Review:

1. Machine Learning
2. Machine Learning sub-categories
3. Different data types and structures in Python

Statistics

Statistics:

We see the application of statistics everywhere around us.

There is no technology that did not use statistics in some forms.

The word come from the German one “statistik” or the Italian word “statista” or Latin “status” where all means ”political state”.

Statistics deals with methods for following different stages:

1. Data collection
2. Preparation, Organisation
3. Processing and analysis
4. Presentation and Draw conclusions

Data collection:

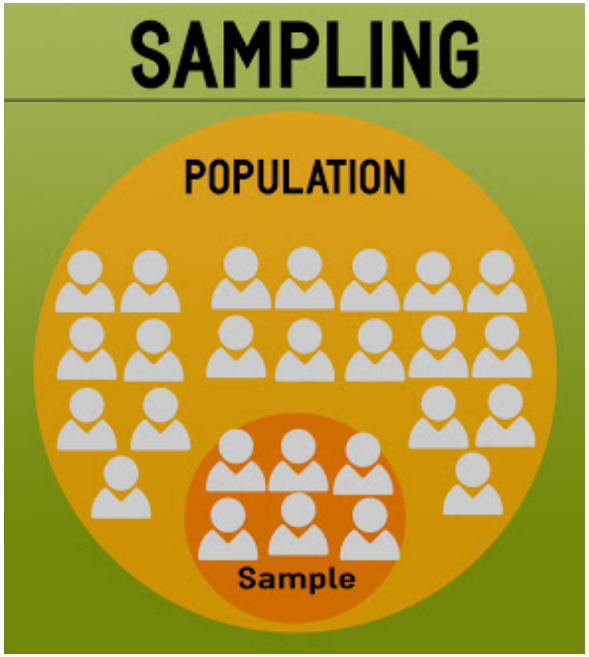
Population and samples:

- We are use to sampling in our day lives, we do it all the time,
- We usually test a sample of a greater collection,
- Population:
  A set of similar objects which falls within the the same category of interest, question or experiment.
- Some time it is hard to (or impossible) to analyse the whole population, then we use sampling.
  For examples:
  Some analysis for the whole population of the world
  Analysis of the account information off all customers of an international bank
  …

Data collection:

Samples:

- Samples is a portion of population,
- It is a finite subset of population,
- The number of unit is sample is sample size, for example what is the sample size in this picture?



Reason for sampling:

Sometimes analysing the whole population is not possible (all data not available or can not be collected, or the amount of data is too large or…)

The problem is time sensitive, results are needed fast,

Data collection:

What is population and sample size in this example:

What is the average age of the human population?
Discuss,

Data collection:

Sampling methods:

- Simple random sampling:
  - Samples are selected randomly so all have the same chance to be selected.
  - Maybe the best sampling method, sample would probably be a fair representative of the whole population.
  - Sometimes really hard to implement (sometime the whole population is not even known), e.g. there are viruses that are not discovered yet

- E.g. What is the average age of the human population?
  Pick random samples from the whole population

Data collection:

Sampling methods:

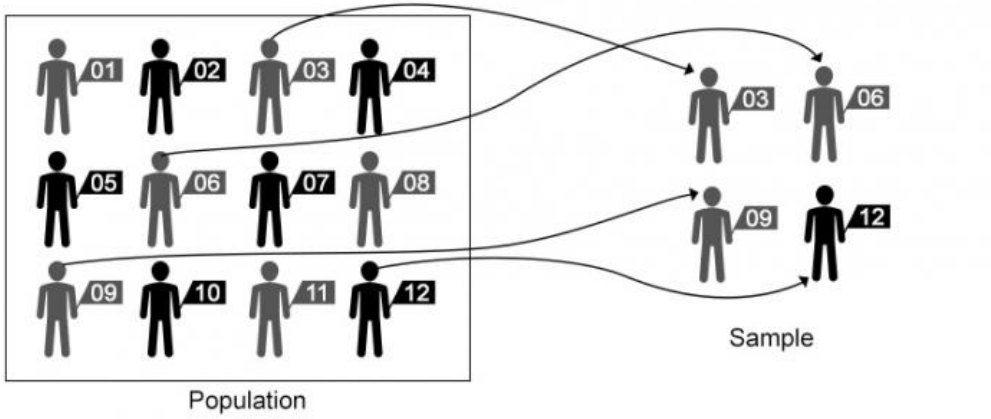
- Stratified Random Sampling:
  - In Stratified Random Sampling the whole population is divided into non-overlapping homogenous groups. Then random sampling is used to pick samples from each group.

- E.g. What is the average age of the human population?
  Divide the whole world into continents (or country) then pick random samples from each group.

Data collection:

Sampling methods:

- Systematic Random Sampling:
  - In Systematic Random Sampling every nth member of the population is picked.

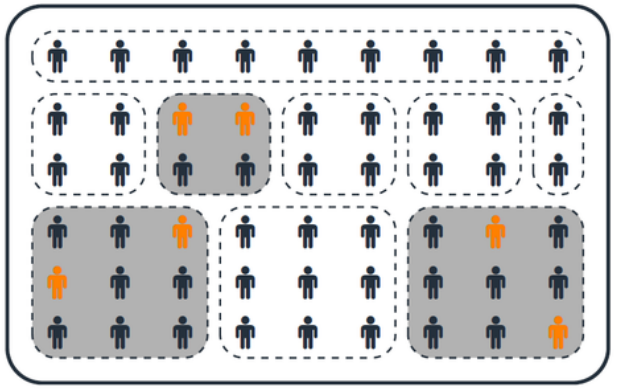


Data collection:

Sampling methods:

- Cluster Random Sampling:
  - In Cluster Random Sampling population is divided into subcategories, called clusters.
  - Then random sampling is used to select clusters,
  - All members of the cluster would be part of the sample,

- E.g. What is the average age of the human population?
  Divide the whole world into continents (or country) then pick random continents (or country).



Data collection:

Sampling methods, review:

- Simple random sampling: random
  - Researchers are not familiar with the population and possible categories
- Stratified Random Sampling: random from groups
  - Researchers are familiar with the population and possible categories and demographics
  - So they can divide the population into smaller groups
- Systematic Random Sampling:
  - Pick every nth sample
- Cluster Random Sampling: pick random clusters
  - When we don't have access to the whole population or it is hard/expensive to get it

- These are all Probability Sampling methods, where there are some randomness in the selection process.
- Non-Probability sampling, are those that not all units have the same chance of being in the sample (Not much randomness)

Some definitions:

Data collection:

We collect data everyday in our lives.

- Interviews
- Questionnaires and surveys
- Observations, sensors
- Documents and records
- Focus groups
- Historical data

Summary:

List
General purpose
Most widely used data structure
Grow and shrink size as needed
Sequence type
Sortable

Tuple
Immutable (can't add/change)
Useful for fixed data
Faster than Lists
Sequence type

Set
Store non-duplicate items
Very fast access vs Lists
Math Set ops (union, intersect)
Unordered

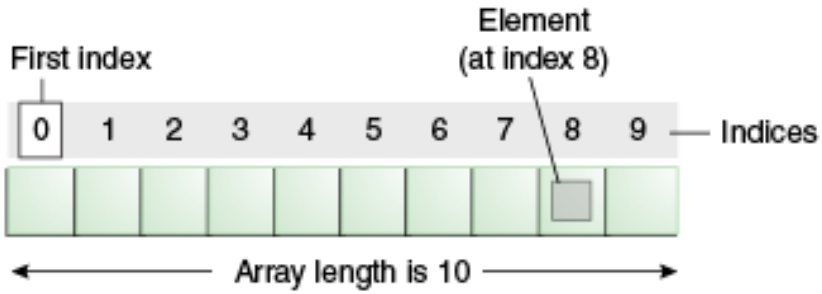
Dict
Key/Value pairs
Associative array, like Java HashMap
Unordered

Python:

Arrays:

Arrays, same as lists, are a collection of objects, however arrays are homogeneous (all elements have the same type).

Arrays are one of the most common data structures among all programing languages. Slicing and indexing are the same as lists.



Pp: Python practice

1. Arrays
   - Create an array
   - Indexing and slicing
   - Reassign values
   - Append,
   - Elementwise operations
   - “arrange” function, ”identity” function
   - Copy
   - …

More on arrays:

https://www.programiz.com/python-programming/array

https://docs.scipy.org/doc/numpy/reference/generated/numpy.array.html

Why "Learn"?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to "learn" to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

What We Talk About When We Talk About"Learning"

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:

  People who bought "Da Vinci Code" also bought "The Five People You Meet in Heaven" (www.amazon.com)
- Build a model that is *a good and useful approximation* to the data.

Data Mining/KDD

Definition := *"KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"* (Fayyad)

Applications:

- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Optimization, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Quality of service optimization
- Bioinformatics: Motifs, alignment
- Web mining: Search engines
- ...

What is Machine Learning?

- Machine Learning
  - Study of algorithms that
  - improve their performance
  - at some task
  - with experience
- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

Growth of Machine Learning

- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
- This trend is accelerating
  - Improved machine learning algorithms
  - Improved data capture, networking, faster computers
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for self-customization to user, environment
  - It turns out to be difficult to extract knowledge from human experts→*failure of expert systems in the 1980's.*

Applications

- Association Analysis
- Supervised Learning
  - Classification
  - Regression/Prediction
- Unsupervised Learning
- Reinforcement Learning

Learning Associations

- Basket analysis:

  $P(Y \mid X)$ probability that somebody who buys X also buys Y where X and Y are products/services.

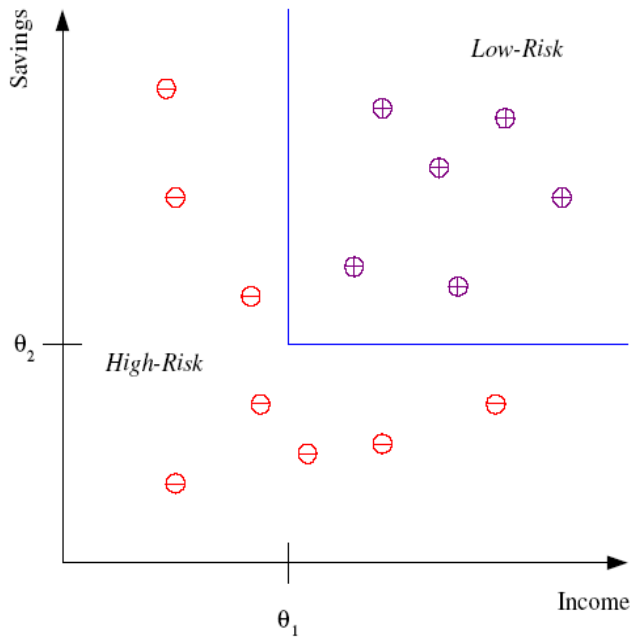
  Example: P (chips | beer) = 0.7

Market-Basket transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* $> \theta_1$ AND *savings* $> \theta_2$
THEN low-risk ELSE high-risk

Model

Classification: Applications

- Aka Pattern recognition
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
  - Use of a dictionary or the syntax of the language.
  - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- Medical diagnosis: From symptoms to illnesses
- Web Advertizing: Predict if a user clicks on an ad on the Internet.

Face image Recognition and creation