SELECTED TOPICS IN ENGINEERING

INTR. TO PROG. FOR DATA SCIENCE
ENGR 350

Tuesday-Thursday 10:00-12:45
ENG B05
2019 Summer

Dr. Banu Yobaş

Structured data

- Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed.
- This includes defining what fields of data will be stored and how that data will be stored:
  - data type (numeric, currency, alphabetic, name, date, address)
  - any restrictions on the data input (number of characters; restricted to certain terms such as Mr., Ms. or Dr.; M or F).
- well organized,
- follows a consistent order,
- can be readily accessed and understood by a person or a computer program.

Structured data

- Structured data has the advantage of being easily
  - entered,
  - stored,
  - queried and
  - analyzed. A
- usually stored in well-defined schemas such as databases.
- It's usually tabular, with columns and rows that clearly define its attributes.
- Relational databases and spreadsheets use structured data
- A classic example of structured data is an Excel spreadsheet with labeled columns.

Unstructured data

- tends to be free-form,
- non-tabular,
- dispersed, and
- not easily retrievable
- such data requires deliberate intervention to make sense of it.
- It’s hard to categorize the content of unstructured data.
- all those things that can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

Unstructured data

- The content of unstructured data is hard to work with or make sense of programmatically.
- Computer programs cannot analyze or generate reports on such data, simply because
  - it lacks structure,
  - has no underlying dominant characteristic, and
  - individual items of data have no common ground.

Unstructured data

- Unstructured data requires more work to make it useful, so it gets more attention — thus tends to consume more time.

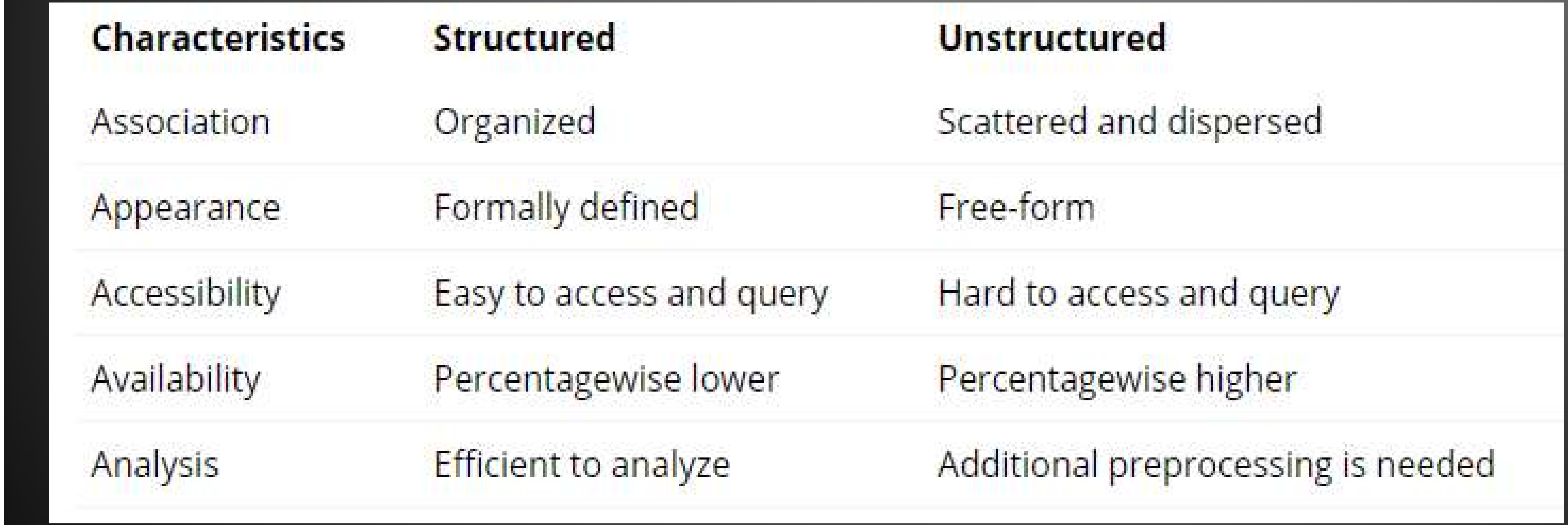
Semi-structured data

- a cross between the two.
- It is a type of structured data, but lacks the strict data model structure.
- With semi-structured data, tags or other types of markers are used to identify certain elements within the data, but the data doesn't have a rigid structure.
- For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text.
- Emails have the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments.
- Photos or other graphics can be tagged with keywords such as the creator, date, location and keywords, making it possible to organize and locate graphics.
- XML and other markup languages are often used to manage semi-structured data.

Structured vs Unstructured

- Don’t underestimate the importance of structured data and the power it brings to your analysis. It’s far more efficient to analyze structured data than to analyze unstructured data.
- Unstructured data can also be costly to preprocess for analysis as you’re building a predictive analytics project. The selection of relevant data, its cleansing, and subsequent transformations can be lengthy and tedious.

Structured vs Unstructured



Characteristics	Structured	Unstructured
Association	Organized	Scattered and dispersed
Appearance	Formally defined	Free-form
Accessibility	Easy to access and query	Hard to access and query
Availability	Percentagewise lower	Percentagewise higher
Analysis	Efficient to analyze	Additional preprocessing is needed

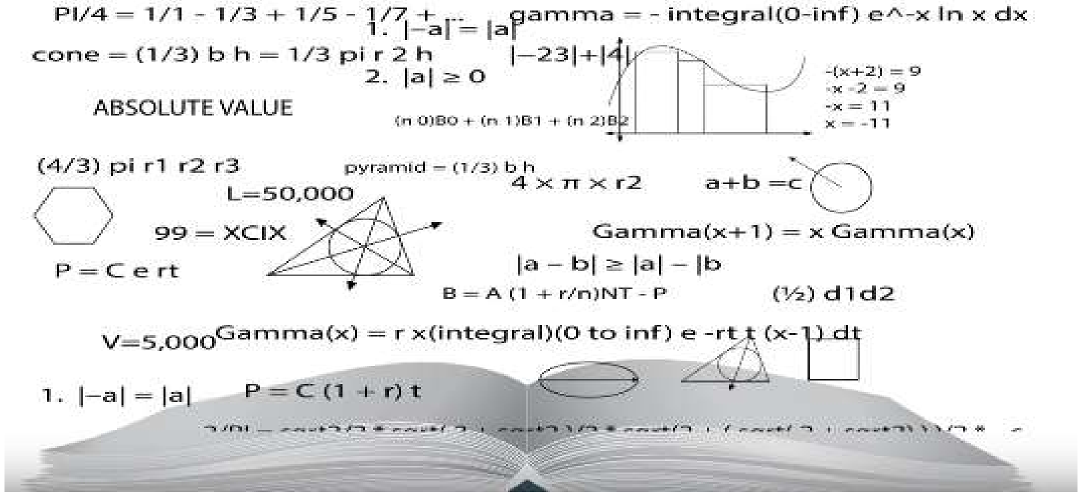
For a successful predictive analytics project, both your structured and unstructured data must be combined in a logical format that can be analyzed.

Data Science

Making better decisions

Data Science

- Dealing with unstructured and structured data,
- Data Science is a field that comprises of everything that related to
  - data cleansing,
  - preparation, and
  - analysis.



Interdisciplinary

A combination of

- statistics,
- mathematics,
- programming,
- problem-solving,

Capturing data in ingenious ways,
The ability to look at things differently, and
the activity of

cleansing,
preparing and
aligning the data.

Data Science & Python

- It is the umbrella of techniques used when trying to extract insights and information from data.

Python

Extended With It's Growing Number Of Data Analytics Libraries

NumPy
SciPy
Pandas
SM StatsModels
Statistics in Python
learn
machine learning in Python

Applications of Data Science

Internet search

Digital Advertisements: The entire digital marketing spectrum uses the data science algorithms - from display banners to digital billboards. This is the mean reason for digital ads getting higher CTR than traditional advertisements.

Recommender systems: The recommender systems not only make it easy to find relevant products from billions of products available but also adds a lot to user-experience. A lot of companies use this system to promote their products and suggestions in accordance with the user's demands and relevance of information. The recommendations are based on the user's previous search results.

Data Analytics:

- The science of examining raw data with the purpose of drawing conclusions about that information.
- Involves applying an algorithmic or mechanical process to derive insights. For example, running through a number of data sets to look for meaningful correlations between each other.
- is used in a number of industries to allow the organizations and companies to make better decisions as well as verify and disprove existing theories or models.

Data Analytics

- The focus lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.

Applications of Data Analysis

- Healthcare: The main challenge for hospitals with cost pressures tightens is to treat as many patients as they can efficiently, keeping in mind the improvement of the quality of care. Instrument and machine data is being used increasingly to track as well as optimize patient flow, treatment, and equipment used in the hospitals. It is estimated that there will be a 1% efficiency gain that could yield more than $63 billion in the global healthcare savings.

Applications of Data Analysis

- Travel: Data analytics is able to optimize the buying experience through the mobile/ weblog and the social media data analysis. Travel sights can gain insights into the customer's desires and preferences. Products can be up-sold by correlating the current sales to the subsequent browsing increase browse-to-buy conversions via customized packages and offers. Personalized travel recommendations can also be delivered by data analytics based on social media data.

Applications of Data Analysis

- Gaming: Data Analytics helps in collecting data to optimize and spend within as well as across games. Game companies gain insight into the dislikes, the relationships, and the likes of the users.

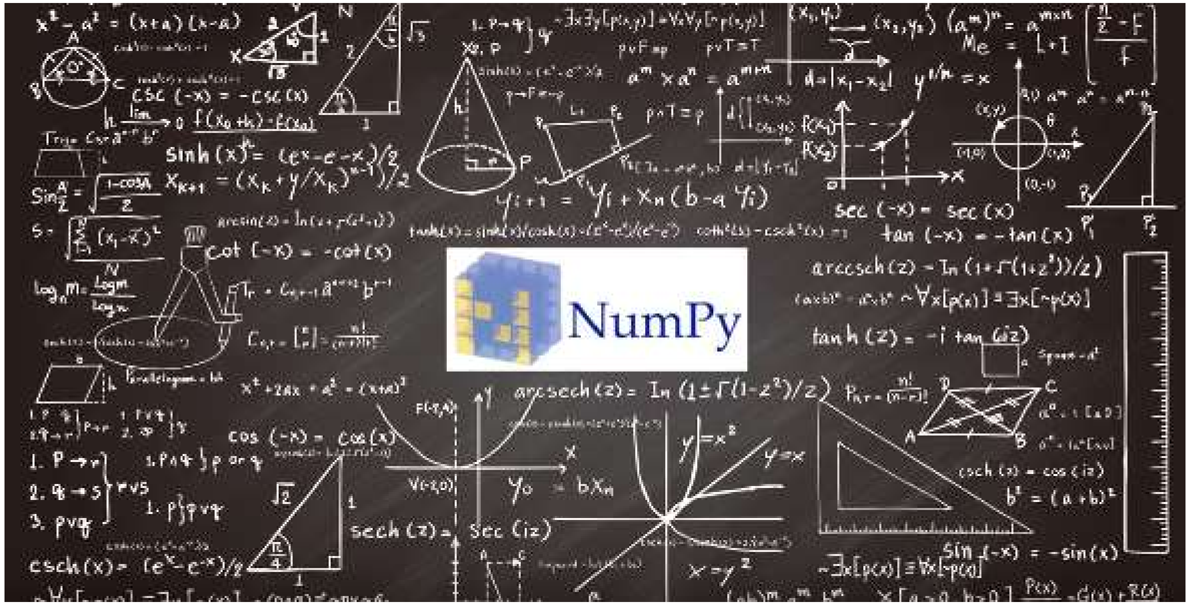
Applications of Data Analysis

- Energy Management: Most firms are using data analytics for energy management, including smart-grid management, energy optimization, energy distribution, and building automation in utility companies. The application here is centered on the controlling and monitoring of network devices, dispatch crews, and manage service outages. Utilities are given the ability to integrate millions of data points in the network performance and lets the engineers use the analytics to monitor the network.

Good to know modules

- numpy, scipy: basics for almost everything
- Matplotlib, a Python 2D plobng library
  - http://matplotlib.org
- NLTK, Natual Language Toolkit
  - http://www.nltk.org
- Pandas, Python Data Analysis Library
  - http://pandas.pydata.org
- mrjob, route to wriGng MapReduce jobs
  - https://pythonhosted.org/mrjob/
- IPython, InteracGve console with IDE--like features
  - http://ipython.org
- Scikit-Learn, ML resource and library
  - http://scikit-learn.org/dev/index.html
- Theano/Pylearn2, deep learning
  - http://deeplearning.net/soXware/theano/
  - http://deeplearning.net/soXware/pylearn2/
- More: mlpy, PyBrain, Orange, Scrapy, ...

Numerical Python – Numpy



Science Packages & Libraries

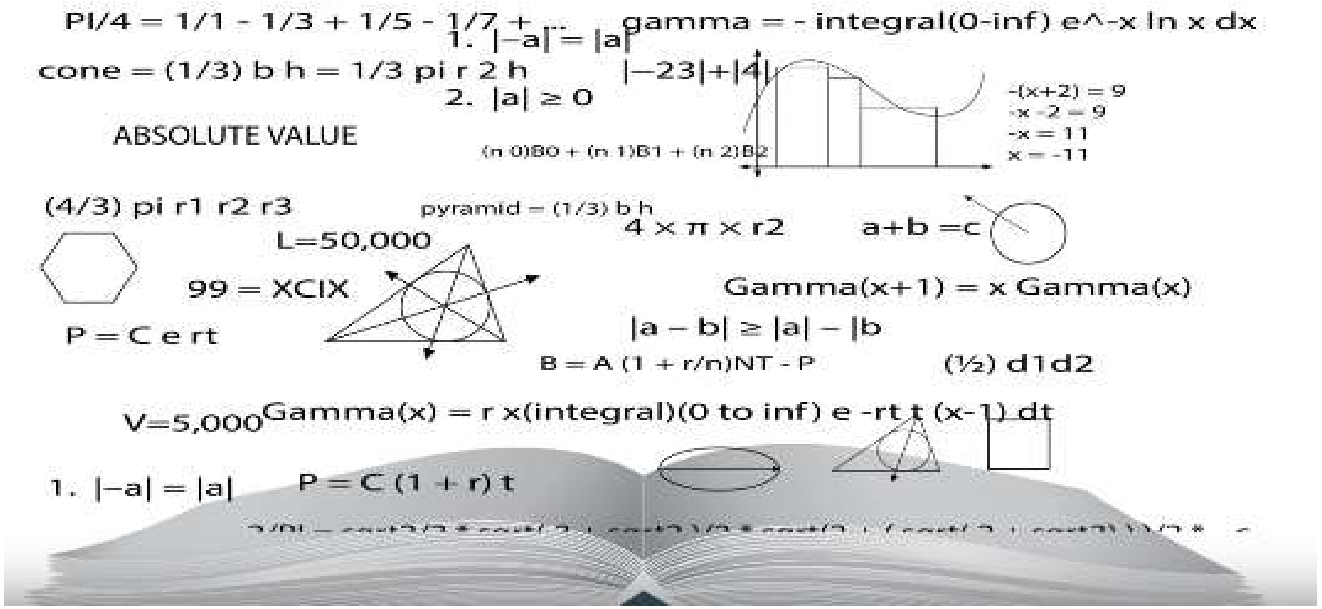
- Optimization,
- Integration,
- Linear algebra,
- Fast fourier transformation,
- Signal & image processing,

…

What is Scikit-learn

- A Python Machine Learning Library
- Focused on modeling data
- Developed by David Cournapeau as a Google summer of code project in 2007.
- First public release (v0.1 beta) published in late January 2010.
- Now has more than 30 acGve contributors and has had paid sponsorship from INRIA, Google, Tinyclues and the Python SoXware FoundaGon.
- The library is built upon the SciPy that must be installed before you can use scikitlearn.

Science Packages&Libraries

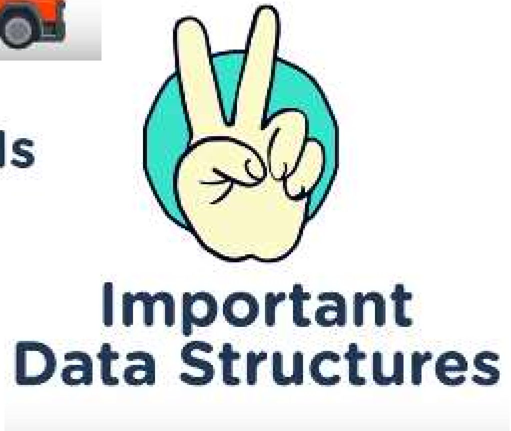


- Optimization,
- Integration,
- Linear algebra,
- Fast fourier transformation,
- Signal & image processing, …

Data Structures

Pandas built on top of NumPy, adds data frames which offer critical data analysis functionality and features

Different From Other Standard Arrays Is Its Ability To Index All Its Elements



Data Structures

2 Dimensional Series
With Indexes For
Both Rows & Columns

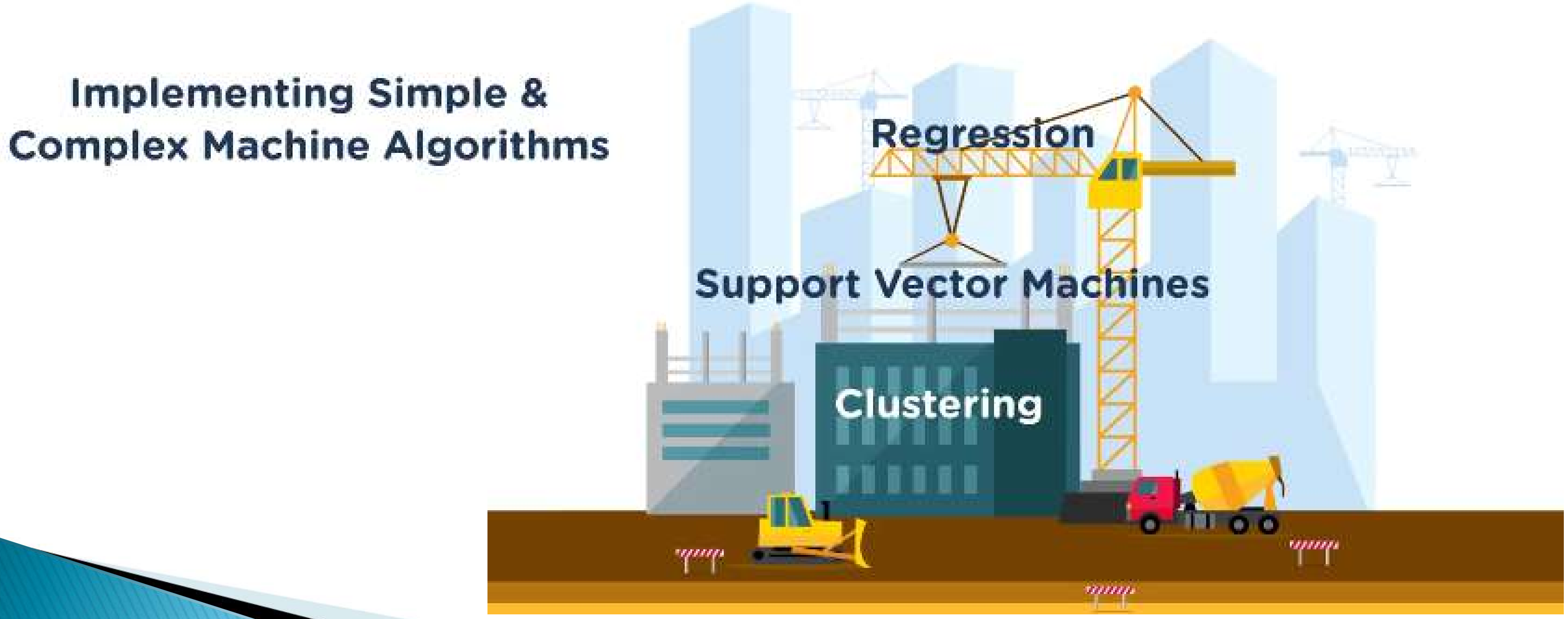
Extracting Excel
Or An SQL Table's Data
Into Python

Executed On Series and Dataframes Like
Average, Sum, Concatenate, Group By & Order By
Among Others

Machine Learning

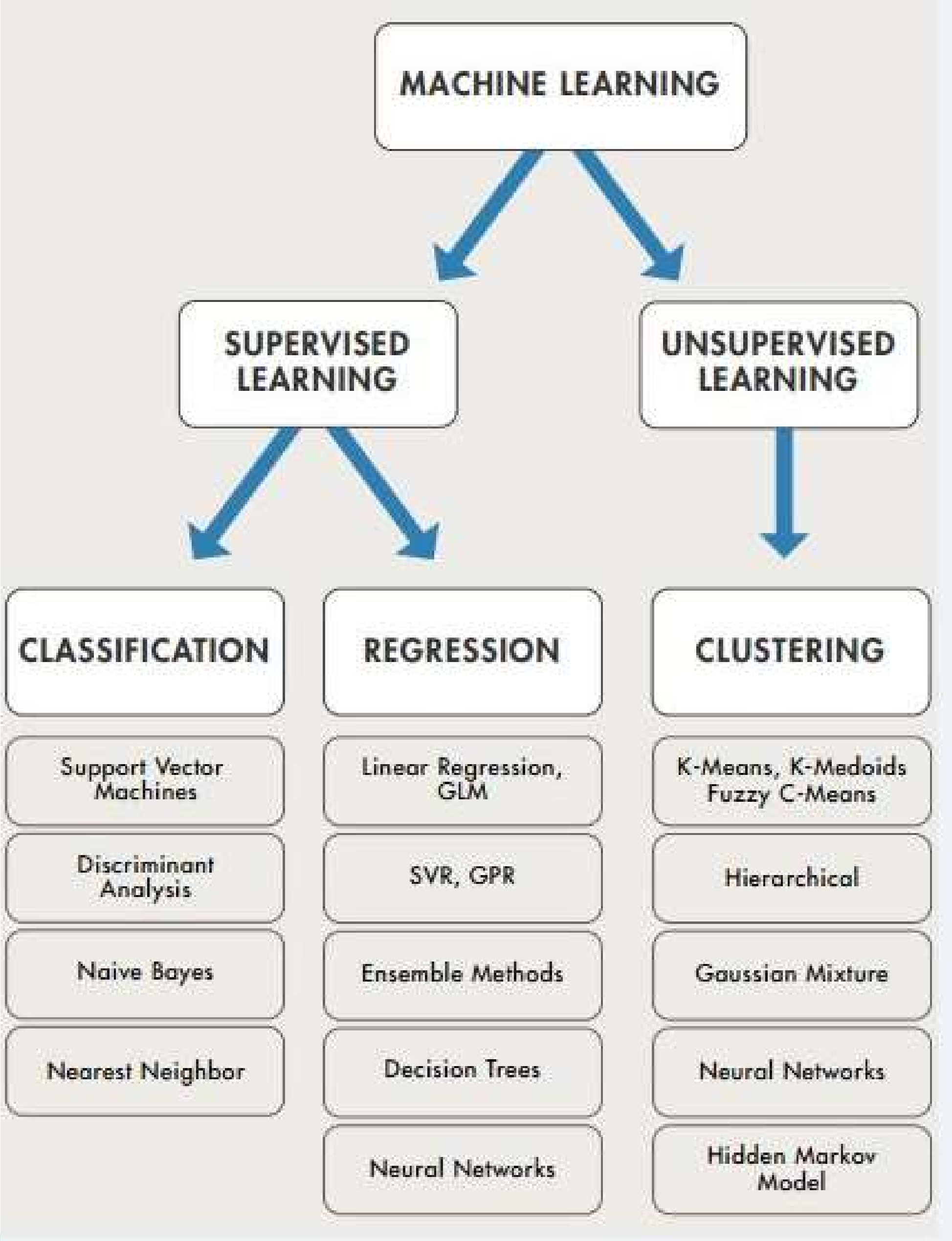
scikits learn

machine learning in Python



Machine Learning

- Organisations seek to make better decisions by examining their data with an aim to discovering and/or drawing conclusions about the information contained within.

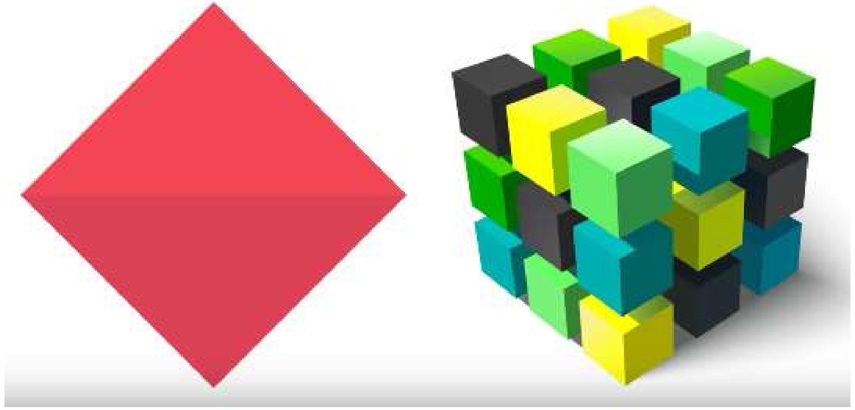

MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISED LEARNING
CLASSIFICATION
REGRESSION
CLUSTERING
Support Vector Machines
Discriminant Analysis
Naive Bayes
Nearest Neighbor
Linear Regression, GLM
SVR, GPR
Ensemble Methods
Decision Trees
Neural Networks
K-Means, K-Medoids Fuzzy C-Means
Hierarchical
Gaussian Mixture
Neural Networks
Hidden Markov Model

Statistics in Python

SM StatsModels

Statistics in Python

Data Visualization



Visualization

Histograms

Power Spectra

Bar Charts

Pie Charts

Box Plots

Scatter Plots

Visualization

Advanced Visualisations,
Matplotlib Requires Too Much Effort

SEABOURN®

Create Complicated Plots With Ease

Heatmap uses visualization
which can create with Seabourn
using one line of code

Seaborn module

- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- https://seaborn.pydata.org/
- See tutorials

python
R
Scala

Jupyter notebook

Install the jupyter notebook
Refer to previous lecture notes and guides for installation

Deep learning libraries

TensorFlow™ theano

Decision Trees (DTs)

- are a non-parametric supervised learning method used for classification and regression.
- The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Some advantages of decision trees

- Simple to understand and to interpret. Trees can be visualised.
- Requires little data preparation. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data. Other techniques are usually specialised in analysing datasets that have only one type of variable.

Some advantages of decision trees

- Able to handle multi-output problems.
- Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret.
- Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.
- Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

The disadvantages of DT

- Decision-tree learners can create over-complex trees that do not generalise the data well. This is called **overfitting**. Mechanisms such as pruning (not currently supported), setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.

The disadvantages of DT

- The problem of learning an optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees in an ensemble learner, where the features and samples are randomly sampled with replacement.
- There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems.
- Decision tree learners create *biased trees* if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.