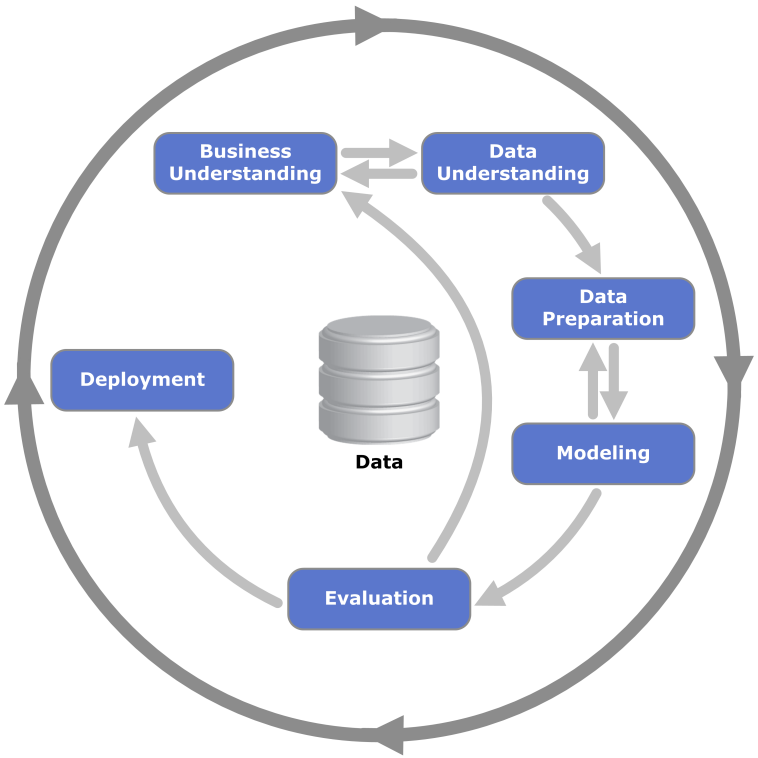
Machine learning

Managing Data



1. Problem Understanding lub Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Data cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

Missing Values

There are several reasons why data could be missing.

- changes in data collection methods,
- human error,
- combining various datasets,
- human bias,
- and others.

Missing Values

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	78,000	C	M	5000
1002	J2S7K7	F	–40,000	40	W	4000
1003	90210		10,000,000	45	S	7000
1004	6269	M	50,000	0	S	1000
1005	55101	F	99,999	30	D	3000

It is important to try to understand if there is a reason or pattern for the missing values.

For example, particular groups of people may not respond to certain questions in a survey.

Handling missing values

- **removal** - remove all instances with features that have a missing value.

- **this is a destructive approach and can result in the loss of valuable information and patterns that would have been useful in the machine learning process;**
- **this approach should be used only when the impact of removing the affected instances is relatively small or when all other approaches to dealing with missing data have been exhausted or are infeasible.**

Handling missing values

- **imputation** - is the use of a systematic approach to fill in missing data using the most probable substitute values.
  - **random imputation** - involves the use of a randomly selected observed value as the substitute for a missing value. Disadvantage with this approach is that it ignores useful information or patterns in the data when selecting substitute values.
  - **distribution-based imputation** approach - the substitute value for a missing feature value is chosen based on the probability distribution of the observed values for the feature. This approach is often used for categorical values, where the **mode** for the feature is used as a substitute for the missing value.
  - **mean or median imputation** - involves the use of the mean or median of the observed values as a substitute for the missing value.
  - **predictive imputation** is the use of a predictive model (regression or classification) to predict the missing value. With this approach, the feature with the missing value is considered the dependent variable (class or response), while the other features are considered the independent variables.

Handling outliers

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	78,000	C	M	5000
1002	J2S7K7	F	–40,000	40	W	4000
1003	90210		10,000,000	45	S	7000
1004	6269	M	50,000	0	S	1000
1005	55101	F	99,999	30	D	3000

A data point is an outlier if it is more than $1.5 \cdot \text{IQR}$ above the third quartile or below the first quartile. Said differently, low outliers are below $Q_1 - 1.5 \cdot \text{IQR}$ and high outliers are above $Q_3 + 1.5 \cdot \text{IQR}$.

Outliers are those data points that are distant from most of the similar points.

Outliers cause problems when it comes to building predictive models, such as long model training times, poor accuracy, an increase in error variance or a decrease in normality.

Handling outliers

- **Box Plot** - we can use a box plot to create a bunch of data points through quartiles. It groups the data points between the first and third quartile into a rectangular box. The box plot also displays the outliers as individual points using the interquartile range.
- **Scatter Plot**
- **Z-Score** the Z-score is a kind of parametric approach to detecting outliers. It assumes a normal distribution of the data. The outlier lies in the tail of the normal curve distribution and is far from the mean.

Transforming the Data

As part of the data preparation process, it is often necessary to modify or transform the structure or characteristics of the data

- to meet the requirements of a particular machine learning approach,
- to enhance our ability to understand the data,
- to improve the efficiency of the machine learning process.

Feature scaling brings all the features to the same level of magnitude.

Z-score standardization

- **z-score, or zero mean normalization** - the approach results in normalized values that have a mean of 0 and a standard deviation of 1.

$$x^* = \frac{x - \bar{x}}{\sigma},$$

where x^* new value, x value from dataset.

Min-Max Normalization

- With **min-max normalization**, we transform the original data to a interval $[0, 1]$.

$$x^* = \frac{x - \min(x_i)}{\max(x_i) - \min(x_i)}.$$

```
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}
```

Log Transformation

- The **logarithmic transformation** is applied to variables with distributions that are either not symmetric or have a wide range of values.

$$x^* = log(x_i).$$

Feature encoding techniques

Machine learning models are mathematical models that required numeric and integer values for computation. Such models can't work on categorical features. That's why we often need to convert categorical features into numerical ones. Machine learning model performance is affected by what encoding technique we use.

Feature encoding techniques

- **Label encoding**
- **One-hot encoding**

- **Label encoding** - is also known as integer encoding. **Integer encoding replaces categorical values with numeric values.** Here, the unique values in variables are replaced with a sequence of integer values. For example, let's say there are three categories: red, green, and blue. These three categories were encoded with integer values; that is, red is 0, green is 1, and blue is 2.

Drive	Code
Front-Wheel Drive	1
Rear-Wheel Drive	2
All-Wheel Drive	3

- **One-hot encoding** - One-hot encoding transforms the categorical column into labels and splits the column into multiple columns. The numbers are replaced by binary values such as 1s or 0s.

Drive	Front-Wheel Drive	Rear-Wheel Drive	All-Wheel Drive
Front-Wheel Drive	1	0	0
Rear-Wheel Drive	0	1	0
All-Wheel Drive	0	0	1

Drive	Front-Wheel Drive	Rear-Wheel Drive
Front-Wheel Drive	1	0
Rear-Wheel Drive	0	1
All-Wheel Drive	0	0

Types of machine learning algorithms

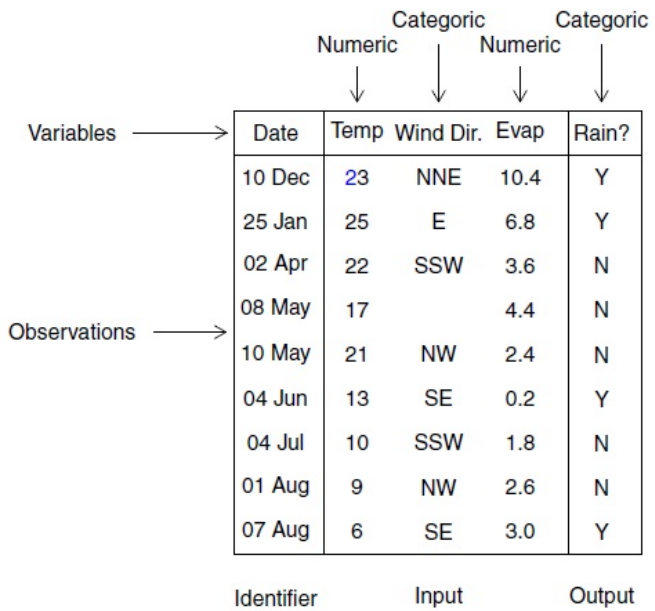
Machine learning algorithms are divided into two categories.

- **supervised learning**
- **unsupervised learning**

http://www.saedsayad.com/data_mining_map.htm

The **supervised learning** algorithm attempts to discover and model the relationship between the **target (output)** y feature (the feature being predicted) and the other features x.

$$x_n \rightarrow (\langle x_{n1}, \ldots, x_{nk} \rangle, y_n)$$



Date	Temp	Wind Dir.	Evap	Rain?
10 Dec	23	NNE	10.4	Y
25 Jan	25	E	6.8	Y
02 Apr	22	SSW	3.6	N
08 May	17		4.4	N
10 May	21	NW	2.4	N
04 Jun	13	SE	0.2	Y
04 Jul	10	SSW	1.8	N
01 Aug	9	NW	2.6	N
07 Aug	6	SE	3.0	Y

Supervised ML algorithms usually take a limited set of labeled data and build models that can make reasonable predictions for new data.

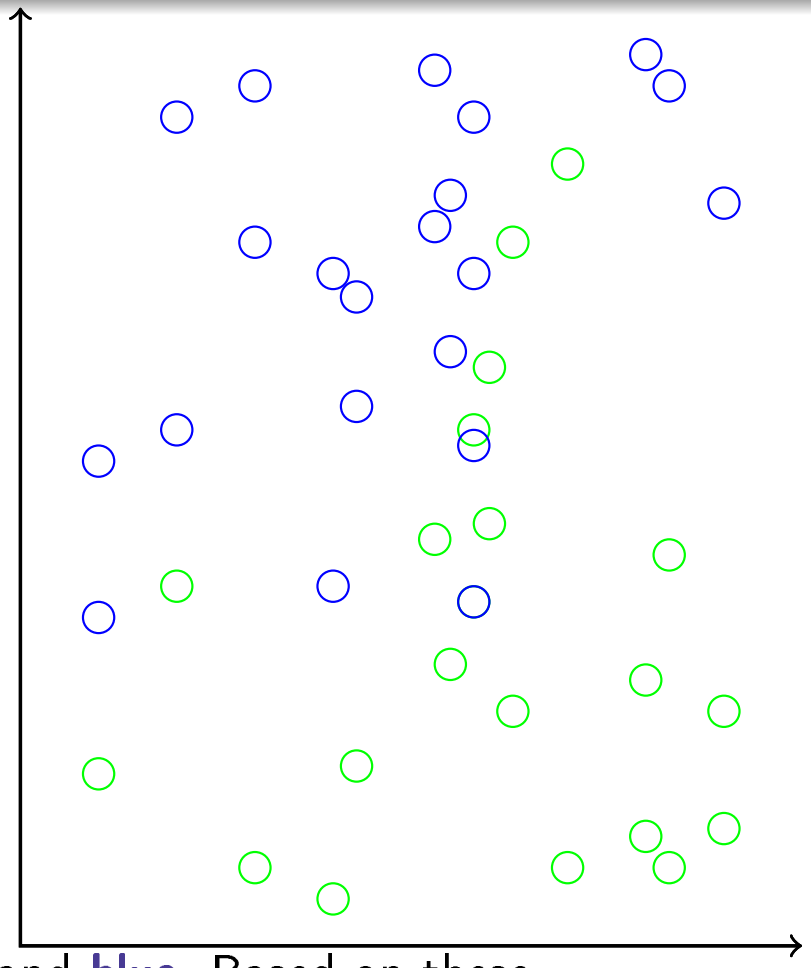
We can split supervised learning algorithms into two main parts:

- **classification techniques** - predict some finite and distinct types of categories. This could be a label that identifies if an email is spam or not, or whether an image contains a human face or not.
- **regression techniques** - predict continuous responses such as changes in temperature or values of currency exchange rates.

Supervised learning

classification - supervised machine learning task of predicting which category an example belongs to.

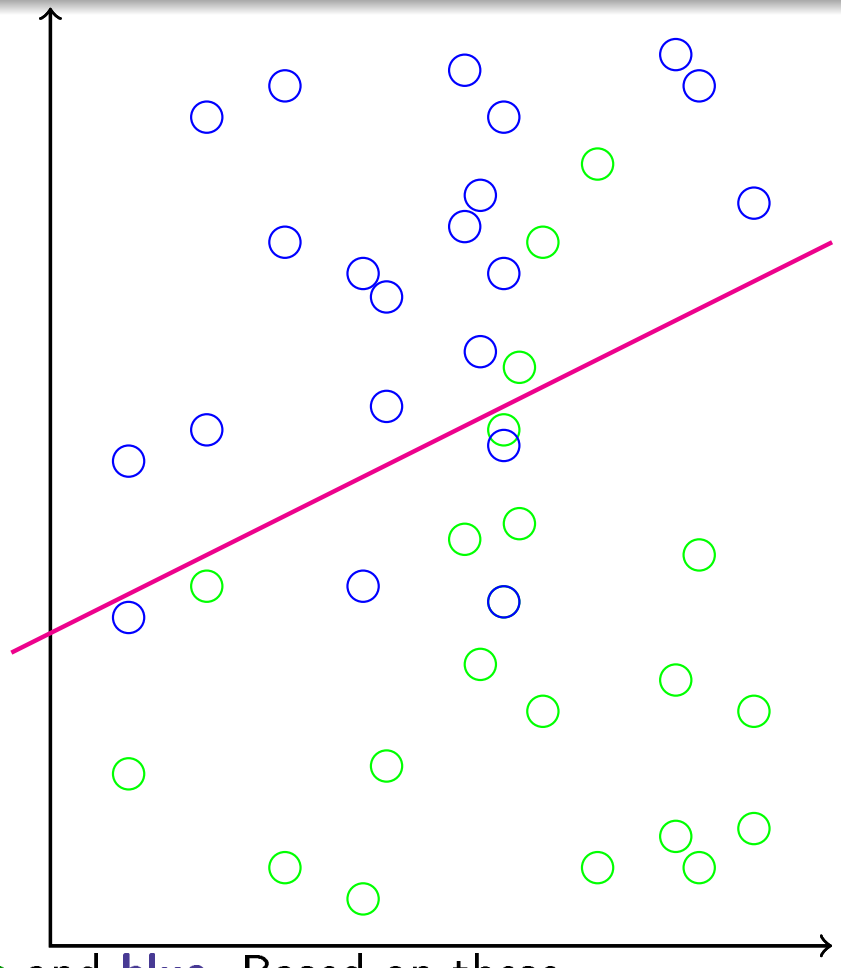
- the target feature to be predicted is a categorical feature known as the **class**, and is divided into categories called **levels**;
- a class can have two or more levels, and the levels may or may not be ordinal.



In the picture we have two classes, **green** and **blue**. Based on these observations, **a model - classifier (here the red line)** was built. We have a new (previously unknown) observation **(red)**. For attributes of red observation, the classifier determines whether it should be included in the **green** or **blue**.

Supervised learning

classification - supervised machine learning task of predicting which category an example belongs to.

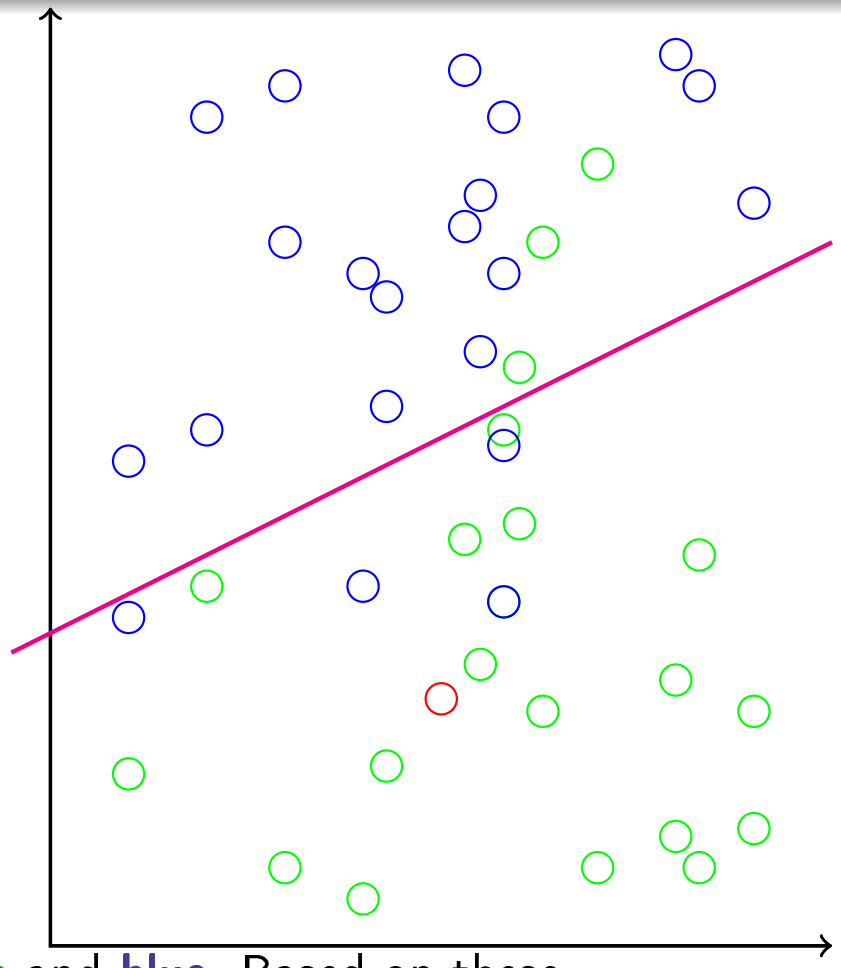


- the target feature to be predicted is a categorical feature known as the **class**, and is divided into categories called **levels**;
- a class can have two or more levels, and the levels may or may not be ordinal.

In the picture we have two classes, **green** and **blue**. Based on these observations, **a model - classifier (here the red line)** was built. We have a new (previously unknown) observation **(red)**. For attributes of red observation, the classifier determines whether it should be included in the **green** or **blue**.

Supervised learning

classification - supervised machine learning task of predicting which category an example belongs to.

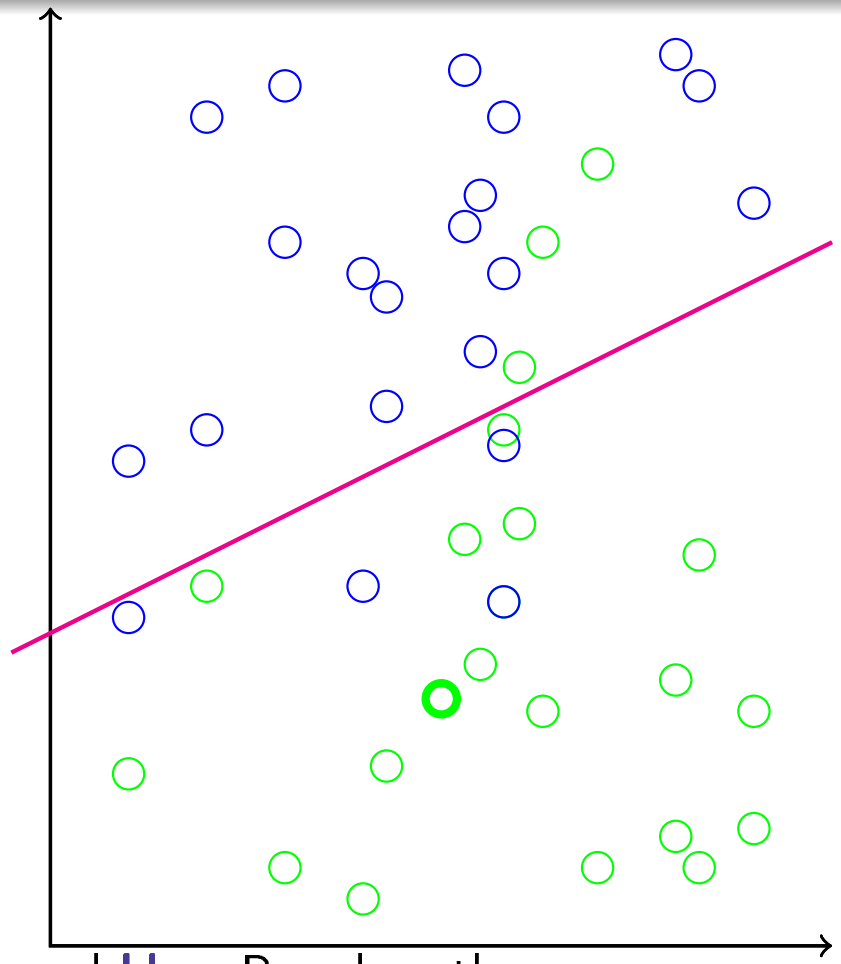


- the target feature to be predicted is a categorical feature known as the **class**, and is divided into categories called **levels**;
- a class can have two or more levels, and the levels may or may not be ordinal.

In the picture we have two classes, **green** and **blue**. Based on these observations, **a model - classifier (here the red line)** was built. We have a new (previously unknown) observation **(red)**. For attributes of red observation, the classifier determines whether it should be included in the **green** or **blue**.

classification - supervised machine learning task of predicting which category an example belongs to.

- the target feature to be predicted is a categorical feature known as the **class**, and is divided into categories called **levels**;
- a class can have two or more levels, and the levels may or may not be ordinal.



In the picture we have two classes, **green** and **blue**. Based on these observations, **a model - classifier (here the red line)** was built. We have a new (previously unknown) observation **(red)**. For attributes of red observation, the classifier determines whether it should be included in the **green** or **blue**.

Supervised learning - examples

Classification models are applied in

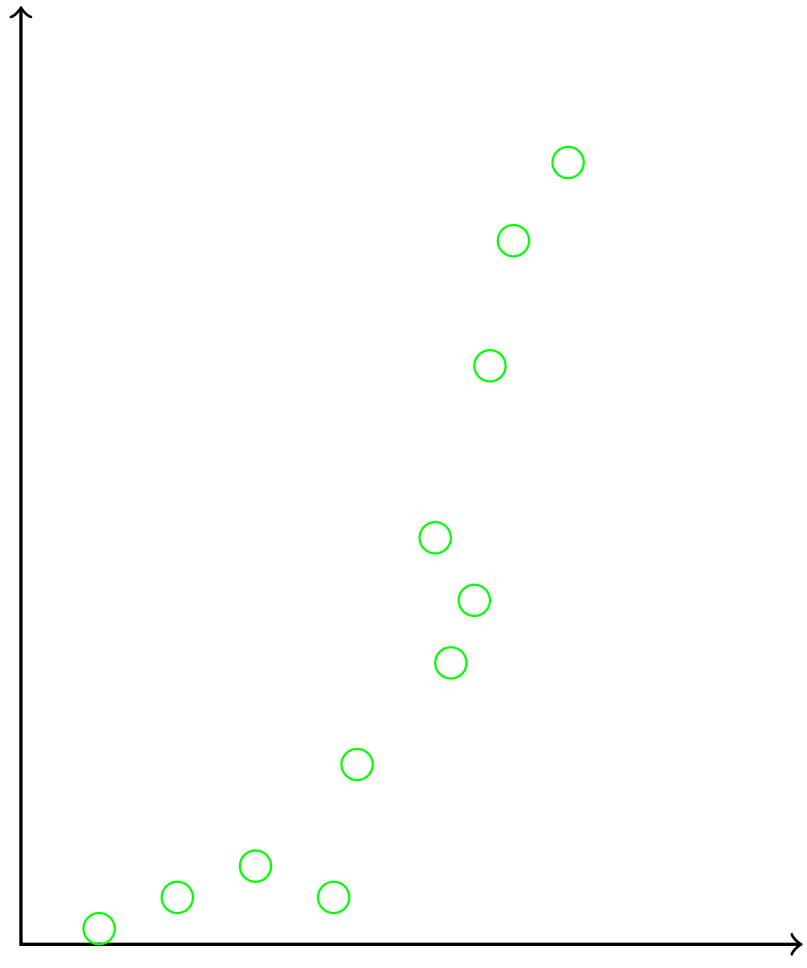
- speech and text recognition,
- object identification on images,
- credit scoring,
- detecting spam in emails, person has cancer

Typical algorithms for creating classification models are

- Support Vector Machine (SVM),
- decision tree approaches,
- k-nearest neighbors (KNN),
- logistic regression,
- Naive Bayes,
- neural networks;

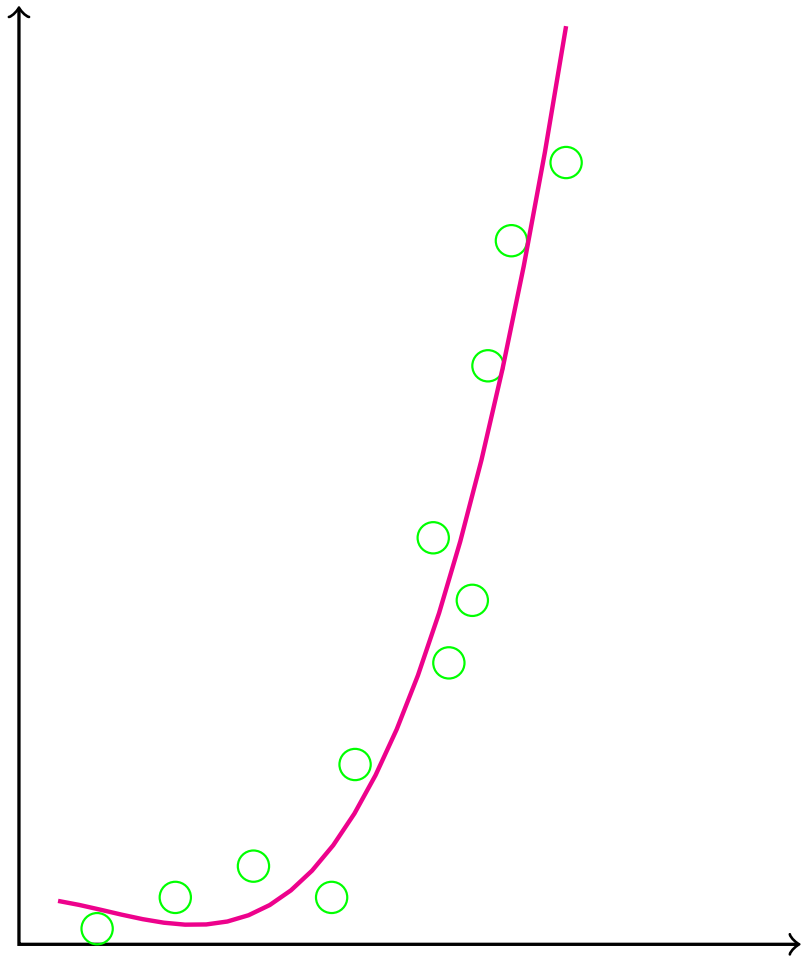
Supervised learning

- supervised learners can also be used to **predict numeric data** such as income, laboratory values, test scores, or counts of items - **regression**;
- we should find the function of variables called the **regression function the red curve**.



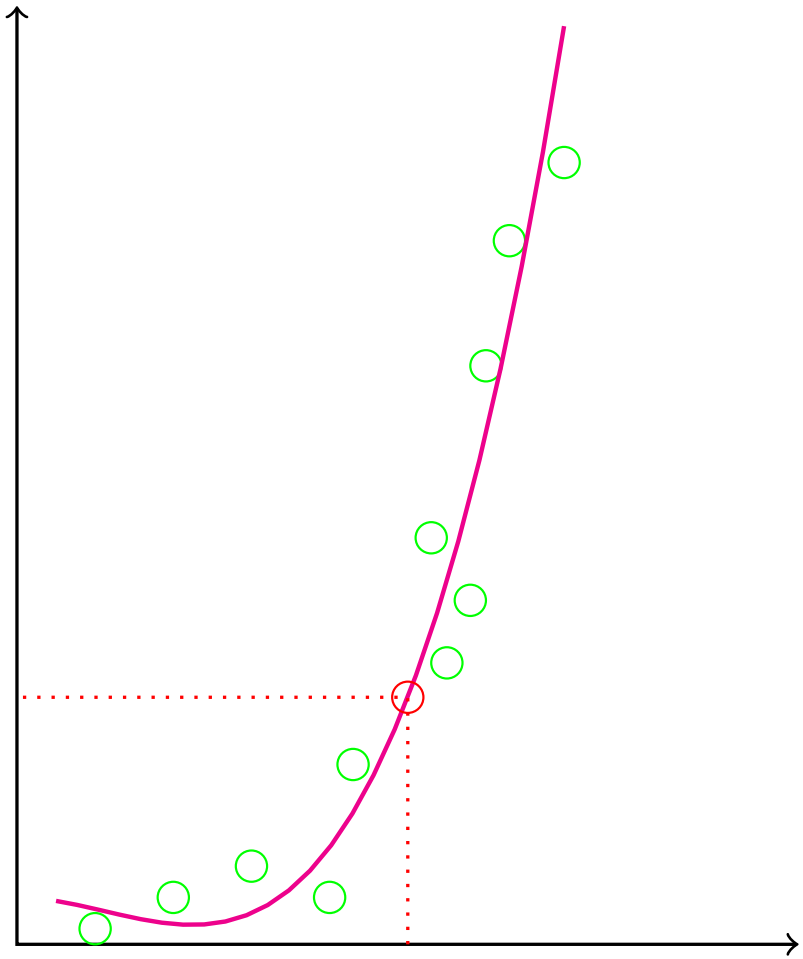
Supervised learning

- supervised learners can also be used to **predict numeric data** such as income, laboratory values, test scores, or counts of items - **regression**;
- we should find the function of variables called the **regression function the red curve**.



Supervised learning

- supervised learners can also be used to **predict numeric data** such as income, laboratory values, test scores, or counts of items - **regression**;
- we should find the function of variables called the **regression function the red curve**.



Supervised learning - examples

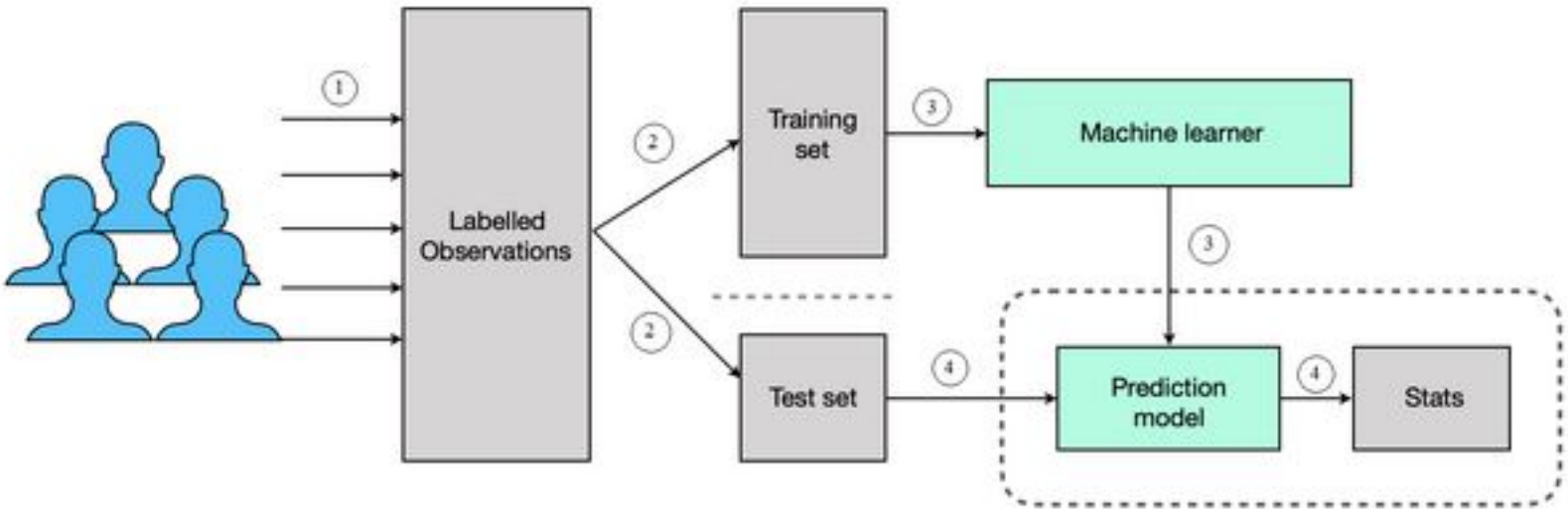
Regression models are applied in

- trading,
- forecasting of electricity load,
- revenue prediction,
- detecting spam in emails, person has cancer,
- the monthly rental of a house.

Creating a regression model usually makes sense if the output of the given labeled data is real numbers. Typical algorithms for creating regression models are

- linear regression,
- regression tree,
- neural networks;

Supervised learning



Unsupervised learning

- **A descriptive model** recognizes patterns or relationships in data.
- Because there is no target to learn, the process of training a descriptive model is called **unsupervised learning**.

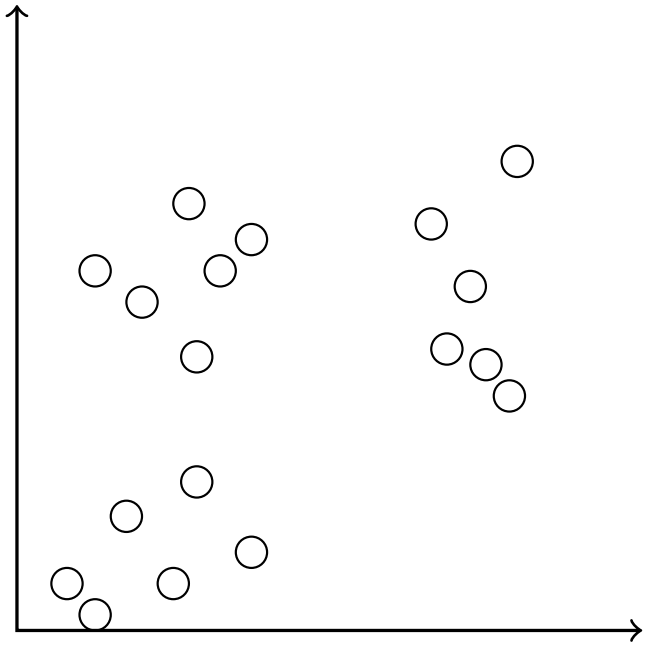
As opposed to predictive models that predict a target of interest, in a descriptive model, no single feature is more important than any other.

$$x_n \rightarrow (x_{n1}, \ldots, x_{nk})$$

Unsupervised learning algorithms do not use labeled datasets.

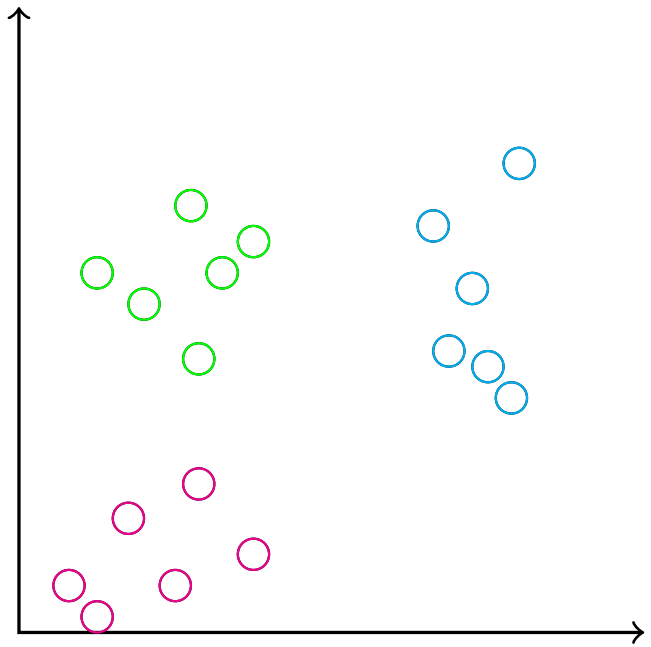
Unsupervised learning

- the descriptive modeling task called **pattern discovery** is used to identify useful associations within data. Pattern discovery is often used for **market basket analysis** on retailers transactional purchase data;
- the descriptive modeling task of dividing a dataset into homogeneous groups is called **clustering**;



Unsupervised learning

- the descriptive modeling task called **pattern discovery** is used to identify useful associations within data. Pattern discovery is often used for **market basket analysis** on retailers transactional purchase data;
- the descriptive modeling task of dividing a dataset into homogeneous groups is called **clustering**;



Unsupervised learning

Unsupervised learning technique is

- clustering,
- PCA (Principal component analysis),
- association rules.

Clustering is applied in

- market researches,
- different types of exploratory analysis,
- image segmentation,
- social networks.

- k-means,
- k-medoids,
- apriori,
- Kohonen networks.

Association Analysis

If you visit any retail website, you'll be exposed to terms such as

- related products,
- because you watched...,
- customers who bought x also bought y,
- recommended for you.

We can establish two techniques

- **association rules,**
- **recommendation engines.**

Association rule analysis is commonly referred to as market basket analysis, as it is concerned with understanding what items are purchased together.

With recommendation engines, the goal is to provide a customer with other items that they will enjoy based on how they have rated items they have viewed or purchased previously.

Market basket analysis

Association rules

Association rules allow us to explore the relationship between items and sets of items.

1. Cherry coke, chips, lemon
2. Cherry coke, chicken wings, lemon
3. Cherry coke, chips, chicken wings, lemon
4. Chips, chicken wings, lemon
5. Cherry coke, lemon, chips, chocolate cake
6. lemon

Association rules

1. Cherry coke, chips, lemon
2. Cherry coke, chicken wings, lemon
3. Cherry coke, chips, chicken wings, lemon
4. Chips, chicken wings, lemon
5. Cherry coke, lemon, chips, chocolate cake

Association analysis is a data mining technique that has the purpose of finding the optimal combination of **products** or services and allows marketers to exploit this knowledge to provide recommendations, optimize product placement, or develop marketing programs that take advantage of cross-selling.

If a customer buys chips, **then** he'll buy a lemon.

If a customer buys chicken wings, **then** he'll buy a chocolate cake.

Association rules

Association analysis is a data mining technique that has the purpose of finding the optimal combination of products or **services** and allows marketers to exploit this knowledge to provide recommendations, optimize product placement, or develop marketing programs that take advantage of cross-selling.

If a patient undergoes treatment A, **then** he'll exhibit symptom X.

If a customer buys a hotel room, **then** he'll rent a car.

If the temperature is high and the wind direction is south, **then** it will not rain.

- $\mathcal{I} = \{i_1, i_2, ..., i_n\}$ - the set of items;
- the set of transactions $T = \{T_1, T_2, ..., T_N\}$, where $T_i \subset \mathcal{I}$, $i = 1, ..., N$;
- X is the subset of the set of items $X \subset \mathcal{I}$;

Subset X can be an empty set.

Frequency of the set X

$$\sigma(X) = |\{T_i : X \subset T_i, T_i \in T\}|$$

The support for an itemset is the proportion among all cases where the itemset of interest is present.

$$support(X) = \frac{\sigma(X)}{N}$$

Association rules

1. Cherry coke, chips, lemon
2. Cherry coke, chicken wings, lemon
3. Cherry coke, chips, chicken wings, lemon
4. Chips, chicken wings, lemon
5. Cherry coke, lemon, chips, chocolate cake

$$\sigma(\{chips\}) = 4$$

$$\sigma(\{lemon, chocolatecake\}) = 1$$

$$support(\{chips\}) = \frac{4}{5}$$

An **association rule** is a relationship in the data, of the form $\mathcal{X} \Rightarrow \mathcal{Y}$, where $X, Y \subset \mathcal{I}$, $X \cap Y = \emptyset$.

example:

If a customer buys chips, then he'll buy a lemon.

$$\{chips\} \Rightarrow \{lemon\}$$

- the result of a market basket analysis is a collection of association rules that specify patterns found in the relationships among items in the itemsets.
- **association rules** are always composed from subsets of itemsets and are denoted by relating one itemset on the **left-hand side (LHS)** of the rule to another itemset on the **right-hand side (RHS)** of the rule.

An **association rule** is a relationship in the data, of the form $\mathcal{X} \Rightarrow \mathcal{Y}$, where $X, Y \subset \mathcal{I}$, $X \cap Y = \emptyset$.

Support - a rule's support is a measurement of how frequently the itemset appears in the dataset;

$$support\,(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{N},$$

The rule r: **if a customer buys a smartphone, then he buys a case**.

Let us assume:

$$support(r) = 2\%$$

confidence - a rule's confidence is a measurement of how often the rule has been found to be true;

$$confidence\,(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

The rule r: **if a customer buys a smartphone, then he buys a case**.

Let us assume:

$$confidence(r) = 100\%$$

lift - is a measure of the improvement of the rule support over what can be expected by chance (measures how much more likely one item or itemset is to be purchased relative to its typical rate of purchase, given that you know another item or itemset has been purchased);

$$lift\,(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)\sigma(Y)}.$$

- **lift=1** $\mathcal{A}$ i $\mathcal{B}$ do not impact on each other;
- **lift<1** $\mathcal{A}$ i $\mathcal{B}$ are negatively related;
- **lift>1** $\mathcal{A}$ i $\mathcal{B}$ are positively correlated (two items are found together more often than expected by chance alone);

The rule r: **if a customer buys a smartphone, then he buys a case**.

Let us assume:

$$lift(r) = 3.7$$

A large lift value is a strong indicator that a rule is important and reflects a true connection between the items.

Association rules

The frequent itemset is an itemset whose support value is greater than a threshold value(support).

Apriori property (heuristics)

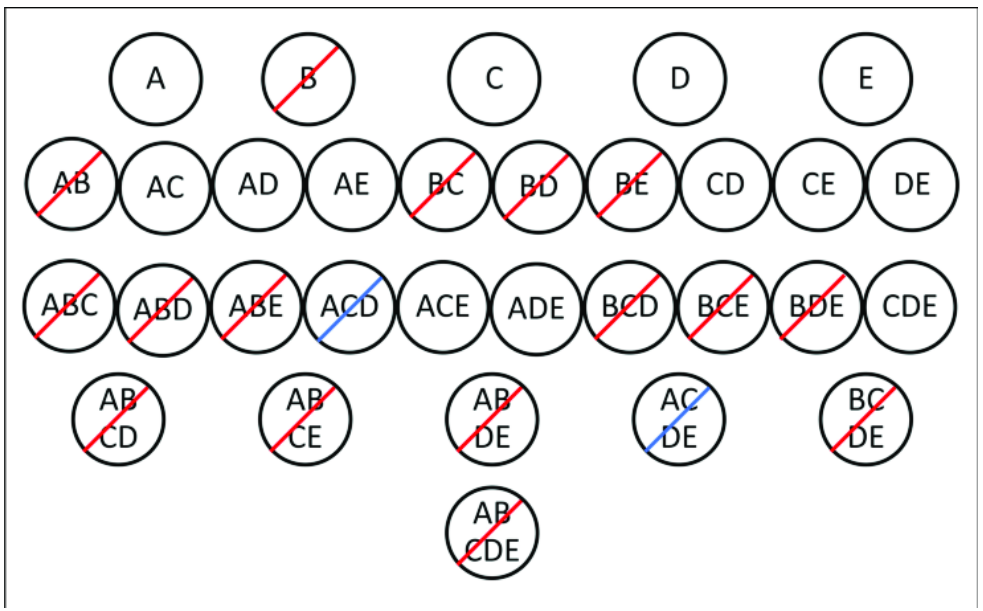
All subsets of a frequent itemset must also be frequent.

- if $\{A, B\}$ is frequent, then $\{A\}$ and $\{B\}$ must both be frequent.
- if we know that $\{A\}$ does not meet a desired support threshold, there is no reason to consider $\{A, B\}$ or any itemset containing $\{A\}$, it cannot possibly be frequent.

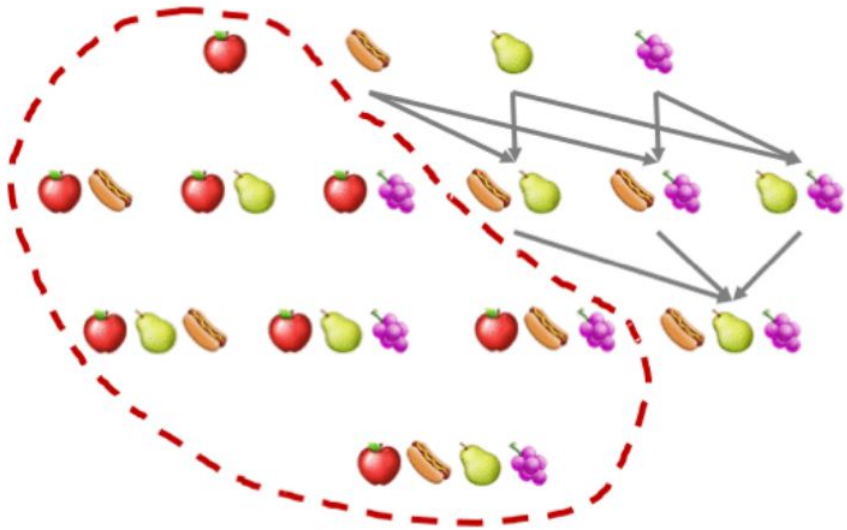
The frequent itemset is an itemset whose support value is greater than a threshold value(support).

Apriori property (heuristics)

All subsets of a frequent itemset must also be frequent.

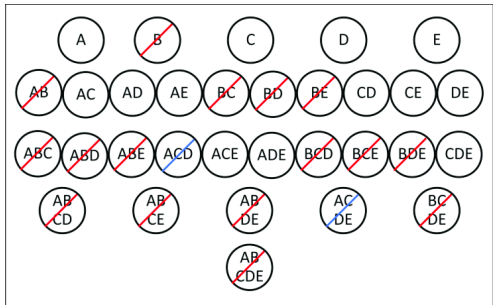


The Apriori algorithm for association rule learning



- transactional datasets can be large in both the number of transactions as well as the number of items or features that are recorded.
- the problem is that the number of potential itemsets grows exponentially with the number of features.
- let us assume, that given k items that can appear or not appear in a set, there are 2^k possible itemsets that could be potential rules.

The Apriori algorithm for association rule learning



The Apriori algorithm:

1. identifying all the itemsets that meet a **minimum support threshold**.
2. creating rules from these itemsets using those meeting a **minimum confidence threshold**.

- the first phase occurs in multiple iterations.
- each successive iteration involves evaluating the support of a set of increasingly large itemsets. For instance, iteration one involves evaluating the set of 1-item itemsets (1-itemsets), iteration two evaluates the 2-itemsets, and so on.
- the result of each iteration i is a set of all the i-itemsets that meet the minimum support threshold.

Thank you for your attention!!!