**DATA ANALYSIS. DATA MANAGEMENT**

**Data analysis bases. Methods of collection, classification and prediction. Decision trees. Processing of large volumes of data. Methods and stages of Data mining. Tasks Data mining. Visualization of data.**

1. Introduction to data
2. Data science process
3. Data collection

3.1 Data collection methods

1. Data processing
2. Data visualization

5.1 Python plotting libraries

1. Data analysis

6.1 Classification

6.2 Regression

6.3 Decision tree

1. Data management

**Introduction to data**

Data is a collection of raw facts and figures, that is need to be processed for information or for the collection of details. Data is used as “Input” for the computer system and data is a basic building for the information.

Information is a processed data and the information described as that form of data which is processed, organized and structured.



There are two types of data: quantitative and qualitative. Qualitative data is data that is non-measurable. It captures feelings, personal experience, attitudes and intensions. This data can be obtained from sources such as focus groups, interviews, narrative texts and reports. Quantitative data is numerical and measurable data, which allows for an objective assessment of a situation in order to compare one situation with another and to track conditions within the same situation over time.



Data type examples

|  |  |  |
| --- | --- | --- |
| Type | Definition | Example |
| Continuous | Data that can be measured | Age, height, weight of a personMoneyTime |
| Discrete | Data that relies on counts | Number of studentsNumber of animals in zoo |
| Nominal | Labeled and named data that can be divided into various group which do not overlap | Kazakh, russian, englishMen, womenWhite, black, red |
| Ordinal | Data involves some order stand in relation to each other in a ranked fashion | Very good, good, poorSatisfied, unsatisfiedHappy, unhappy, ok |
| Binary | Data that can take exactly two possible values | Yes, noSuccess, failureTrue, false |

**Data science process**

Data science process roughly consists from five activities: collection, processing, visualization, analysis and management. For working with data, firstly we need to collect them. It includes identification of and authenticated access to all related data. Then we need to explore data and process them. The first step in data preparation involves literally looking at the data to understand its nature, what it means, its quality and format. Processing activity includes cleaning data, sub-setting or filtering data, creating data, which programs can read and understand, such as modeling raw data into a more defined data model, or packaging it using a specific data format. The next step is visualization of getting results. The prepared data then would be passed onto the analysis step, which involves selection of analytical techniques to use, building a model of the data and analyzing results. After analysis you can manage data as you wish.



**Data collection**

Data collection is the systematic approach to gathering and measuring information from a variety of sources to get a complete and accurate picture of an area of interest. There are many methods of gathering information, and a wide variety of information sources. Listed below are some of the most common data collection methods:

1. *Survey, questionaires*.Survey is choosing the most appropriate information or evidence to answer the author’s questions.

2. *Interviews*. Interviewing is a great way to learn detailed information from a single individual or small number of individuals. It is useful when someone wants to gain expert opinions on the subject or talk to someone knowledgeable about a topic.

3. *Document review*. The author may collect some important data from various databases, meeting minutes, reports, attendance logs, financial records, newsletters and etc.

4. *Observation*. The author can make direct observations in offices, in various station were visited, it helps to understand how the systems behave in the real office environment.

5. *Focus group*.A facilitated group interview with individuals that have something in common.

6. *Case studies*.This method is focused on specific and interesting cases.

**Data processing**

Data processing occurs when data is collected and translated into usable information. It converts raw data into a readable format that can be interpreted, analyzed and used for a variety of purposes. It denotes the actual data manipulation techniques such as cleaning, classifying, sorting, calculating, summarizing, comparing, etc that is convert data into information.

Professional scientists usually spend a lot of time for data cleaning part. It includes:

1. Remove unwanted observations

2. Fix structural errors

3. Filter unwanted outliers

4. Handle missing data

**Data visualization**

The best data is the data that each can see and understand. Therefore all data scientists want to create and build the most comprehensive and understandable visualizations. It is not always simple: it is need to find the data, read it, clean it, filter it, and then use the right tool to visualize it.

Data visualization is the graphical representation of information. It can be done by using visual elements like charts, graphs and maps, which helps to provide an accessible way to see and understand trends, outliers and patterns in data.

Python is one of the most popular programming languages for data science and therefore enjoys a large number of useful libraries developed by its great developer and open-source community. Python offers multiple plotting libraries for data visualization that come packed with a lots of different features. It is matplotlib, seaborn, ggplot, plotly, pandas visualization.

**Data analysis**

Data analysis is a process of evaluating data using analytical and logical reasoning to examine each component of the data provided. More advanced types of data analysis includes data mining, which involves sorting through large data sets to identify trends, patterns and relationships.

Data mining is the analysis step that involves methods at the intersection of machine learning, statistics and database systems. It includes pre-processing tasks like data extraction, data cleaning, data fusion, data reduction and feature construction, as well as pre-processing steps like pattern and model interpretation, hypothesis confirmation and generation.

Machine learning is a method of data analysis that automates analytical model building. It provides systems the ability to automatically learn and improve from experinece without being explicity programmed. The five main steps that are involved in training a machine learning algorithms can be summarized as follows:

 1. Selecting features and collecting training samples

 2. Choosing a performance metric

 3. Chooshing a classifier and optimization algorithm

 4. Evaluating the performance of the model

 5. Tuning the algorithm

There are two types of machine learning: supervised and unsupervised learning.

The main goal in supervised learning is to learn a model from labeled training data that allows us to make predictions about unseen or future data.

Considering the example of email spam filtering, we train a model using supervised machine learning algorithm on a corpus of labeled emails, that correctly predict whether a new email belongs to inbox or to spam folder. A supervised learning task with discrete class labels, such as in the previous email spam filtering example, is also called a classification.



According to the picture email software uses text classifier to determine whether imcoming email is sent to the inbox or filtered to the spam folder. This example represents a typical binary classification task, where the machine learning algorithms learns a set of rules in order to distinguish between two possible classes: spam and not-spam emails.



Another category of supervised learning is regression, which tries to estimate or predict a target value for numerical variables. In regression, it is given a number of predictor (explanatory) variables and a continuous variable (target) and we need to find a relationship between those variables that allows us to predict an outcome.



One of the classification algorithms in machine learning is decision tree. Decision tree is the model as breaking down our data by making decision based on asking a series of questions.



Based on the features in our training set, the decision tree model learns a series of questions to infer the class labels of the samples. The preceding figure illustrates the banking loan concept of a decision tree based on categorical variables and real numbers.Using the decision algorithm, we start at the tree root and split the data on the feature, we can then repeat this splitting procedure at each child node until the leaves are pure. This means that the samples at each node all belong to the same class.

Clustering uses unsupervised learning to group data into distinct clusters or segments. In other words, clustering tries to find natural grouping in the data.

**Data management**

Data management is a process that includes all data processes to meet ongoing data lifecycle needs: data collection, data processing, data visualization, data analysis and etc.