ELECTRONIC PUBLISHING
AND DIGITAL STORYTELLING
Lesson 6

DATA ANALYSIS

TABLE OF CONTENTS

ELECTRONIC PUBLISHING
AND DIGITAL STORYTELLING
Lesson 6

DATA SCIENCE TASKS

DATA SCIENCE

“Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets.”

The objective is to improve decision-making processes by means of data analysis.

Kelleher D., Tierney B. Data science. MIT Press. 2018

DATA SCIENCE

Born in late 90s, it was meant to **redefine statistics** by means of technology (and vice-versa).

After 2000, with the growth of **data on the web**, computer scientists and statisticians had to develop new methods to gather, scrape, merge, clean, store and query online data.

To date, the role of data scientists is commonly associated to the usage of **big data**.

DATA SCIENCE TASKS

CLUSTERING

Highlight groupings of relevant data, e.g. identify customers personas by a common behaviour.

OUTLIER DETECTION

Extract patterns that identify anomalies, e.g. online frauds. Different from clustering, it looks for **differences** rather than similarities.

ASSOCIATION-RULE MINING

Extract patterns that identify groups, e.g. products frequently bought together.

TYPES OF CLUSTERING ALGORITHMS

Non deterministic - every time you run it, it gives different results

Partitional clustering divides data objects into non overlapping groups. Requires the user to specify the number of clusters, indicated by the variable *k*.

Deterministic - every time you run it, it gives the same results

Hierarchical clustering determines cluster assignments by building a hierarchy.
Produce a tree-based hierarchy of points called a dendrogram. The number of clusters (*k*) is often predetermined by the user.

Deterministic - every time you run it, it gives the same results

Density-based clustering determines cluster assignments based on the density of data points in a region. It does not require to choose the number of clusters

CLUSTERING

There are plenty of clustering algorithms, with different strengths and weaknesses!

CLUSTERING

Highlight groupings of relevant data, e.g. identify customers personas by a common behaviour.

Instances of the dataset are sorted in a (pre-defined) number of **subgroups**

Subgroups are created by algorithms that matches **similarity between attribute values**

K-MEANS ALGORITHM

because of its versatility, clustering is often used for exploratory purposes

An unsupervised machine learning algorithm whose focus is on **k clusters** that a developer wants to extract (k is defined in a trial-and-error experiment, until data make sense, or by using common methods, e.g. elbow method) from **unlabelled data** which is data without defined categories or groups.

All data attributes must be **numeric** (nominal must be transformed in numeric). Each instance of the dataset is treated as a point in a **point cloud** (scatterplot) and the algorithm finds the **centers** (means) of every cloud.

K-MEANS ALGORITHM

The initial step is the **random selection** of a point as the center of the first point cloud. The subsequent k-1 centers are calculated on a probabilistic basis. Once defined the k centers, the algorithm **assigns** the other instances to the closest center and then **repositions** the centers to be in the middle of the point clouds.

Since the process is **non-deterministic** (every time it is run gives different results) it is run several times and the analyst compares results to see which ones are the most sensible. Once the clusters are found, these are **named** with a meaningful label representing its characteristics (attributes).

DATA SCIENCE TASKS

OUTLIER DETECTION

Extract patterns that identify anomalies, e.g. online frauds. Different from clustering, it looks for **differences** rather than similarities.

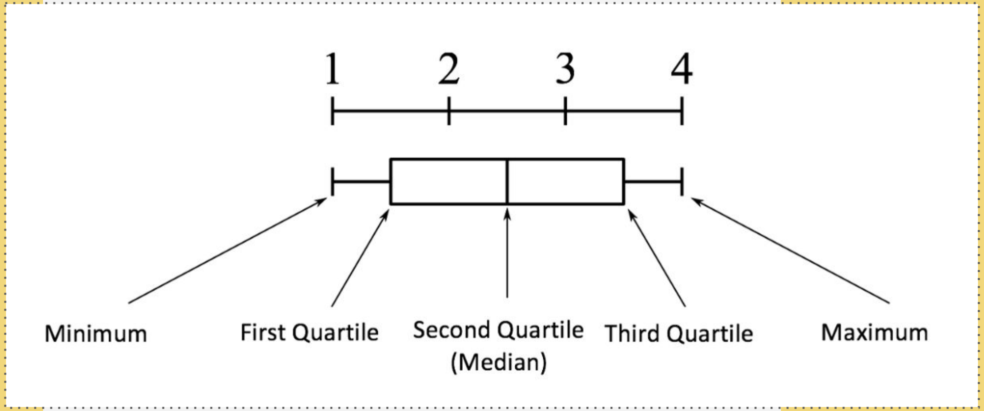
it requires a training dataset including both normal and anomalous records

It can be based on a prediction model, such as a **decision tree**, for classifying anomalies.

it requires the anomaly to be already discovered, it cannot be predicted

It can be based on the manual definition of **rules that characterise anomalies as patterns** (e.g. card transactions in weird places)

BOX PLOTS



Box plots are a graphical depiction of numerical data through their **quantiles**.

It is a very simple but effective way to visualize outliers. Think about the lower and upper whiskers as the boundaries of the data distribution. **Any data points that show above or below the whiskers, can be considered outliers or anomalous.**

DATA SCIENCE TASKS

ASSOCIATION-RULE MINING

Extract patterns that identify groups, e.g. products frequently bought together.

It looks for groups of entities that often **co-occur together**

It does not look for similarities (clusters) or differences (outliers) but for **correlations** between attributes.

APRIORI ALGORITHM

One of the main algorithms to produce association rules.
It first looks for **frequent itemsets**, i.e. all the combinations of items in the dataset that co-occur with a minimum predefined frequency.
Secondly, it generates a rule that represents the **probability** of the co-occurrence within frequent itemsets, based on the presence of other items.

APRIORI ALGORITHM

Rules are in the form: **IF {antecedent} THEN {consequent}**
Even a small dataset creates a huge number of association rules, which have to be **pruned**. Pruned rules include **trivial** rules and **inexplicable** rules.*

you must see this (when it's up again):
http://tylervigen.com/spurious-correlations

APRIORI ALGORITHM MEASURES

Given a rule "A -> C",

Rules with both high support and high confidence are usually interesting.

Measures how frequently **items in the dataset** occur together

SUPPORT

Measures the conditional **probability** that the consequent C will occur when there is the antecedent A.

CONFIDENCE

APRIORI ALGORITHM MEASURES

Measures **how much more often the antecedent and consequent of a rule A->C occur together** than we would expect if they were statistically independent.If A and C are independent, the Lift score is 1.

LIFT

Computes the **difference between the observed frequency** of A and C appearing together **and the frequency that would be expected if A and C were independent**. An leverage value of 0 indicates independence.

LEVERAGE

ELECTRONIC PUBLISHING
AND DIGITAL STORYTELLING
Lesson 6

HANDS-ON

INSTALL PYTHON LIBRARIES

Sklearn

```
pip install scikit-learn
```

Double-check if you have matplotlib

```
python -c "import matplotlib"
```

Otherwise

```
pip install matplotlib
```

mlxtend

```
pip install mlxtend
```

CREATE A NEW JUPYTER NOTEBOOK

Launch Jupyter from the shell

```
jupyter notebook
```

Open the browser and create a new notebook called **data_analysis_tutorial**

MOVE TO THE TUTORIAL

Open the course repository on the browser https://github.com/marilenadaquino/epds

Go to the folder tutorials/ and open in the browser the file called **data_analysis_tutorial.ipynb**

HOMEWORK

Fill in the questionnaire! https://forms.gle/FSndYWU1srx6HcGF7

THANKS

Does anyone have any questions?

marilena.daquino2@unibo.it
github