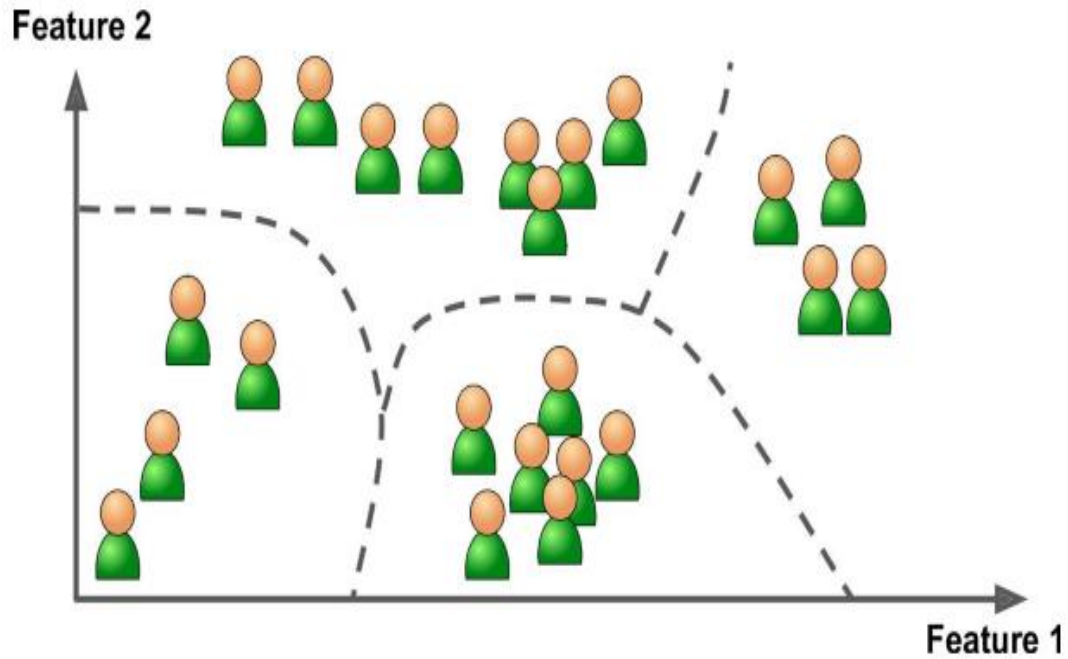
Unsupervised learning-1/2

- There are no labels assigned in the training dataset. The system tries to learn within a teacher.
- Examples are:
 - Clustering techniques (e.g. KMeans/Hierarchical clustering analysis)
 - Principle components analysis(PCA)
 - Association
 - Anomaly detection

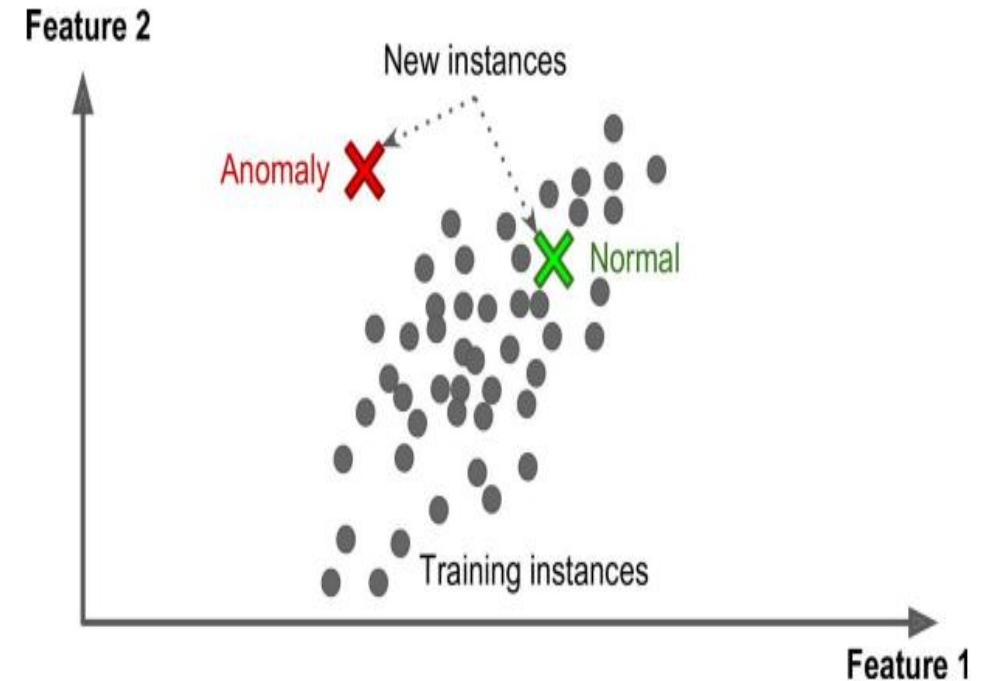


An unlabeled training dataset for unsupervised learning

Unsupervised learning-2/2



Clustering techniques



Anomaly detection techniques

Example of Association

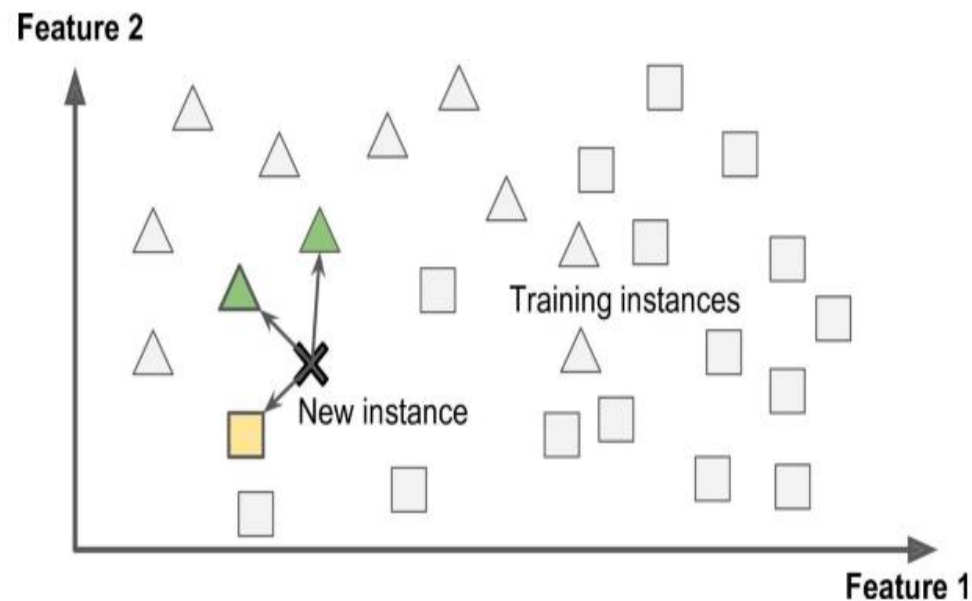
	Transactions List			
1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

With Association algorithm it try to answer the following Question:
What could be the buying pattern of the users
In the supermarket learnt from the transaction list ?

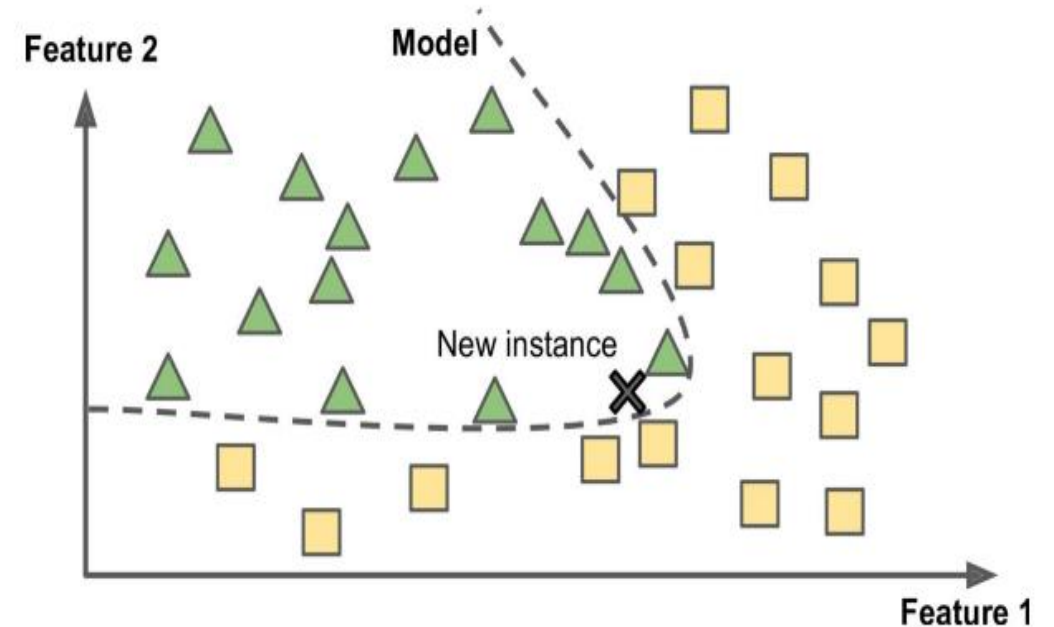
Instance based or model based learning

- One more way to categorize Machine Learning systems is by how they generalize.
- Most Machine Learning tasks are about making predictions. This means that given a number of training examples, the system needs to be able to generalize to examples it has never seen before.
- Having a good performance measure on the training data is good, but insufficient; the true goal is to perform well on new instances.
- There are two main approaches to generalization: instance-based learning and model-based learning.

Instance based or model based learning



Instance based modeling



Model based modeling

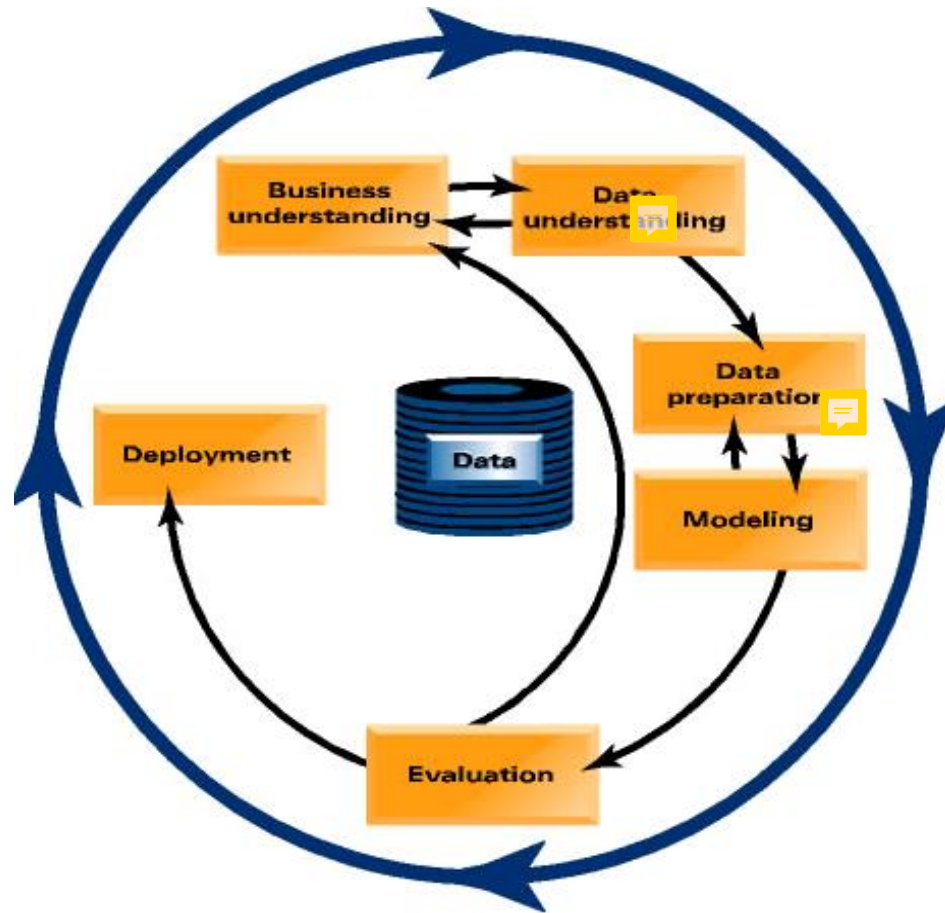
Main challenges of machine learning

- Data related issues
 - Insufficient quantity of data
 - Nonrepresentative training data
 - Poor quality data
 - Outliners
 - Missing values->Impute missing values
- Algorithm related issues
 - Overfitting
 - Underfitting
 - Need to evaluate the performance

Machine learning project checklist

- This checklist can guide you through your Machine Learning projects.
- There are eight main steps:
- Frame the problem and look at the big picture.
- Get the data.
- Explore the data to gain insights.
- Prepare the data to better expose the underlying data patterns to Machine Learning algorithms.
- Explore many different models and short-list the best ones.
- Fine-tune your models and combine them into a great solution.
- Present your solution.
- Launch, monitor, and maintain your system.

CRISP-DM: Overview



- **Data Mining methodology**
- **Process Model**
- **For anyone**
- **Provides a complete blueprint**
- **Life cycle: 6 phases**

CRISP-DM: the six phases (1/2)

- Business/research understanding phase
 - First, clearly enunciate the project objectives and requirements in terms of the business or research unit as a whole
 - Then, translate these goals and restrictions into the formulation of a data mining problem definition
 - Example: Develop a classification model that will maximize profits for direct-mail marketing
- Data understanding phase
 - First collect the data
 - Then, use exploratory data analysis to familiarize yourself with the data and discover initial insights
 - Evaluate the quality of the data
- Data preparation phase
 - This labor intensive phase covers all aspects of preparing the final data set, which shall be used for subsequent phases, from the initial, raw and dirty data
 - Perform transformation on certain variables if needed
 - Clean the raw data so that it is ready for the modeling work

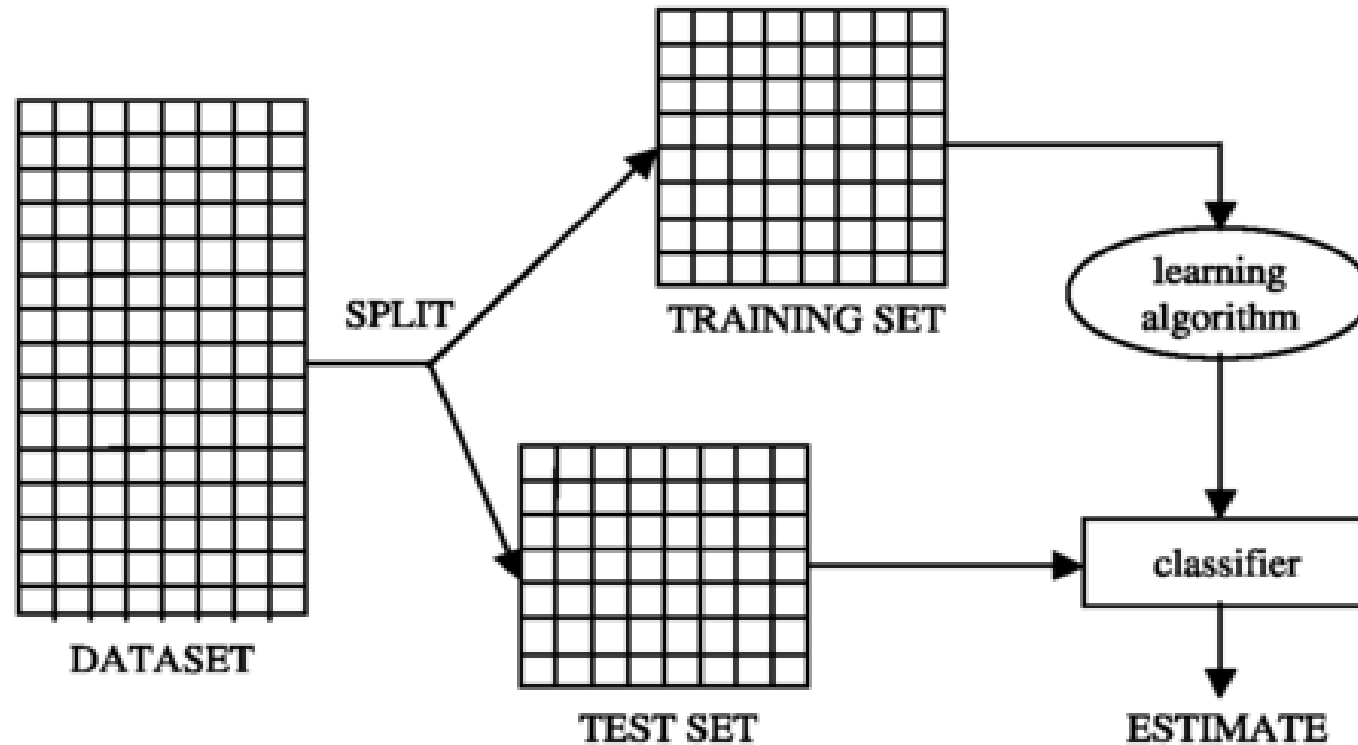
CRISP-DM: the six phases (2/2)

- Modeling phase
 - Select and apply appropriate modeling techniques
 - Calibrate model settings to optimize results
 - Often, several different techniques may be applied for the same data mining problem
 - May require looping back to data preparation phase, in order to bring the form of data into line with the specific requirements of a particular data mining technique
- Evaluation phase
 - The modeling phase has delivered one or more models. These models must be evaluated for quality and effectiveness, before we deploy them for use in the field
- Deployment phase
 - Need to make use of the models
 - Example of a simple deployment: Generate a report

Common use of data mining

- To describe the data pattern
- To model and generalize the criteria of the output cases
- To predict/forecast future output cases

Evaluation of the performance of the model



To measure the performance of the classification model

Definition of true and false positives and negatives

		Predicted class		Total instances
		+	-	
Actual Class	+	TP	FN	P
	-	FP	TN	N

TP: True positives

The number of positive instances that are classified as positive

FP: False positives

The number of negative instances that are classified as positive

FN: False negatives

The number of positive instances that are classified as negatives

TN: True negatives

The number of negative instances that are classified as negatives

P=TP+FN

The total number of positive instances

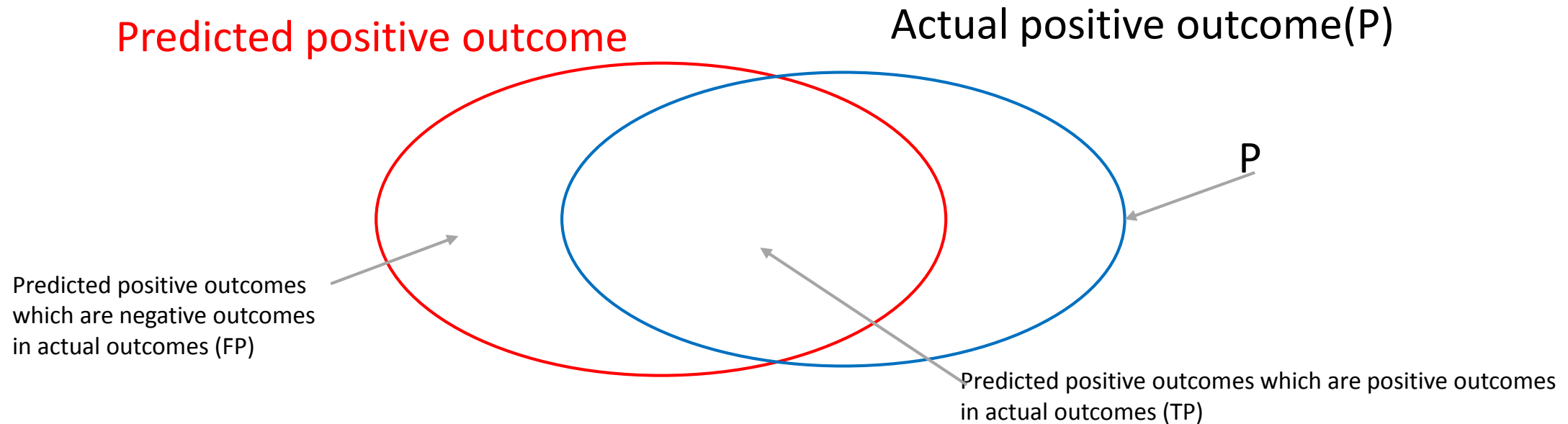
N=FP+TN

The total number of negative instances

Some performance measure for a classifier

Name	Formula	Description
True positive rate (or sensitivity or Recall)	TP/P	The proportion of positive instances that are correctly classified as positive. It gives the probability of getting a positive classification when the true outcome is positive.
False positive rate	FP/N	The proportion of negative instances that are erroneously classified as positive.
True negative rate (or specificity)	TN/N	The proportion of negative instances that are correctly classified as negative. It gives the probability of getting a negative classification when the true outcome is negative.
False negative rate	FN/P	The proportion of positive instances that are erroneously classified as negative
Positive predictive value (or precision)	$TP/(TP+FP)$	Proportion of instances classified as positive that are really positive. It gives the probability that an observation with a positive classification is correctly identified as positive
Accuracy	$(TP+TN)/(P+N)$	Proportion of observations correctly identified

Definition of Precision and True Positive Rate



Precision:

To measure the ability to generate the predicted positive outcomes which are really positive in actual outcomes
 $=TP/(TP+FP)$

True positive rate(recall):

To measure how many of the actual positive outcomes can be detected by the predicted positive outcomes
 $=TP/P$

Quiz

- What is supervised learning ?
- Which of the followings are supervised learning
 1. Decision tree algorithms
 2. KNN
 3. Regression
 4. Clustering
 5. Neural network

Exercise 1

- For each of the following meetings, explain which phase in the CRISP-DM process is represented:
 - Managers want to know by next week whether deployment will take place. Therefore, analysts meet to discuss how useful and accurate their model is
 - The data mining project manager meets with the data warehousing manager to discuss how the data will be collected
 - The data mining consultant meets with the vice president for marketing, who says that he would like to move forward with customer relationship management
 - The data mining project manager meets with the production line supervisor to discuss the implementation of changes and improvement

Agenda

- Introduction to machine learning
- Overview of data mining process
- **Introduction and setup of the data mining tools (Python)**

Brief History of Python

- Invented in the Netherlands, early 90s by Guido van Rossum
- Named after Monty Python
- Open sourced from the beginning
- Considered a scripting language, but is much more
- Scalable, object oriented and functional from the beginning
- Used by Google from the beginning
- Increasingly popular

Python

Download Python

- <https://www.python.org/downloads/>
- IDE for Python
- Default: IDLE
- Others: Pycharm
<https://www.jetbrains.com/pycharm/download/#section=windows>
- Jupyter notebook
<https://www.anaconda.com/download/>

Common Python's packages-1/2

- Pandas: Data Analysis Library
- Django: High Level Web Framework
- NumPy: Package for numerical computing
- SciPy: Routines for numerical Integration and Optimization

Common Python's packages-2/2

- Matplotlib: 2-D plotting
- Flask: Microframework
- Mechanize: Programatic Web Browsring
- BeautifulSoup: Html Scrapping
- PyQt: GUI building

Python vs R

- Python
 - Open source with good user community
 - Support object oriented programming (OOP)
 - Speed
 - Lots of useful packages available (e.g. Pandas/NumPy/SciPy/Matplotlib etc)
 - Easy to write codes
- R
 - Open source with good user community
 - Emphases on statistical analysis
 - Lots of packages support for different purposes (e.g. ggplot2/quantmod/rpart/rvest)
- Use **BOTH** to leverage each other in data analytic work