Data Mining: Data

Lecture Notes for Chapter 2

Introduction to Data Mining , 2nd Edition

by

Tan, Steinbach, Kumar

Outline

- Attributes and Objects
- Types of Data
- *Data Exploration*
- Data Quality
- Data Preprocessing

ATTRIBUTES AND OBJECTS

What is Data?

- Collection of ***data objects*** and their ***attributes***
- An ***attribute*** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or **feature**
- A collection of attributes describe an ***object***
  - Object is also known as record, point, case, sample, entity, or **instance**

Attributes

Objects

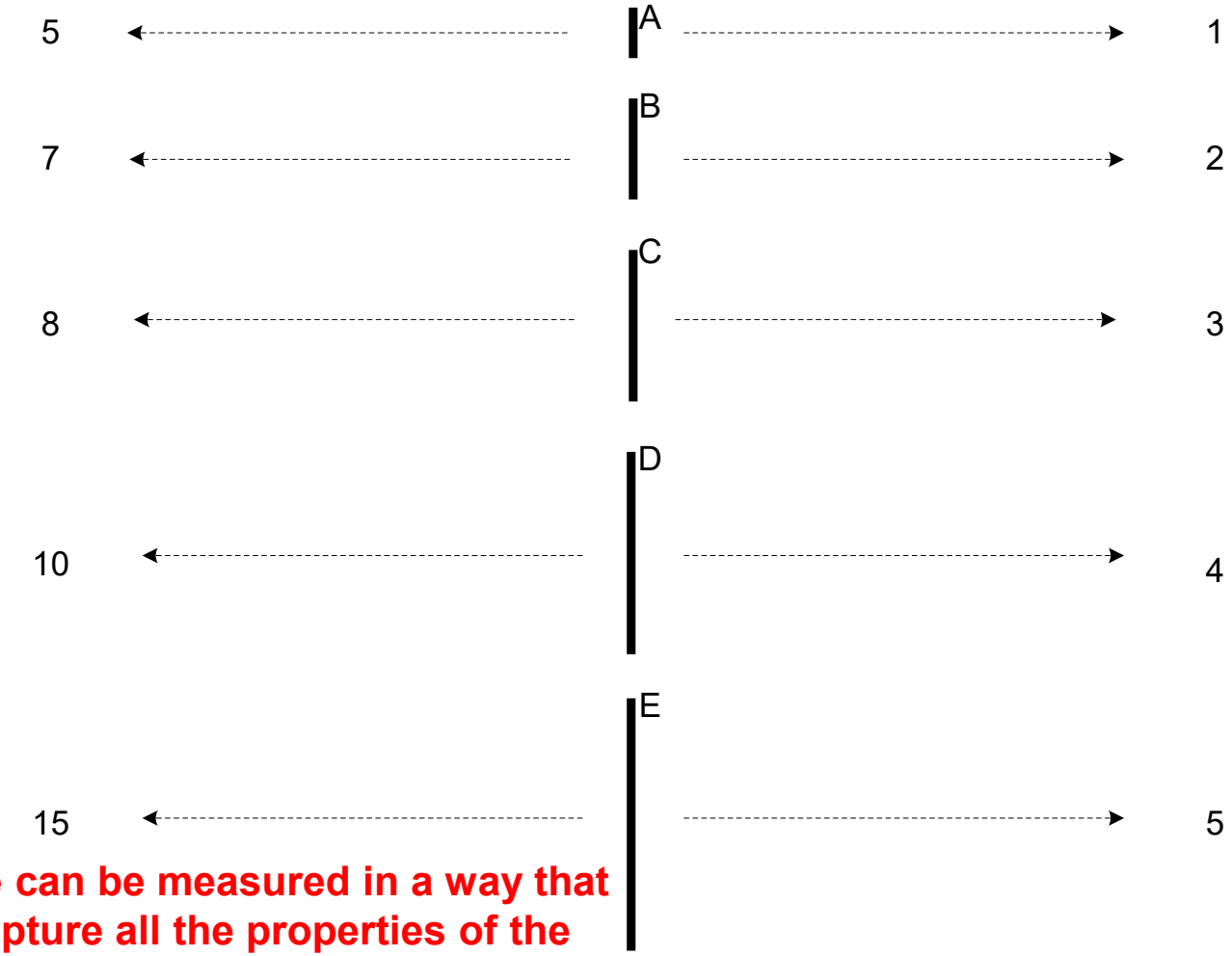
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- ***Attribute values*** are numbers or symbols assigned to an attribute for a particular object
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values – atentar para as propriedades do mesmo
    - Example: height can be measured in feet or meters (ver próximo slide)
  - Different attributes can be mapped to the same set of values – semântica é diferente
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different (não faz sentido calcular a media dos IDs (=,!=), mas das idades sim)

Measurement of Length

- The way you measure an attribute may not match the attributes properties.



This scale preserves only the ordering property of length.

This scale preserves the ordering and additvity properties of length.

An attribute can be measured in a way that does not capture all the properties of the attribute

Types of Attributes

The type of an attribute should tell us what properties of the attribute are reflected in the values used to measure it

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval (no true zero (não representa a ausência de valor))
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit (0° não significa ausência de calor) (10C=50F and 20C=68F).
  - Ratio (true zero exists (representa a ausência de valor))
    - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race), weight (peso=0 = sem peso = ausência de uma determinada quantidade)

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Differences are meaningful : + -
  - Ratios are meaningful * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & meaningful differences
  - Ratio attribute: all 4 properties/operations

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Differences are meaningful : + -
  - Ratios are meaningful * /

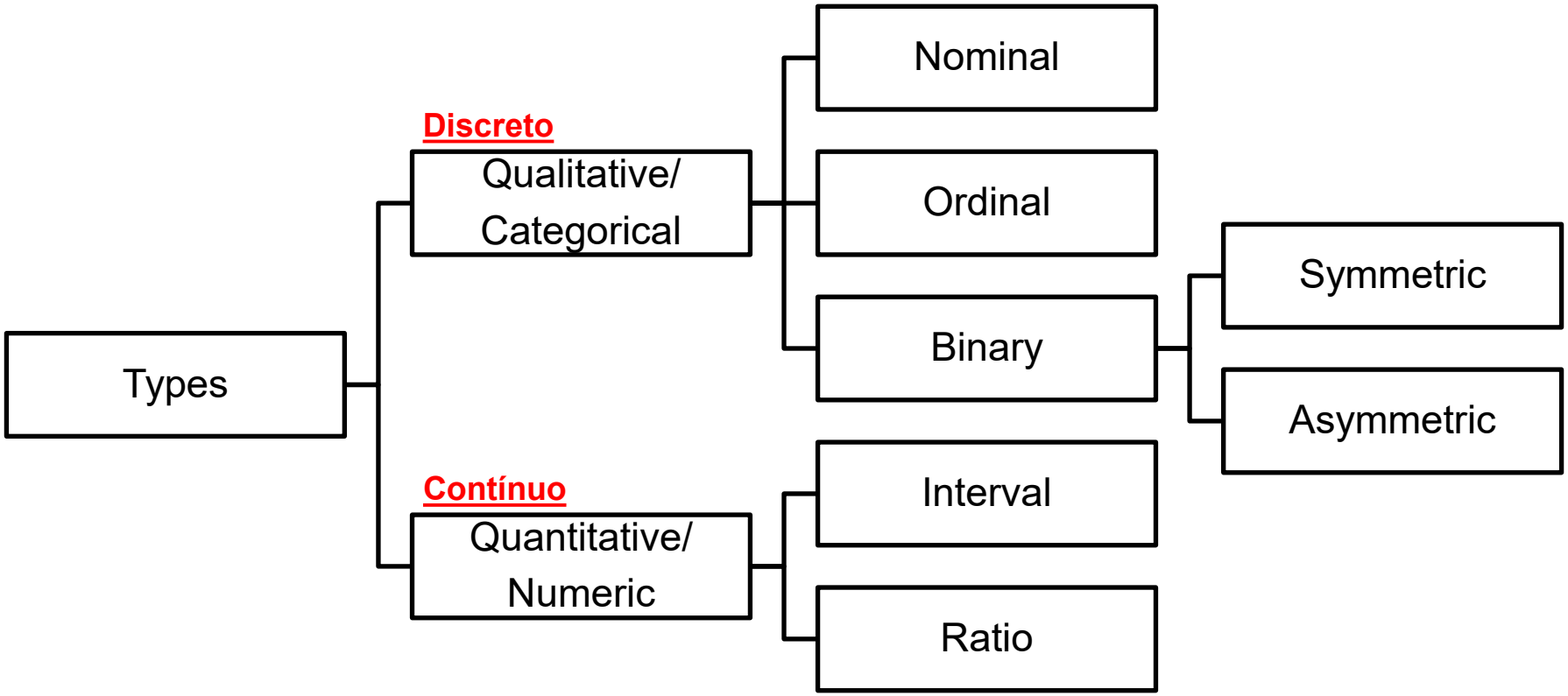
- *Knowing the measurement scale for your variables can help prevent mistakes like taking the average of a group of zip (postal) codes, or taking the ratio of two pH values.*

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

Binary Attributes

- Nominal subset – presents only two values/states (yes/no, true/false, etc.)
  - Symmetric: Both values are equally important (Gender)
  - Asymmetric: Both values are not equally important:
    - Words present in documents
    - Items present in customer transactions
      - If we met a friend in the grocery store would we ever say the following?
        "I see our purchases are very similar since we didn't buy most of the same things."
      - Association analysis uses asymmetric attributes

Summary


Types
Discreto
Qualitative/
Categorical
Nominal
Ordinal
Binary
Symmetric
Asymmetric
Contínuo
Quantitative/
Numeric
Interval
Ratio

TYPES OF DATA

Types of data sets

Han, Kamber, and Pei (2011)

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data
  - Video data

***Foram organizados em grupos. Não cobre todas as possibilidades. Outros agrupamentos (organizações) são possíveis.**

Types of data sets

- Record
  - **Relational records**
  - Data matrix, e.g., numerical matrix
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia
  - Spatial data: maps
  - Image data
  - Video data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Types of data sets

- Record
  - Relational records
  - **Data matrix, e.g., numerical matrix**
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia
  - Spatial data: maps
  - Image data
  - Video data

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

Types of data sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix
  - Document data: text documents: term-frequency vector
  - **Transaction data**
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data

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

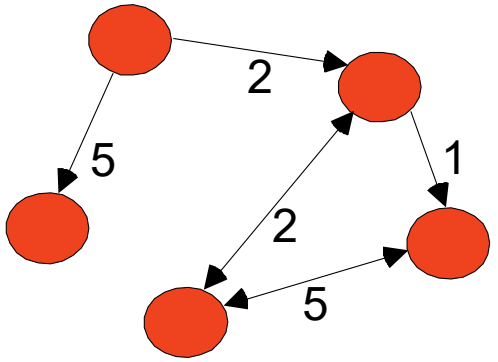
Types of data sets

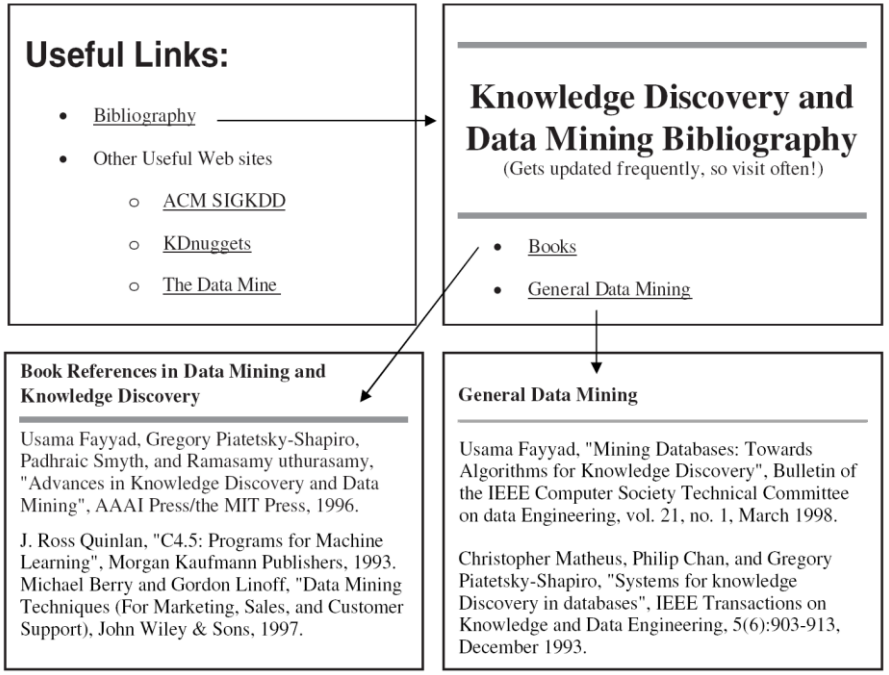
- Record
  - Relational records
  - Data matrix, e.g., numerical matrix
  - **Document data: text documents: term-frequency vector**
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Types of data sets

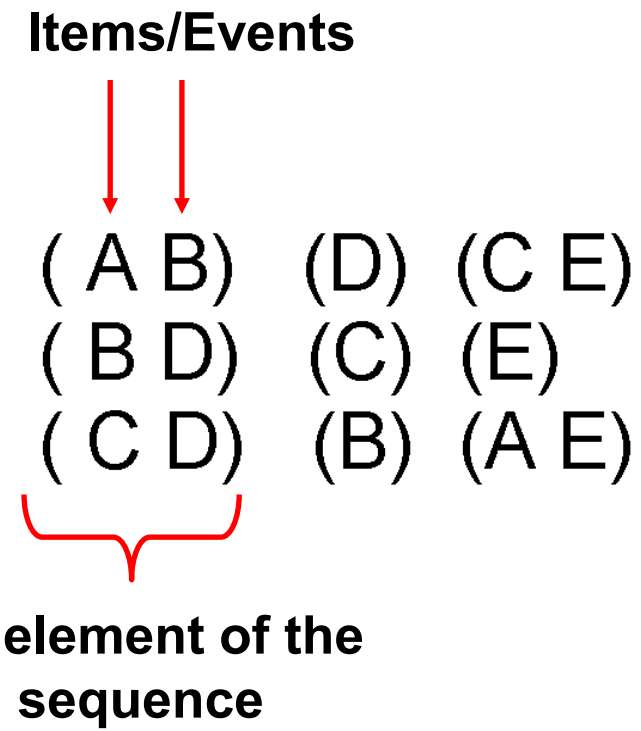
- Record
  - Relational records
  - Data matrix, e.g., numerical matrix
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - **World Wide Web**
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data
  - Video data





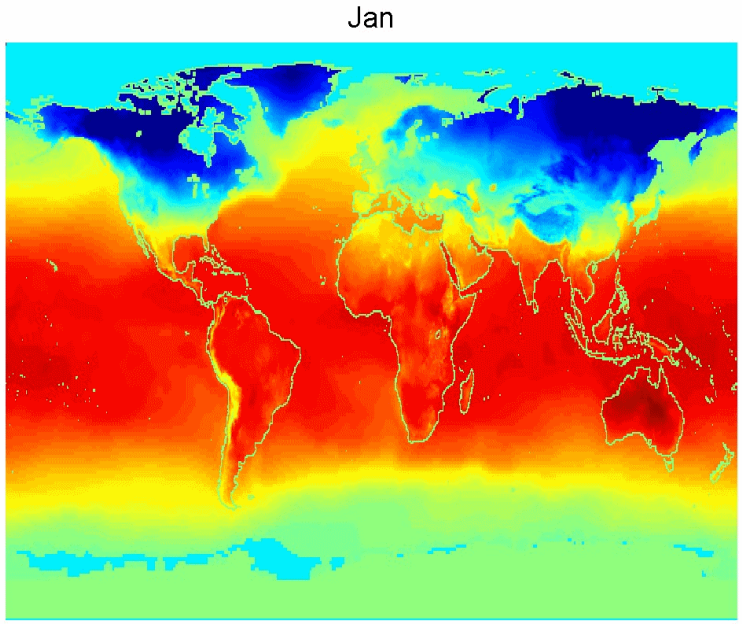
Types of data sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - **Sequential Data: transaction sequences**
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data
  - Video data



Types of data sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - **Temporal data: time-series**
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - **Spatial data: maps**
  - Image data
  - Video data



Average Monthly Temperature of land and ocean [Spatio-Temporal]

DATA EXPLORATION

What is data exploration?

A preliminary exploration of the data to better understand its characteristics.

- Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools
- In our discussion of data exploration, we focus on
  - Summary statistics
  - Visualization

Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Summary Statistics

- Summary statistics are numbers that summarize properties of the data
  - Summarized properties include frequency, location and spread
    - Examples: location - mean
      spread - standard deviation
  - Most summary statistics can be calculated in a single pass through the data

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.
- The mode of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

Percentiles (Percentis)

- For continuous data, the notion of a percentile is more useful.

Given an ordinal or continuous attribute *x* and a number *p* between 0 and 100, the p^{th} percentile x_p is a value of x such that *p*% of the observed values of x are less than x_p.

- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less or equal than $x_{50\%}$.

Percentiles

Table 3.2. Percentiles for sepal length, sepal width, petal length, and petal width. (All values are in centimeters.)

Percentile	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	2.0	1.0	0.1
10	4.8	2.5	1.4	0.2
20	5.0	2.7	1.5	0.2
30	5.2	2.8	1.7	0.4
40	5.6	3.0	3.9	1.2
50	5.8	3.0	4.4	1.3
60	6.1	3.1	4.6	1.5
70	6.3	3.2	5.0	1.8
80	6.6	3.4	5.4	1.9
90	6.9	3.6	5.8	2.2
100	7.9	4.4	6.9	2.5

https://medium.com/@juliodelimas/percentis-e-sua-import%C3%A2ncia-nos-testes-de-performance-ea83e3bba462

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean is also commonly used.

https://pt.wikipedia.org/wiki/M%C3%A9dia_truncada

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Spread: Range and Variance

- Range is the difference between the max and min
  - Amplitude

https://brasilescola.uol.com.br/matematica/medidas-dispersao-amplitude-desvio.htm

$$\text{range}(x) = \max(x) - \min(x)$$

https://brasilescola.uol.com.br/matematica/medidas-dispersao-variancia-desvio-padrao.htm

- The variance or standard deviation is the most common measure of the spread of a set of points.

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

Quanto menor é a variância, mais próximos os valores estão da média; mas quanto maior ela é, mais os valores estão distantes da média.

O desvio padrão é capaz de identificar o "erro" em um conjunto de dados, caso quiséssemos substituir um dos valores coletados pela média aritmética.

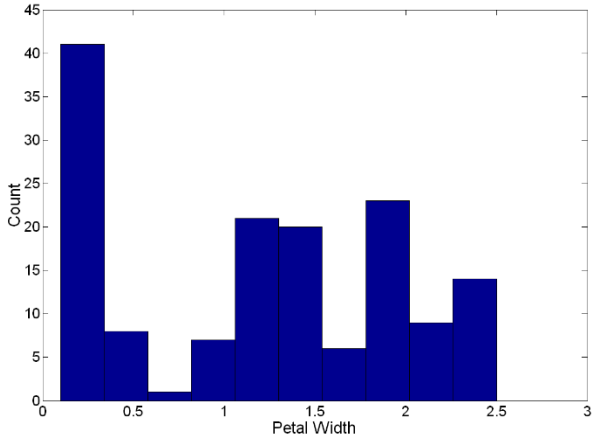
Visualization

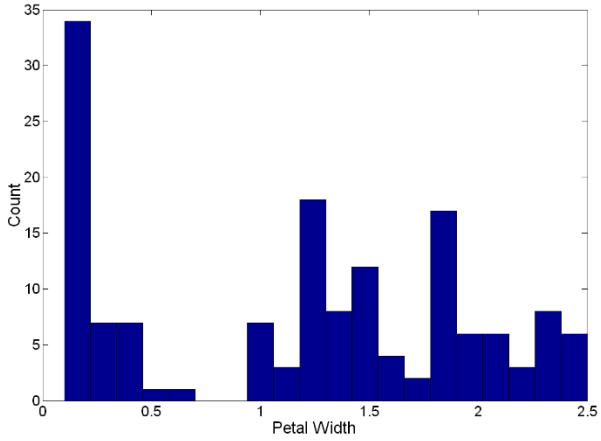
Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Technique: Histograms

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)

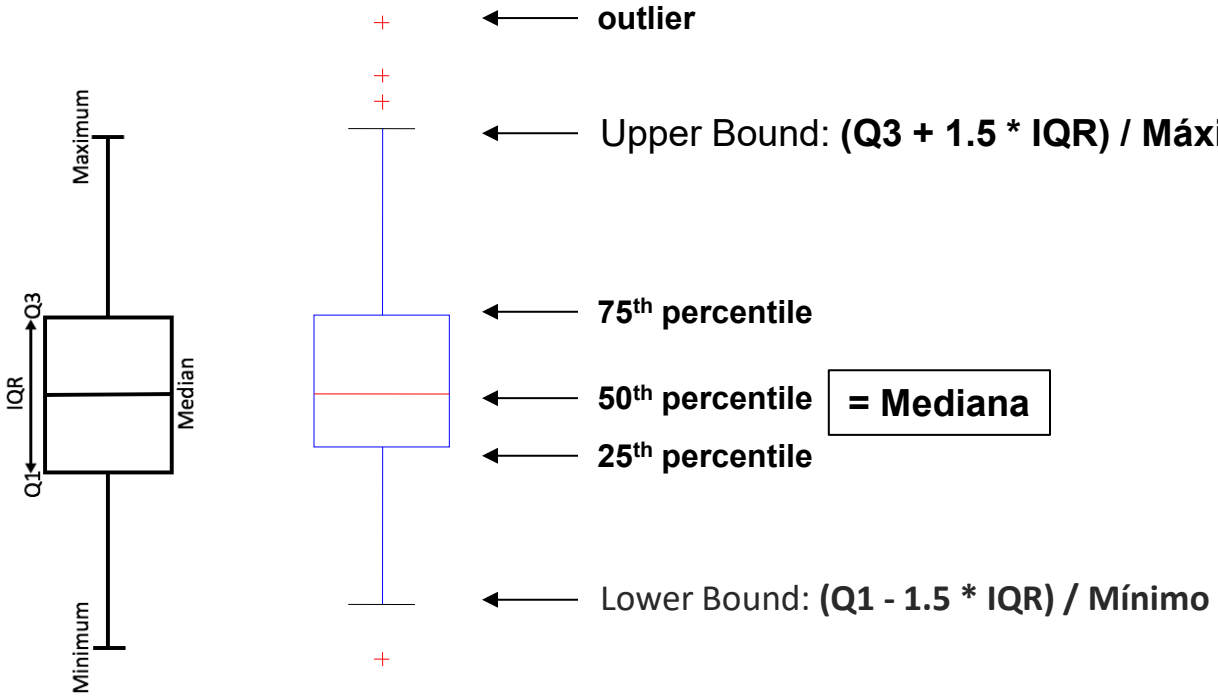




Technique: Box Plots

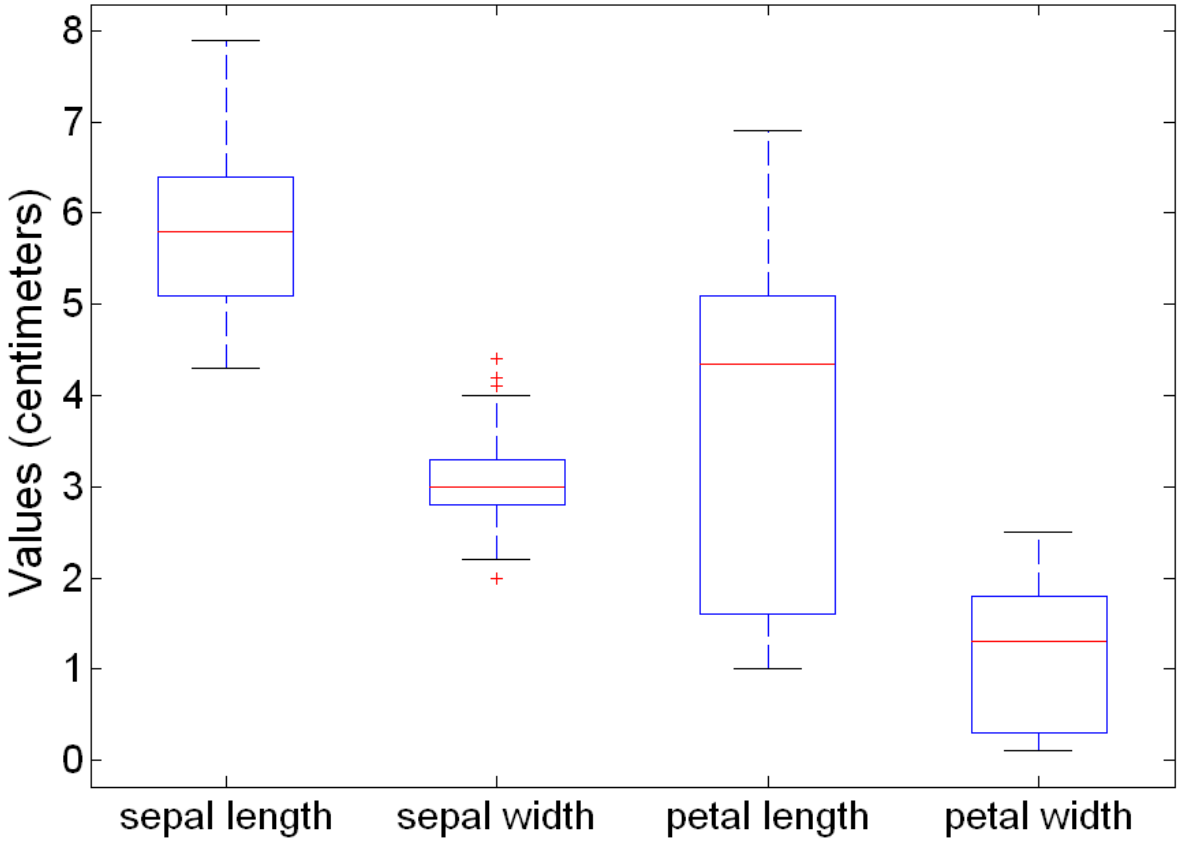
https://pt.wikipedia.org/wiki/Diagrama_de_caixa

- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



Example of Box Plots

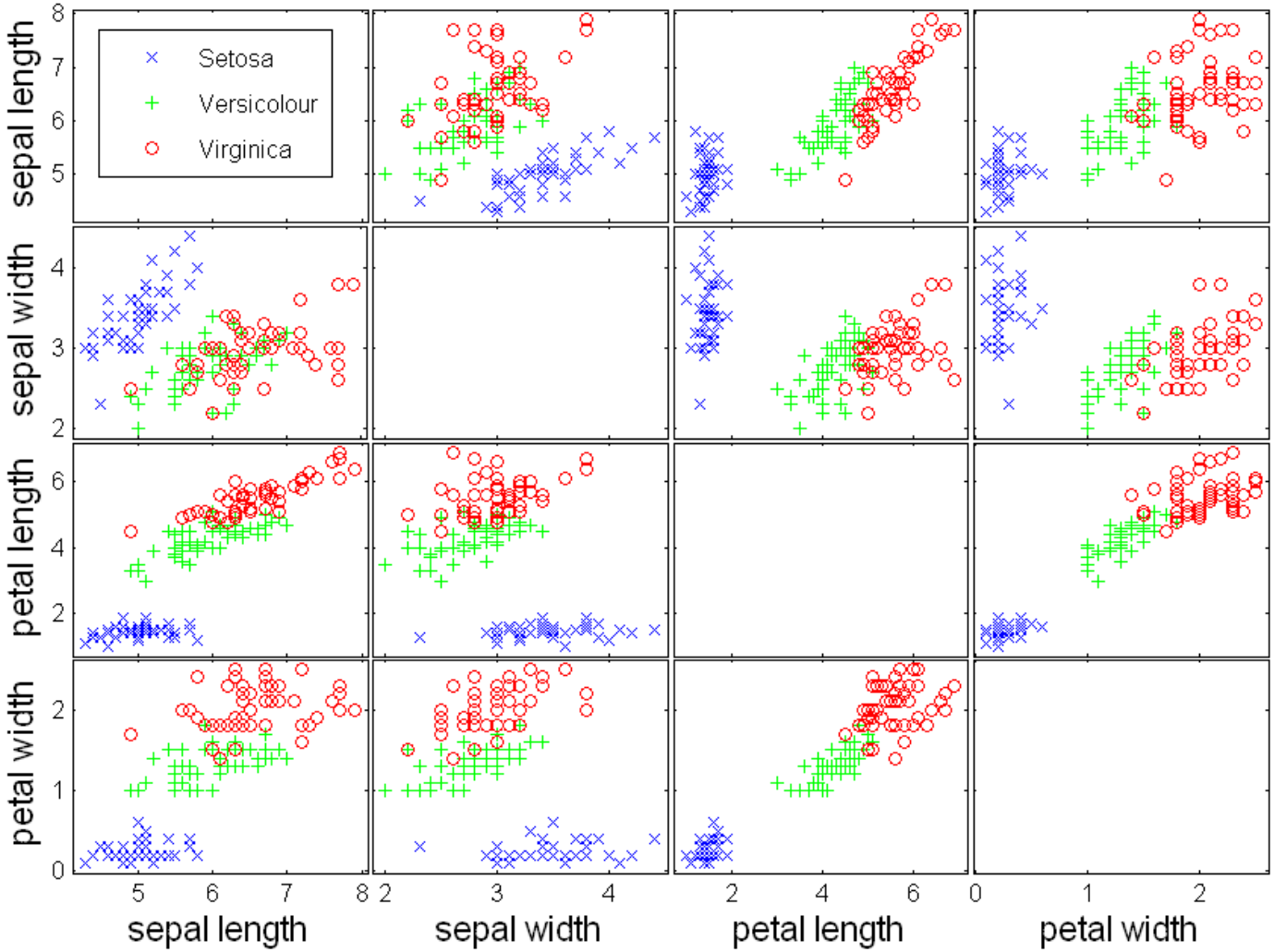
- Box plots can be used to compare attributes



Technique: Scatter Plots

- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
    - See example on the next slide

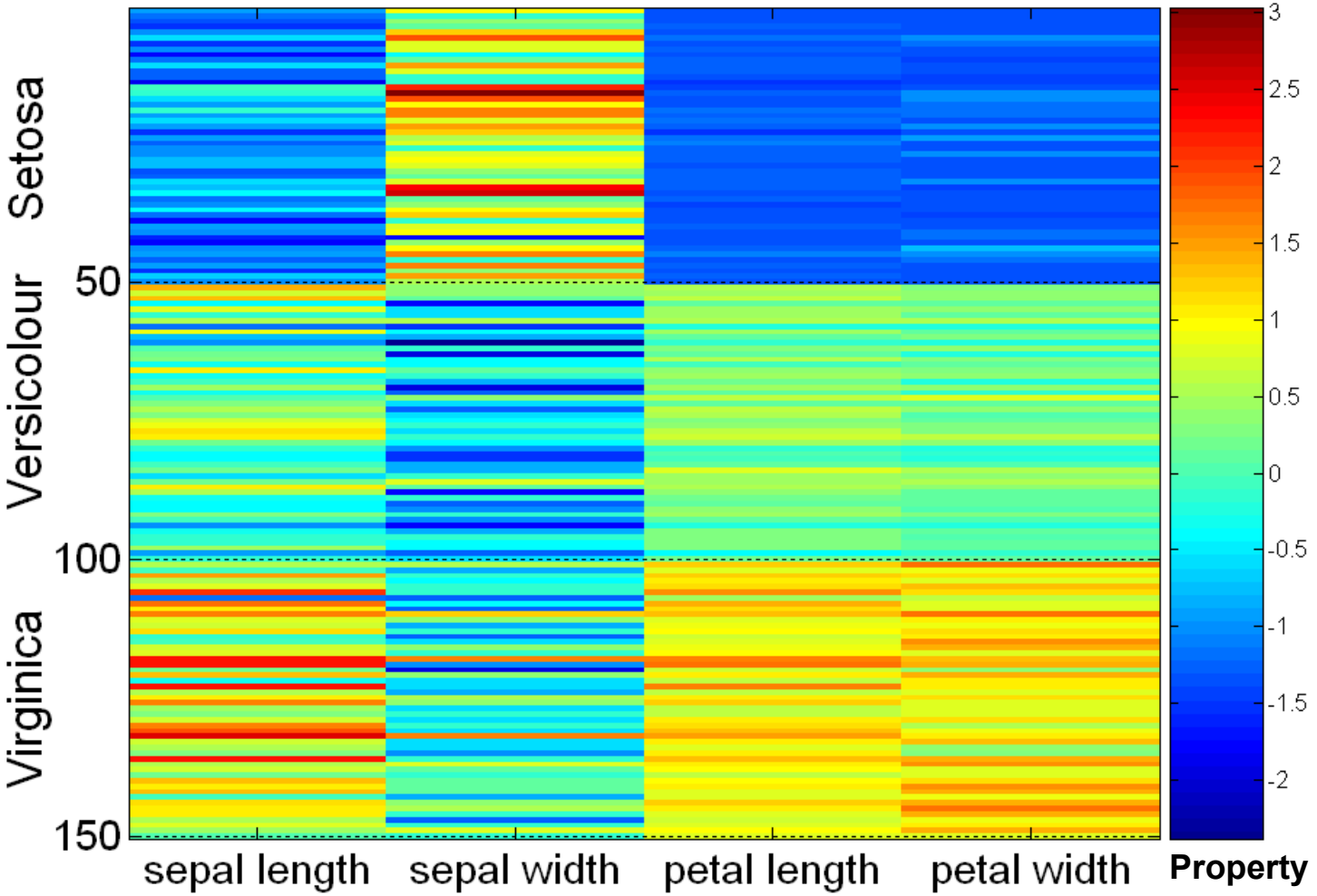
Scatter Plot Array of Iris Attributes



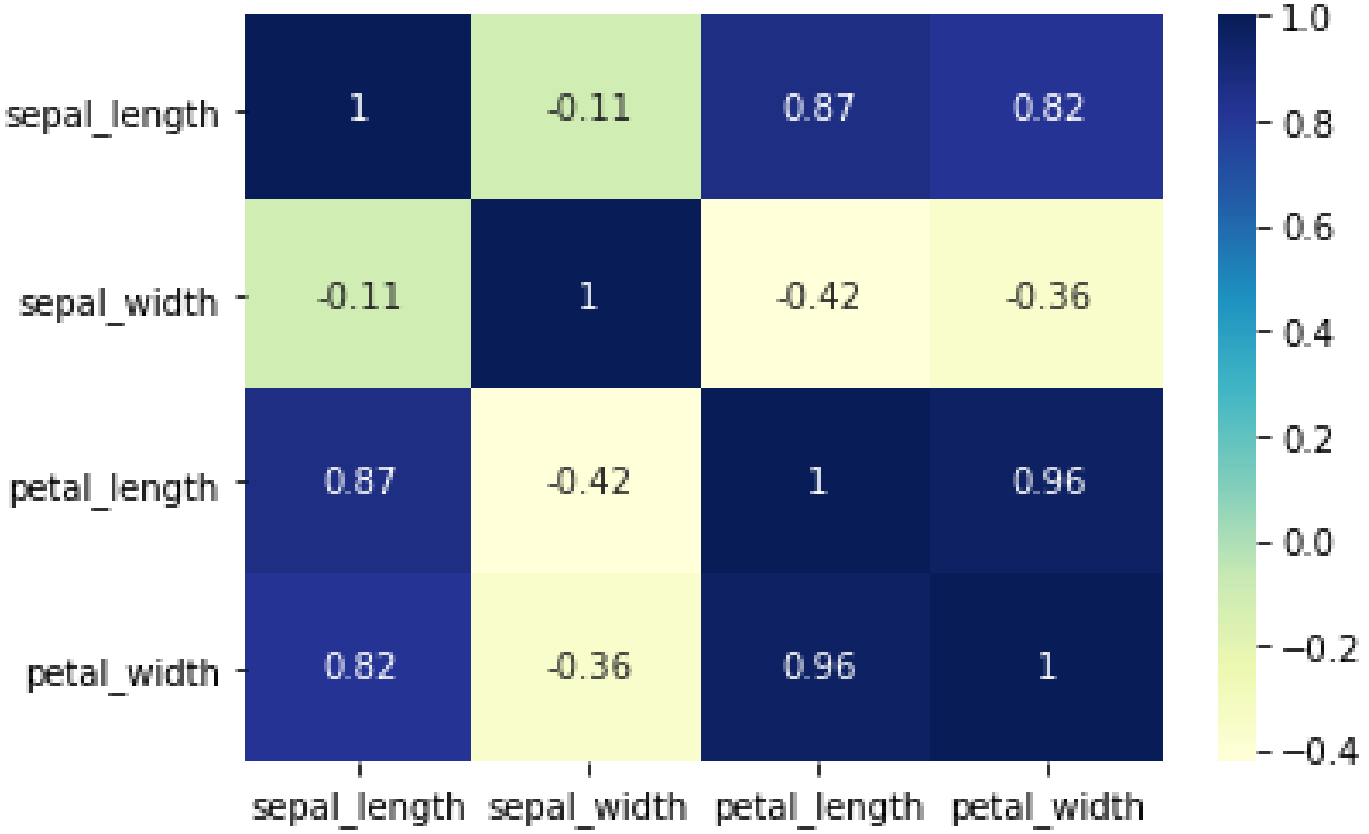
Technique: Matrix Plots

- Matrix plots
  - Can plot the data matrix
  - This can be useful when objects are sorted according to class
  - Typically, the attributes are normalized to prevent one attribute from dominating the plot
  - Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects
    - See example on the next slide

Visualization of the Iris Data Matrix



Visualization of the Iris Correlation Matrix

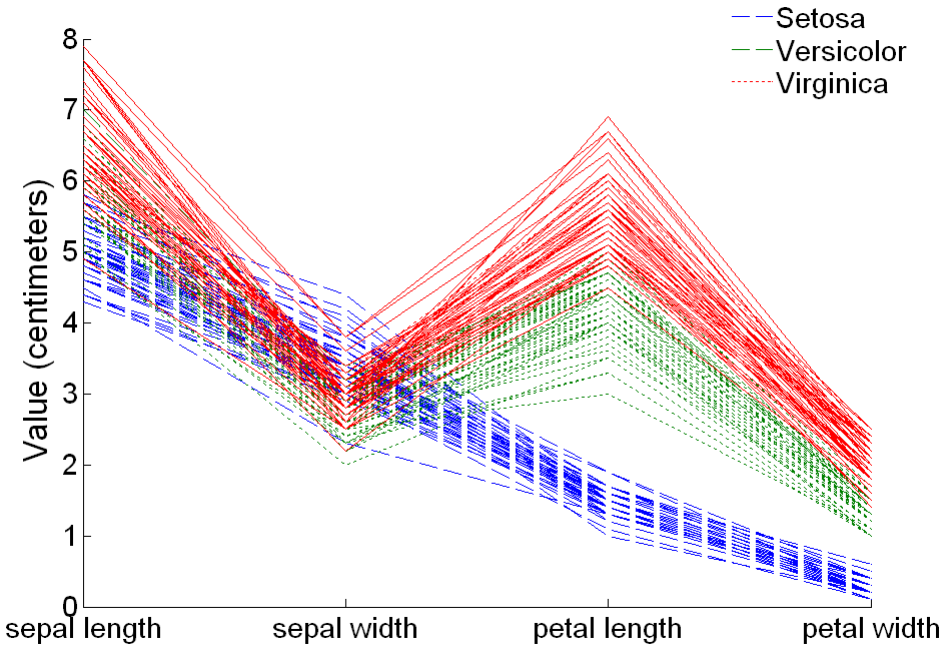


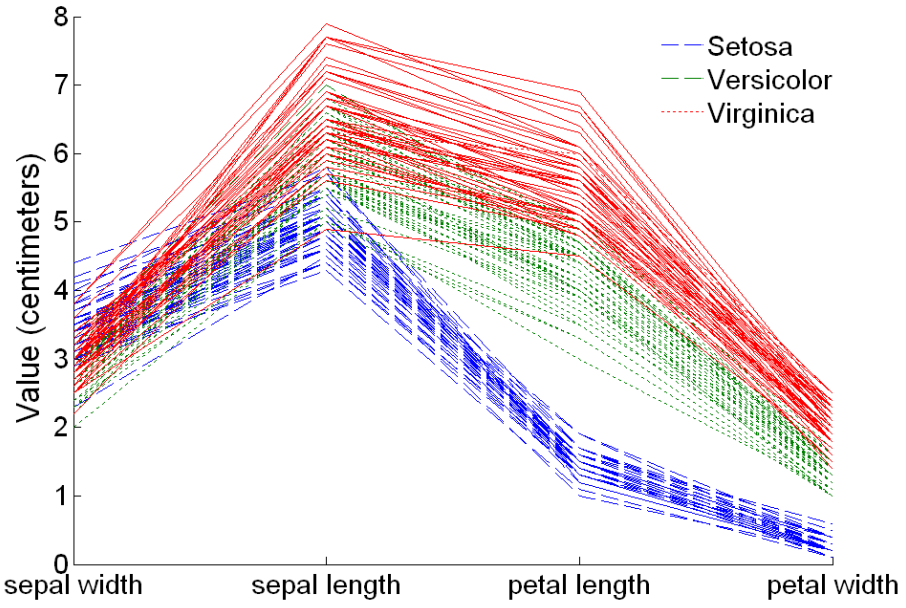
https://medium.com/@knoldus/how-to-find-correlation-value-of-categorical-variables-23de7e7a9e26

Technique: Parallel Coordinates

- Parallel Coordinates
  - Used to plot the attribute values of high-dimensional data
  - Instead of using perpendicular axes, use a set of parallel axes
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
  - Thus, each object is represented as a line
  - Often, the lines representing a distinct class of objects group together, at least for some attributes
  - Ordering of attributes is important in seeing such groupings

Parallel Coordinates Plots for Iris Data





DATA QUALITY

***O primeiro passo, referente a detecção e correção de problemas de qualidade de dados, é chamado de limpeza de dados**

Data Quality

- Poor data quality negatively affects many data processing efforts

- Data mining example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

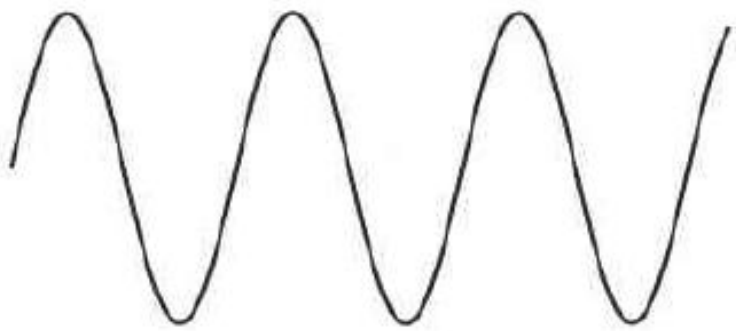
Data Quality

- What kinds of data quality problems?

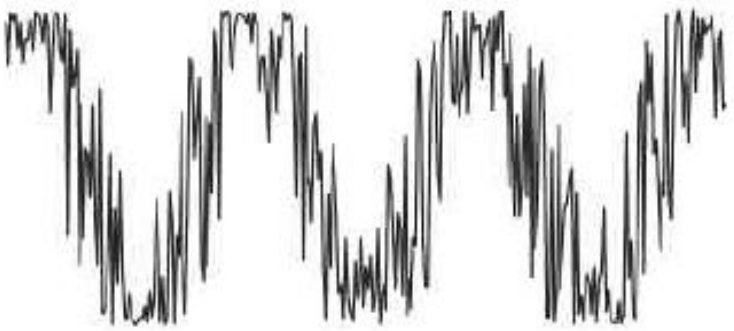
- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



(a) Time series.

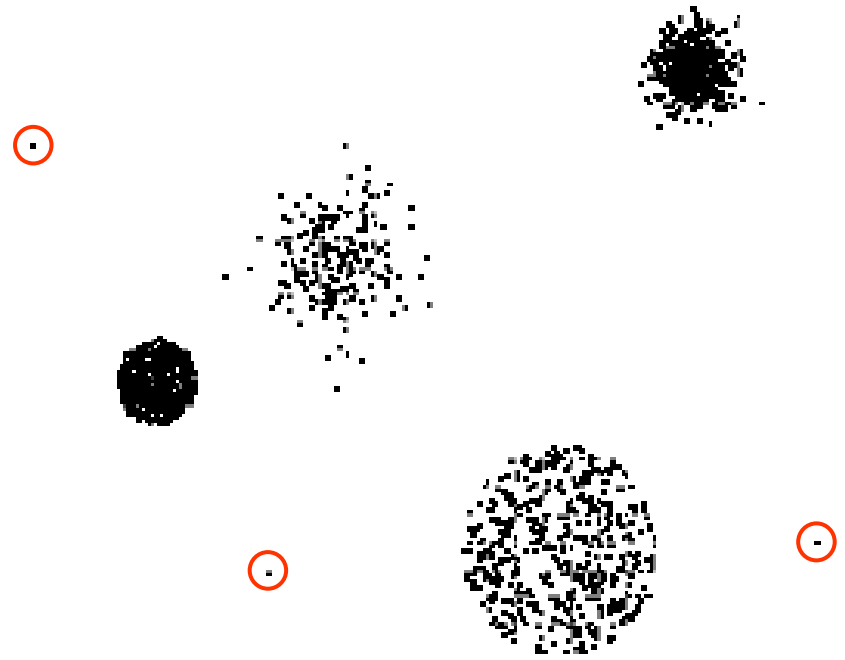


(b) Time series with noise.

Figure 2.5. Noise in a time series context.

Outliers

- ***Outliers*** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection



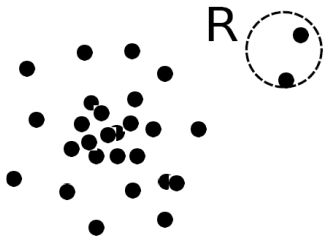
Outliers (Types)

Han, Kamber, and Pei (2011)

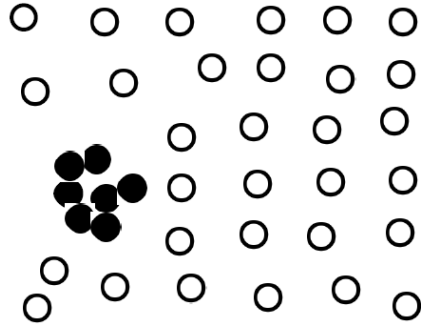
- **Global outlier** (or **point anomaly**)
  - Object is O_g if it significantly deviates from the rest of the data set
  - Ex. Intrusion detection in computer networks
- **Contextual outlier** (or *conditional outlier*)
  - Object is O_c if it deviates significantly based on a selected context
  - Ex. 80^o F in Urbana: outlier? (depending on summer or winter?)
- **Collective Outliers**
  - A subset of data objects *collectively* deviate significantly from the whole data set, even if the individual data objects may not be outliers
  - Ex. Intrusion detection: when a number of computers keep sending denial-of-service packages to each other. The computers involved may be suspected of being compromised by an attack.

Outliers (Types)

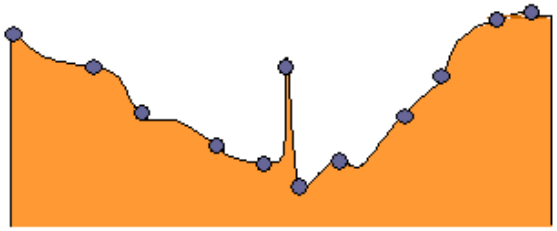
Han, Kamber, and Pei (2011)



Global Outlier



Collective Outlier



Local Outlier

A local outlier is a sample point that has a value within the normal range for the entire dataset, but if you look at the surrounding points, it is unusually high or low

https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/looking-for-global-and-local-outliers.htm#:~:text=A%20global%20outlier%20is%20a,may%20be%20a%20global%20outlier.

A data set may have multiple types of outlier

One object may belong to more than one type of outlier

https://blog.dp6.com.br/t%C3%A9cnicas-de-detec%C3%A7%C3%A3o-de-anomalias-3d9e216bf82e

Outliers

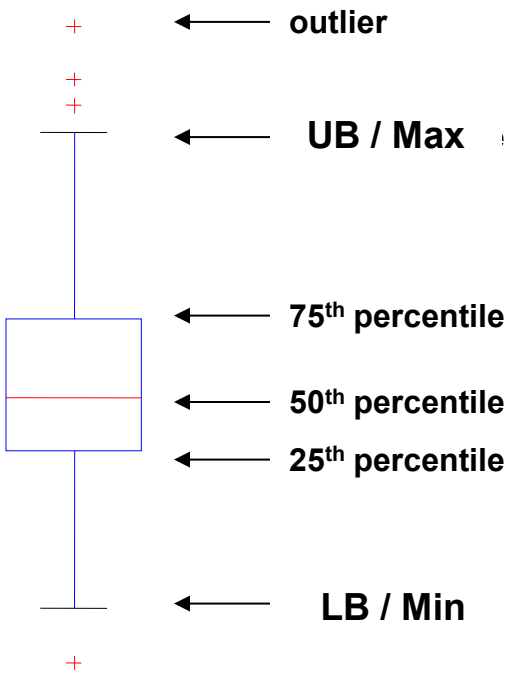
Han, Kamber, and Pei (2011)

- Two ways to categorize outlier detection methods:
  - Based on whether user-labeled examples of outliers can be obtained:
    - Supervised, semi-supervised vs. unsupervised (clustering) methods
  - Based on assumptions about normal data and outliers:
    - Statistical (Z-Score), proximity-based (Local Outlier Factor (LOF)), and clustering-based methods (**DBSCAN**)

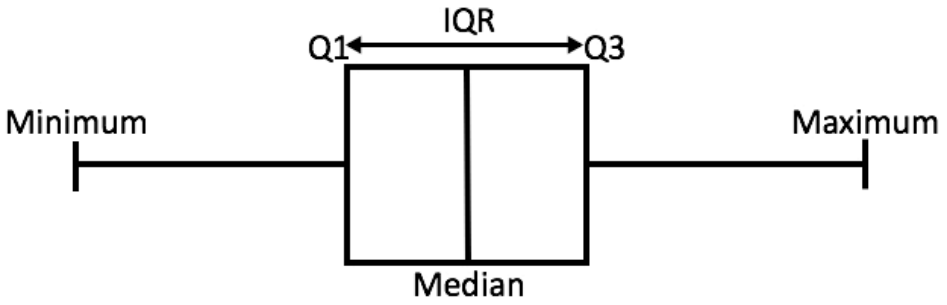
https://pyod.readthedocs.io/en/latest/

Outliers: IQR (Graphical Outlier Detection)

- Box Plot



Inter-Quartile Range or IQR



Lower Bound: **(Q1 - 1.5 * IQR) → Oulier < LB**
Upper Bound: **(Q3 + 1.5 * IQR) → Outlier > UB**

Missing Values

- It is not unusual for an object to be missing one or more attribute values
- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate data objects or variables
  - Estimate missing values
  - Ignore the missing value during analysis

Missing Values

- Handling missing values
  - Eliminate data objects or variables
    - Cuidado para não eliminar demais e tornar a análise inviável – mesmo objetos/variáveis com valores parciais armazenam algum tipo de informação
  - Estimate missing values (imputação, próximo slide)
    - Média (numérico), Moda (categórico), k-NN – porém depende do mecanismo associado a ocorrência – depende do tipo do dado ausente
  - Ignore the missing value during analysis
    - Por exemplo, em um agrupamento, no cálculo da distância, poderia se ignorar atributos com valores ausentes – desde que o número de atributos não seja tão pequeno ou o número de valores ausentes entre os objetos não seja muito alto – distância aproximada

Missing Values

- Proposta por Rubin, há três mecanismos distintos: MCAR, MAR, MNAR
  - A classificação está relacionada ao impacto da ausência de informação (motivo que levou ao não preenchimento da informação) e à escolha da abordagem mais apropriada para análise dos dados

- Missing completely at random (MCAR)
  - Missingness of a value is independent of attributes
  - Fill in values based on the attribute
  - Ex. a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck

Missing Values

Dados ausentes em avaliações educacionais: Comparação de métodos de tratamento [http://publicacoes.fcc.org.br/index.php/eae/article/view/3916/3560]

- Missing at Random (MAR)
  - Missingness is related to other variables
  - Fill in values based other values
  - Ex. if men are more likely to tell you their weight than women, weight is MAR (peso e sexo estão relacionados)

- Missing Not at Random (MNAR)
  - Missingness is related to unobserved measurements – somente as variáveis observadas não explicam a ausência de dados
    - É necessário incorporar o motivo na modelagem para realizer a imputação
  - Ex. sickest people are most likely to drop out of the study
    - Geralmente se avalia outros aspectos (notas, frequências), mas não de uma característica/condição da pessoa que não foi mensurada

https://cran.r-project.org/web/packages/finalfit/vignettes/missing.html

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources

- Examples:
  - Same person with multiple email addresses
    - If there are two objects that actually represent a single object, then the values of corresponding attributes may differ, and these inconsistent values must be resolved
    - Care needs to be taken to avoid accidentally combining data objects that are similar, but not duplicates, such as two distinct people with identical name

DATA PREPROCESSING

Data Preprocessing

- Aggregation
- Sampling
- Feature subset selection
- Dimensionality Reduction
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Data Preprocessing

- Aggregation
- Sampling
- Feature subset selection
- Dimensionality Reduction
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

- Combining two or more objects into a single object - foco em uma visão agregada (mais geral) - OLAP - perdem-se os detalhes
  - Agregar as transações de vendas de uma única loja – visão por loja de uma dada rede
  - Atribtuos são agregados, por exemplo, somando-se valores (preço) ou gerando-se conjuntos (itens)
- Purpose
  - Data reduction ("consequência")
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - More “stable” data
    - Aggregated data tends to have less variability

Aggregation

- Combining two or more objects into a single object - foco em uma visão agregada (mais geral) - OLAP - perdem-se os detalhes
  - Agregar as transações de vendas de uma única loja – visão por loja de uma dada rede
    - Atributos são agregados, por exemplo, somando-se valores (preço) ou gerando-se conjuntos (itens)

Table 2.4. Data set containing information about customer purchases.

Transaction ID	Item	Store Location	Date	Price	...
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	...
101123	Battery	Chicago	09/06/04	$5.99	...
101124	Shoes	Minneapolis	09/06/04	$75.00	...
⋮	⋮	⋮	⋮	⋮	

Conjunto **Valor único**

Sampling

- Sampling is the main technique employed for data reduction
  - It is often used for both the preliminary investigation of the data and the final data analysis

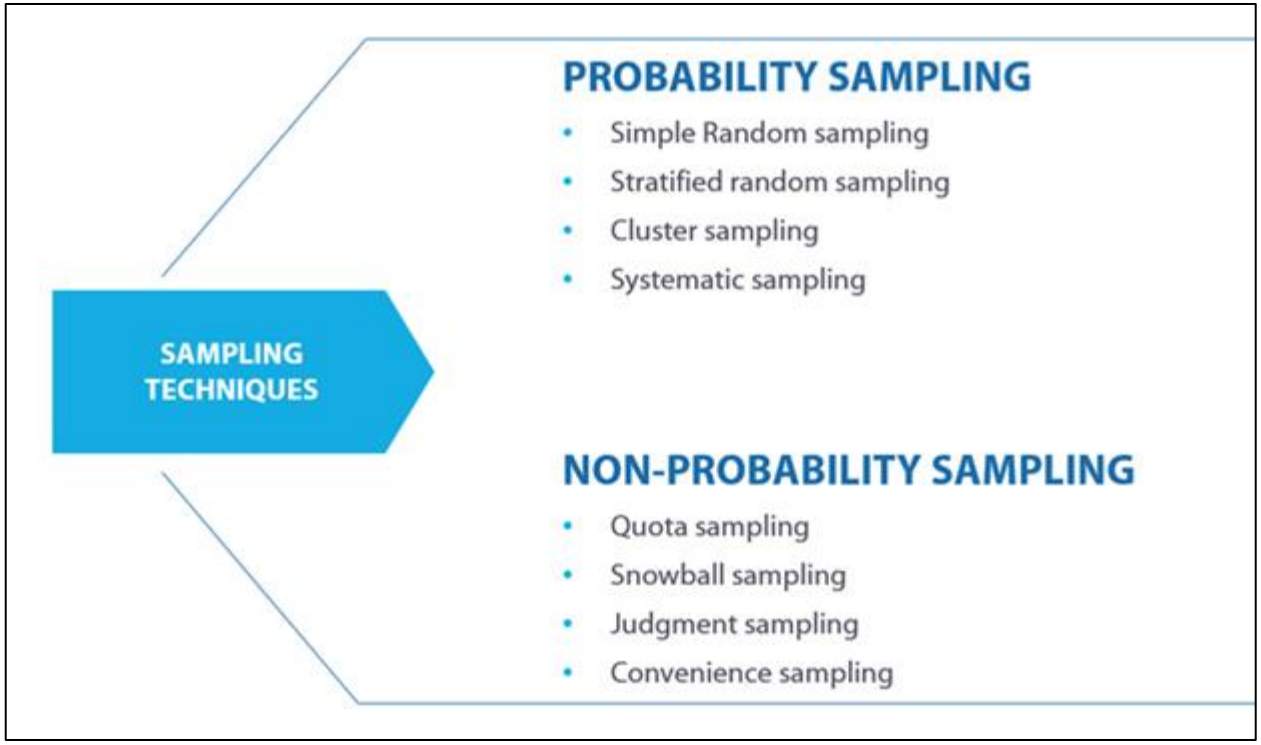
- Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming

- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming

Sampling

- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set, if the sample is representative
  - A sample is representative if it has approximately the same properties (of interest) as the original set of data
    - If the mean (average) of the data objects is the property of interest, then a sample is representative if it has a mean that is close to that of the original data

Types of Sampling



https://www.analyticsvidhya.com/blog/2021/09/a-complete-guide-on-sampling-techniques/

https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/

Types of Sampling

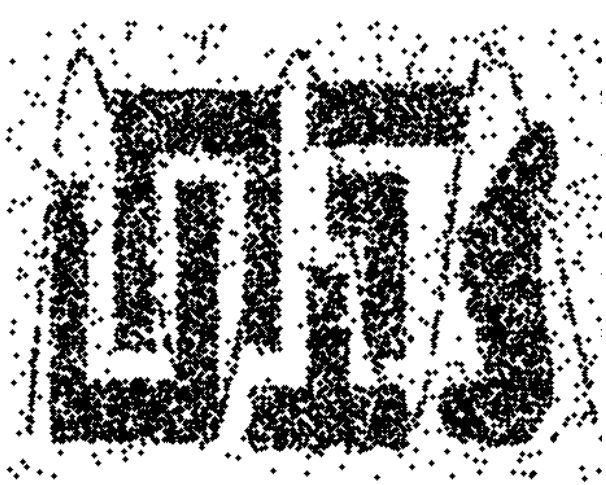
- Simple Random Sampling
  - There is an equal probability of selecting any particular item
  - Sampling without replacement
    - As each item is selected, it is removed from the population
  - Sampling with replacement
    - Objects are not removed from the population as they are selected for the sample
    - In sampling with replacement, the same object can be picked up more than once

Types of Sampling

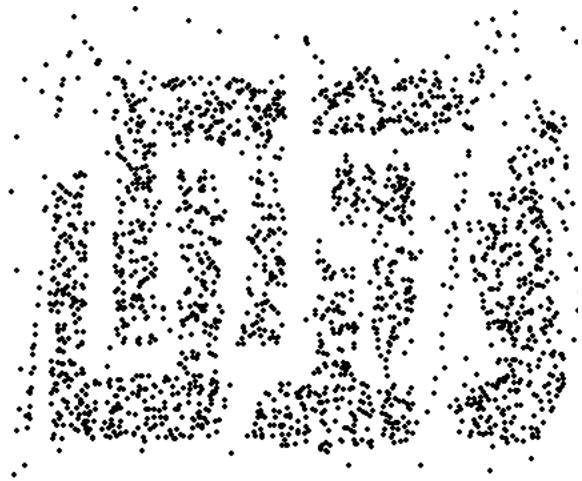
- Stratified sampling
  - Used when the population consists of different types of objects, with widely different numbers of objects, simple random sampling can fail to adequately represent those types of objects that are less frequent (imbalanced problem)
  - Split the data into several partitions; then draw random samples from each partition
    - In the simplest version, equal numbers of objects are drawn from each group even though the groups are of different sizes
    - In another variation, the number of objects drawn from each group is proportional to the size of that group

Sample Size

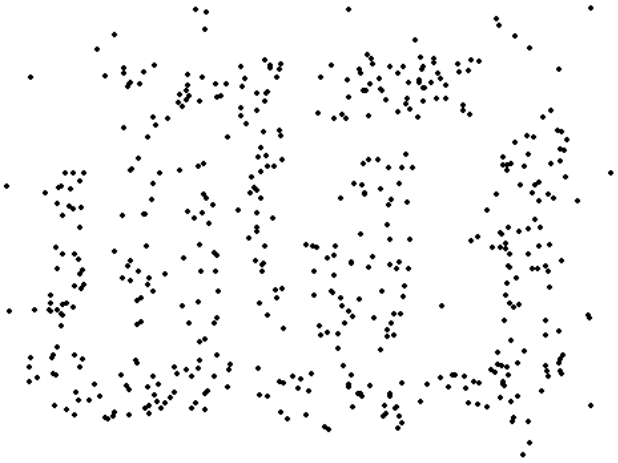
Quanto maior a amostra, mais representativa ela é...



8000 points



2000 Points



500 Points

Relação com o tamanho da amostra

Feature Subset Selection

- Another way to reduce dimensionality of data
- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid [preço de compra de um produto e o valor do imposto pago sobre as vendas]
- Irrelevant features
  - Contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA
- Many techniques developed, especially for classification

Feature Subset Selection

- There are several more reasons to complete feature selection, such as:
  - Simple models are easier to interpret
  - Shorter training time
  - Enhanced generalization by reducing overfitting
  - Variable redundancy
- Techniques:
  - Brute-force
  - Filter
  - Wrapper
  - Embedded

https://heartbeat.comet.ml/hands-on-with-feature-selection-techniques-an-introduction-1d8dc6d86c16

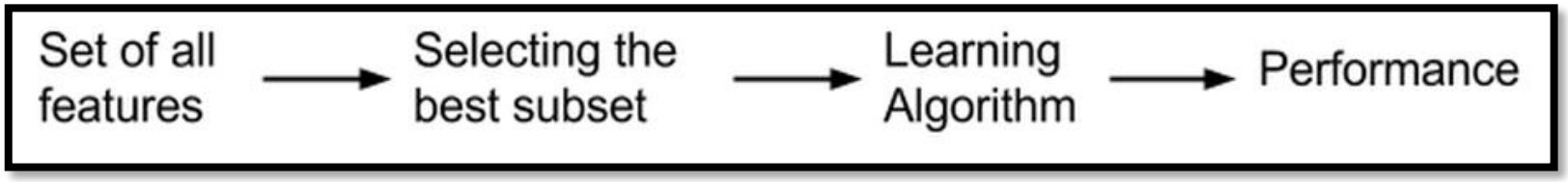
https://itmo-fs.readthedocs.io/en/latest/

Feature Subset Selection

- Techniques:
  - Brute-force approach:
    - Try all possible feature subsets as input to data mining algorithm
      - Somente para um número pequeno de atributos

Feature Subset Selection

- Techniques:
  - Filter approaches:
    - Features are selected before data mining algorithm is run



    - Basic Filter Methods (atributos com valores constantes, etc.)
    - Correlation Filter Methods (Pearson, etc.)
    - Statistical & Ranking Filter Methods (ganho informação, informação mútua, etc.)

Feature Subset Selection

- Techniques:
  - Filter approaches:
    - **Correlation** Filter Methods (Pearson, etc.)

Correlation

https://operdata.com.br/blog/coeficientes-de-correlacao/

- Os coeficientes de correlação são métodos estatísticos para se **medir as relações entre variáveis** e o que elas representam
- O que a correlação procura entender é como uma variável se comporta em um cenário onde outra está variando, visando identificar se existe alguma relação entre a variabilidade de ambas
- Embora não implique em causalidade, o coeficiente de correlação exprime em números essa relação, ou seja, quantifica a relação entre as variáveis

Correlation

https://heartbeat.comet.ml/hands-on-with-feature-selection-techniques-an-introduction-1d8dc6d86c16

- **Pearson correlation coefficient [-1, +1]**
  - It's used to summarize the strength of the **linear relationship** between two data variables
  - The assumptions that the Pearson correlation coefficient makes:
    - Both variables should be normally distributed
    - A straight-line relationship between the two variables
    - Data is equally distributed around the regression line

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$$

https://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_Pearson

Correlation

https://heartbeat.comet.ml/hands-on-with-feature-selection-techniques-an-introduction-1d8dc6d86c16

- **Pearson correlation coefficient [-1, +1]**
  - A covariância mede a relação linear entre duas variáveis. A covariância é semelhante à correlação entre duas variáveis; contudo:
    - Os coeficientes de correlação são padronizados (covariância normalizada)
    - Os valores de covariância não são padronizados ([-inf, +inf])

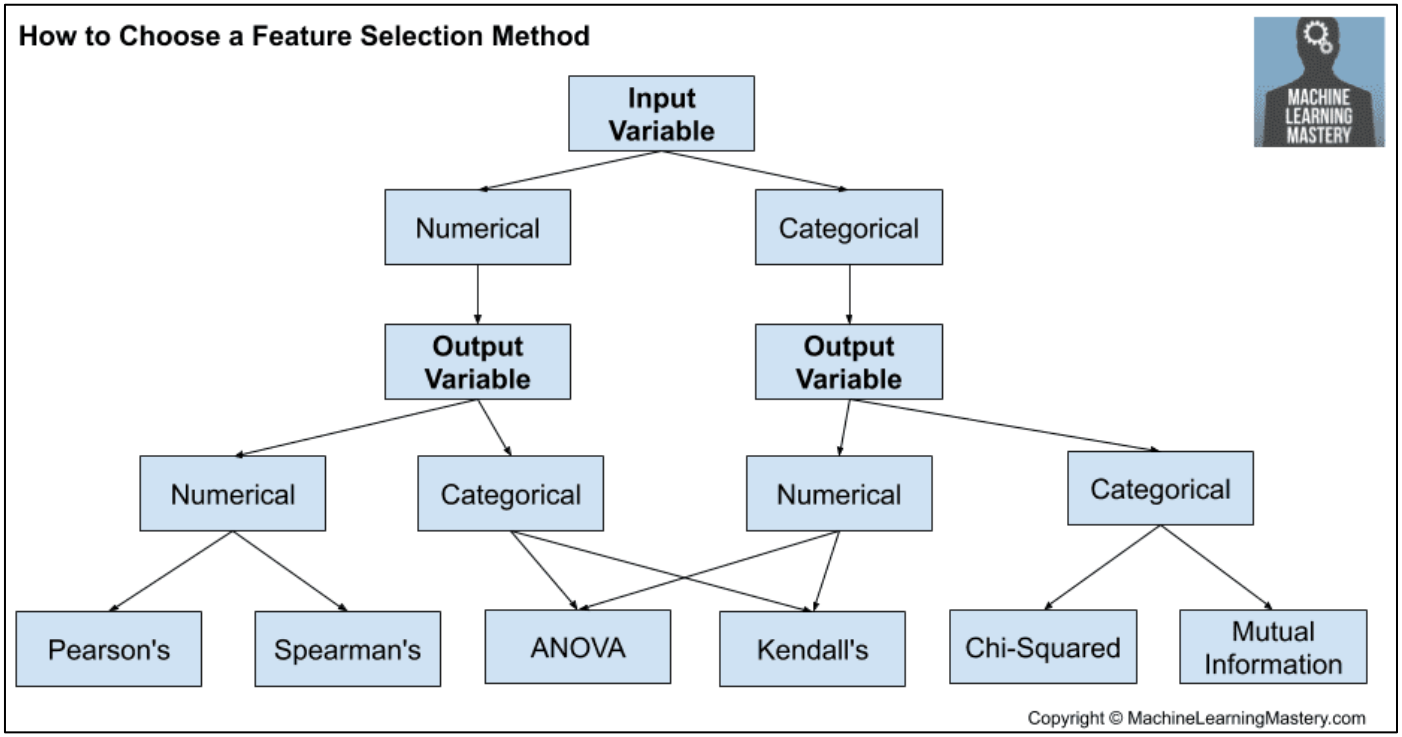
https://support.minitab.com/pt-br/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/correlation-and-covariance/what-is-covariance/#:~:text=A%20covari%C3%A2ncia%20mede%20a%20rela%C3%A7%C3%A3o,um%20coeficiente%20de%20correla%C3%A7%C3%A3o%201.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$$

https://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_Pearson

Correlation

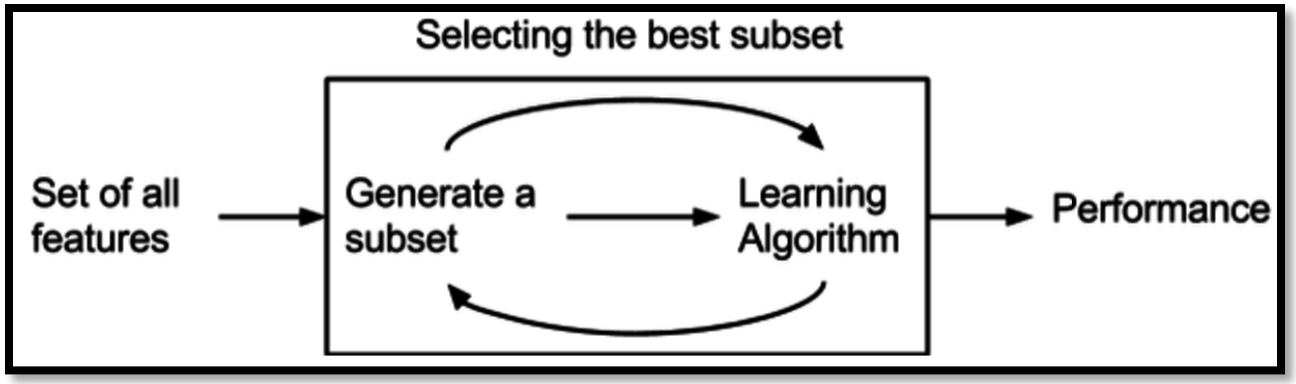
https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/#:~:text=Feature%20selection%20methods%20are%20intended,redundant%20predictors%20from%20the%20model



https://medium.com/@knoldus/how-to-find-correlation-value-of-categorical-variables-23de7e7a9e26

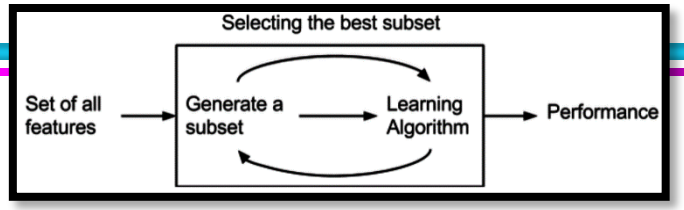
Feature Subset Selection

- Techniques:
  - Wrapper ("empacotadas") approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes



  - Very computationally expensive. However, they provide the best performing feature subset for a given machine learning algorithm (próximo slide).
  - Abordagem gulosa: uma vez tomada uma decisão não se volta atrás.

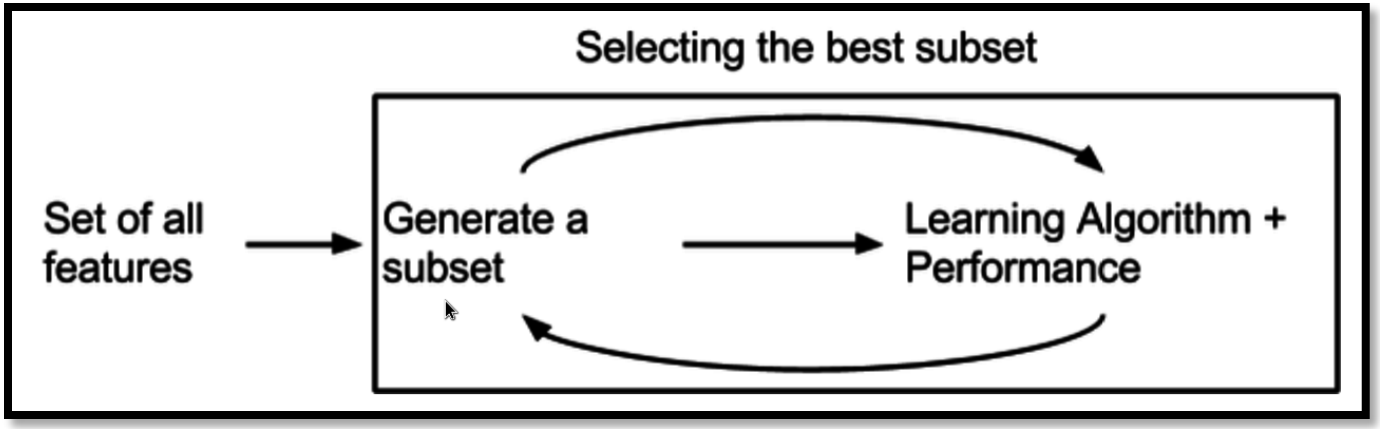
Feature Subset Selection



- Techniques:
  - Wrapper (“empacotadas”) approaches steps:
    - Search for a subset of features: using a search method (forward (de 1 a n), backward (de n a 1), bidirectional (ambos os sentidos)), select a subset of features from the available ones
    - Build a machine learning model: a chosen ML algorithm is trained on the previously-selected subset of features
    - Evaluate model performance: evaluate the newly-trained ML model with a chosen metric
    - Repeat: The whole process starts again with a new subset of features, a new ML model trained, and so on
      - Critério de parada: model performance decreases/increases, a predefined number of features is reached, etc.

Feature Subset Selection

- Techniques:
  - Embedded ("embutidas") approaches:
    - Feature selection occurs naturally as part of the data mining algorithm

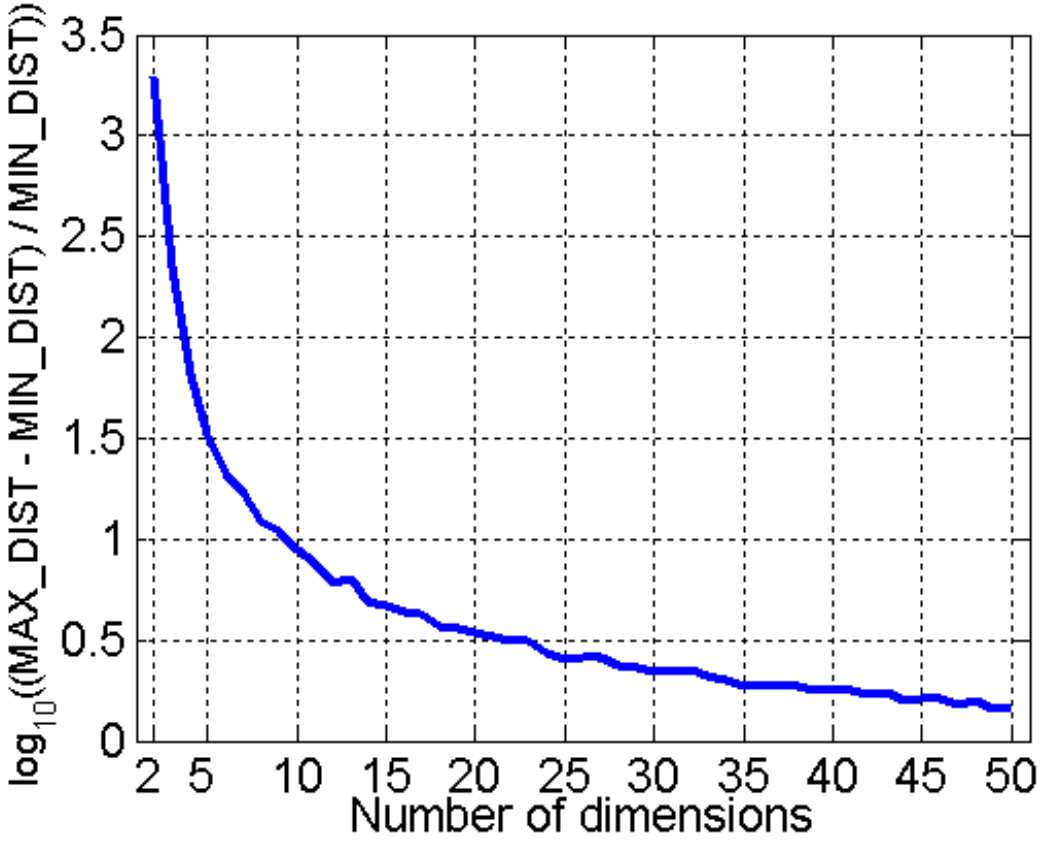


Figuras extraídas de https://en.wikipedia.org/wiki/Feature_selection#cite_note-42

More on https://heartbeat.fritz.ai/hands-on-with-feature-selection-techniques-an-introduction-1d8dc6d86c16

Dimensionality Reduction

- **Curse of Dimensionality**
  - When dimensionality increases, data becomes increasingly sparse in the space that it occupies
  - Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful



- **Randomly generate 500 points**
- **Compute difference between max and min distance between any pair of points**

Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

Dimensionality Reduction

- Techniques
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Others: supervised and non-linear techniques

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.decomposition
https://scikit-learn.org/stable/modules/classes.html#module-sklearn.manifold

- The term dimensionality reduction is often reserved for those techniques that reduce the dimensionality of a data set by creating new attributes that are a combination of the old attributes (“criação”)
- The reduction of dimensionality by selecting new attributes that are a subset of the old is known as feature subset selection or feature selection (“seleção”)

Dimensionality Reduction

- Principal Components Analysis (PCA)
  - Dadas n variáveis, V_1, V_2, …, V_n, busca-se encontrar combinações lineares entre elas para produzir varáveis não correlacionadas entre si (cada componente C_i representa uma variável)
    - $C_i = coef_1 * V_1 + coef_2 * V_2 + \ldots + coef_n * V_n$
    - As componentes são direções ortogonais qua capturam variação nos dados. Cada componente “cobre” um total da variância dos dados (todas=100%)

Principal Components Analysis (PCA)

- Used for reducing the dimensionality of data while preserving as much as possible of the information contained in the original data
  - PCA achieves this goal by projecting data onto a lower-dimensional subspace that retains most of the variance among the data points

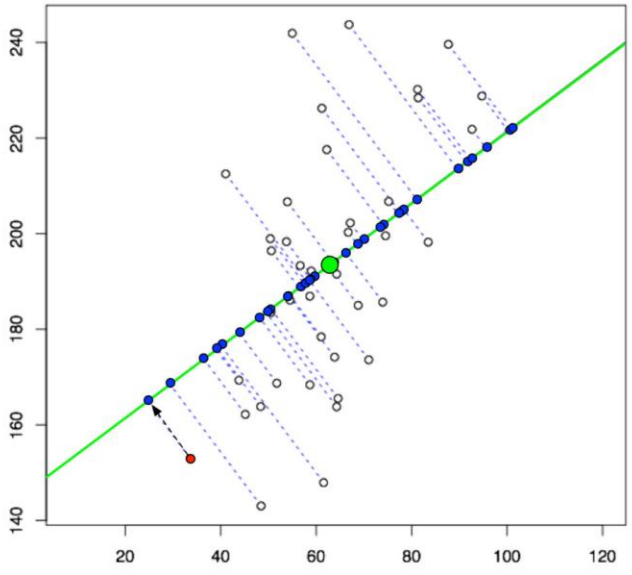
https://programmathically.com/principal-components-analysis-explained-for-dummies/

Principal Components Analysis (PCA)

- What is dimensionality reduction, and what is a subspace?
  - If you have data in a 2-dimensional space, you could project all the data points onto a line using PCA

You have essentially reduced the dimensionality of your data from 2D to 1D

The 1D space (your line) is a subspace of the 2D coordinate system



https://programmathically.com/principal-components-analysis-explained-for-dummies/

Principal Components Analysis (PCA)

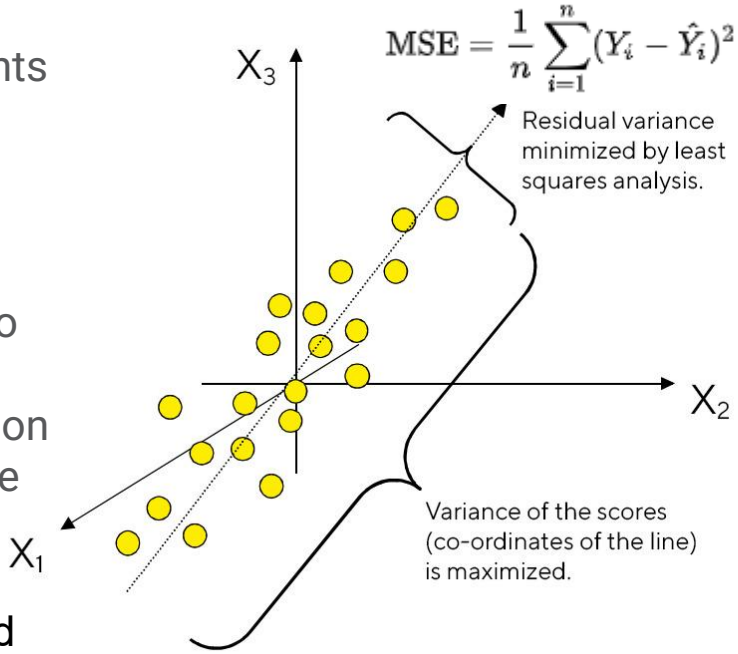
- What is dimensionality reduction, and what is a subspace?

1.The total distance among the projected points is maximum. This means they can be distinguished from one another as clearly as possible

2.The total distance from the original points to their corresponding projected points is minimum. This means we have a representation that is as close to the original data as possible

In other words, the best line must convey the *maximum variation* among data points and contain *minimum error*

https://blog.bioturing.com/2018/06/14/principal-component-analysis-explained-simply/



https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186

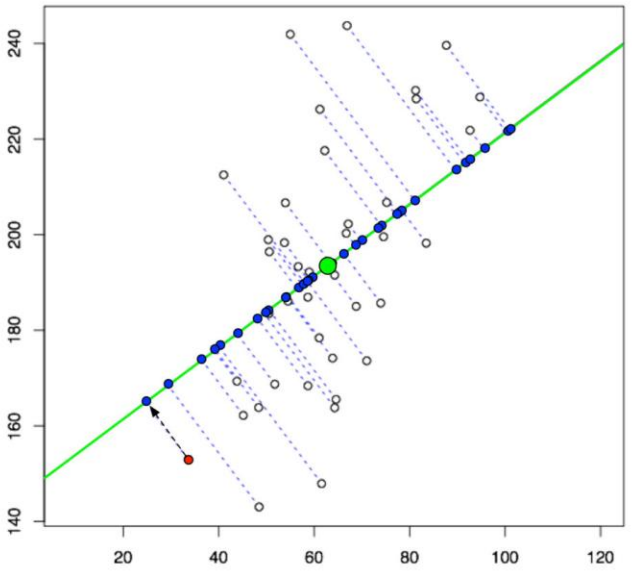
Principal Components Analysis (PCA)

- The green line has been constructed using mathematical optimization to maximize the variance between the data points as much as possible along that line

The points on the line are still closer to each other than in the original 2D space because you are losing a dimension to distinguish them

However, the simplification achieved by dimensionality reduction outweighs the loss in information

Principal components are less interpretable and don't have any real meaning since they are constructed as linear combinations of the initial variables



https://programmathically.com/principal-components-analysis-explained-for-dummies/

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

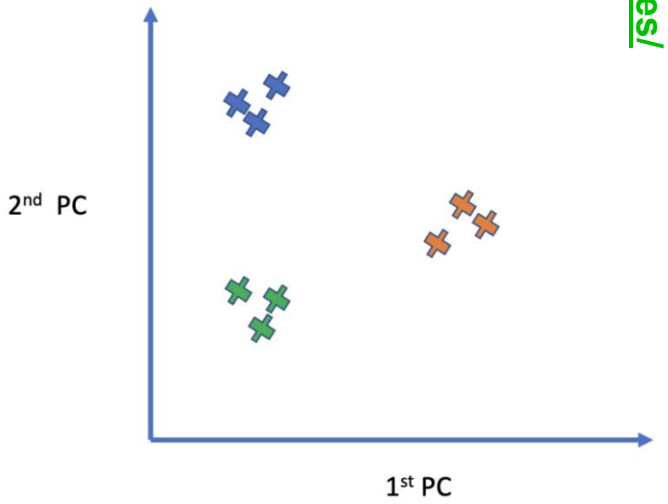
Principal Components Analysis (PCA)

- You'll likely use several principal components because the variance explained by one principal component is often insufficient
- Principal components are vectors that are orthogonal to each other
  - This means they form a 90-degree angle. Mathematically, orthogonal vectors are independent, meaning the variance explained by the second principal component does not overlap with the variance of the first. So they represent information as efficiently as possible.
- The first principal component will capture most of the variance; the second principal component will capture the second-largest part of the variance that has been left unexplained by the first one, etc.

https://programmathically.com/principal-components-analysis-explained-for-dummies/

Principal Components Analysis (PCA)

- If you plot your data in your lower-dimensional subspace with the principal components as your axes, similar data points should cluster together.
  - This happens because you are explicitly focusing on axes that maximize the variance

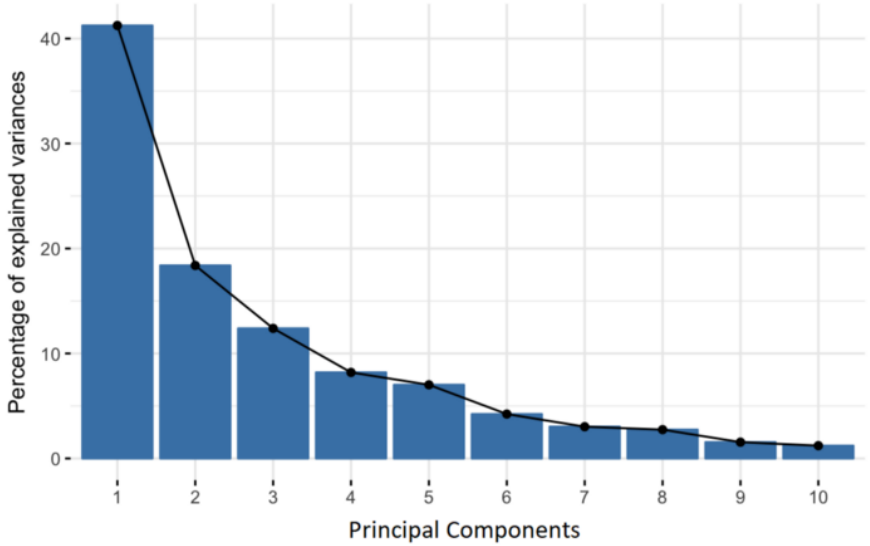


https://programmathically.com/principal-components-analysis-explained-for-dummies/

Principal Components Analysis (PCA)

- In highly dimensional datasets, the vast majority of the variance in the data is often captured by a small number of principal components.
  - A plot of the distribution of the variance across principal components may look like this.

https://programmathically.com/principal-components-analysis-explained-for-dummies/



https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal Components Analysis (PCA)

- Step 1: Standardization
  - The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis
    - Once the standardization is done, all the variables will be transformed to the same scale

$$z = \frac{value - mean}{standard\ deviation}$$

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal Components Analysis (PCA)

- Step 2: Covariance Matrix Computation
  - The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other
    - The idea is to see if there is any relationship between them

    [Identify if variables are highly correlated in such a way that they contain redundant information]

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal Components Analysis (PCA)

- Step 3: Compute The Eigenvectors And Eigenvalues Of The Covariance Matrix To Identify The Principal Components
  - Geometrically speaking, principal components represent the directions of the data that explain a **maximal amount of variance**, that is to say, the lines that capture most information of the data
  - The relationship between variance and information here, is that, the larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal Components Analysis (PCA)

- Step 3: Compute The Eigenvectors And Eigenvalues Of The Covariance Matrix To Identify The Principal Components
  - The eigenvectors of the Covariance matrix are *the directions of the axes where there is the most variance* (most information) and that we call Principal Components
  - The eigenvalues are simply the coefficients attached to eigenvectors, which give the *amount of variance carried in each Principal Component*
  - By ranking your eigenvectors in order of their eigenvalues, highest to lowest, you get the principal components in order of significance

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal Components Analysis (PCA)

- Step 3: Compute The Eigenvectors And Eigenvalues Of The Covariance Matrix To Identify The Principal Components

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

1° Componente
1.28/1.33 = 0.96 = 96% variância

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

2° Componente
0.05/1.33 = 4% variância

λ1 + λ2 = 1.28 + 0.05 = 1.33

* For a 2-dimensional data set, there are 2 variables, therefore there are 2 eigenvectors with 2 corresponding eigenvalues

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal Components Analysis (PCA)

- Step 4: Feature Vector
  - The feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep
    - This makes it the first step towards dimensionality reduction, because if we choose to keep only ***p*** eigenvectors (components) out of ***n***, the final data set will have only ***p*** dimensions

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix}$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{bmatrix} \Rightarrow \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix}$$

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal Components Analysis (PCA)

- Step 5: Recast The Data Along The Principal Components Axes
  - The aim is to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the principal components

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal Components Analysis (PCA)

- PCA transformations are linear transformations
  - PCA is a linear method
  - Kernel PCA was developed for data whose decision boundaries are described by non-linear function
    - https://www.geeksforgeeks.org/ml-introduction-to-kernel-pca/
- PCA from Scratch in Python
  - https://machinelearningmastery.com/calculate-principal-component-analysis-scratch-python/
  - https://www.kaggle.com/code/nirajvermafcb/principal-component-analysis-explained/notebook

Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies:
  - Feature extraction (domain-specific): Creation of a new set of features from the original raw data
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: PCA, Fourier and wavelet analysis

Discretization

- Discretization is the process of converting a numeric attribute into an categorical attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification and association analysis
  - Many classification algorithms work best if both the independent and dependent variables have only a few values
  - Discretization involves two subtasks: (a) decide how many categories to have; (b) determine how to map the values of the continuous attribute to these categories

Discretization

- Discretization is the process of converting a numeric attribute into an categorical attribute
  - Involves two subtasks:
    - (a) decide how many categories to have
      - After the values of the continuous attribute are sorted, they are then divided into n intervals by specifying n-1 split points
    - (b) determine how to map the values of the continuous attribute to these categories
      - All the values in one interval are mapped to the same categorical value (trivial step)
    - Therefore, the problem of discretization is one of deciding how many split points to choose and where to place them

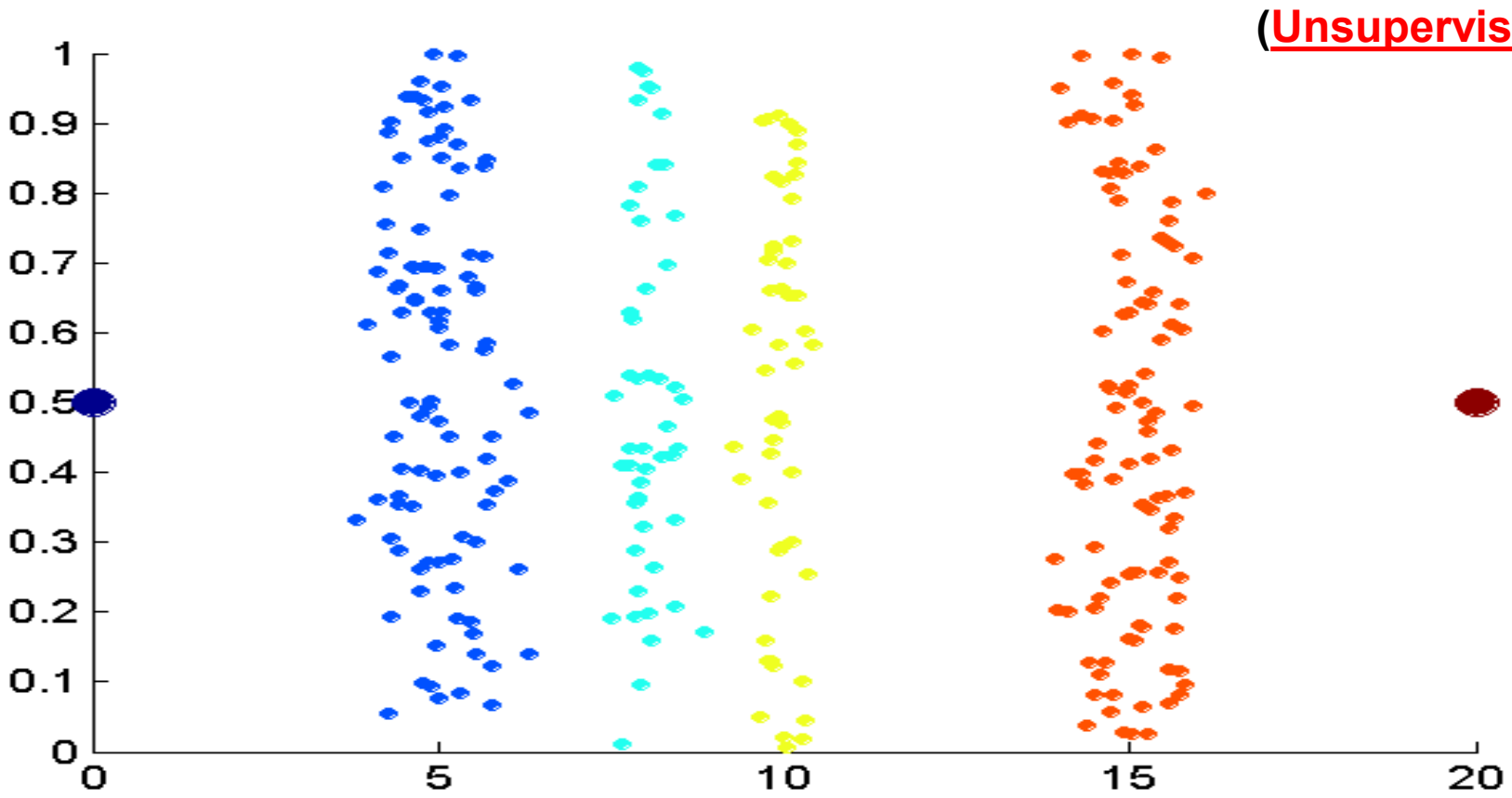
Discretization

- Discretization is the process of converting a numeric attribute into an categorical attribute
  - Therefore, the problem of discretization is one of deciding how many split points to choose and where to place them.
    - Unsupervised discretization: find breaks in the data values
    - Supervised discretization: use class labels to find breaks (CAIM, ur-CAIM)

https://sci2s.ugr.es/keel/algorithms.php#discretization

More on "A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning", 2013

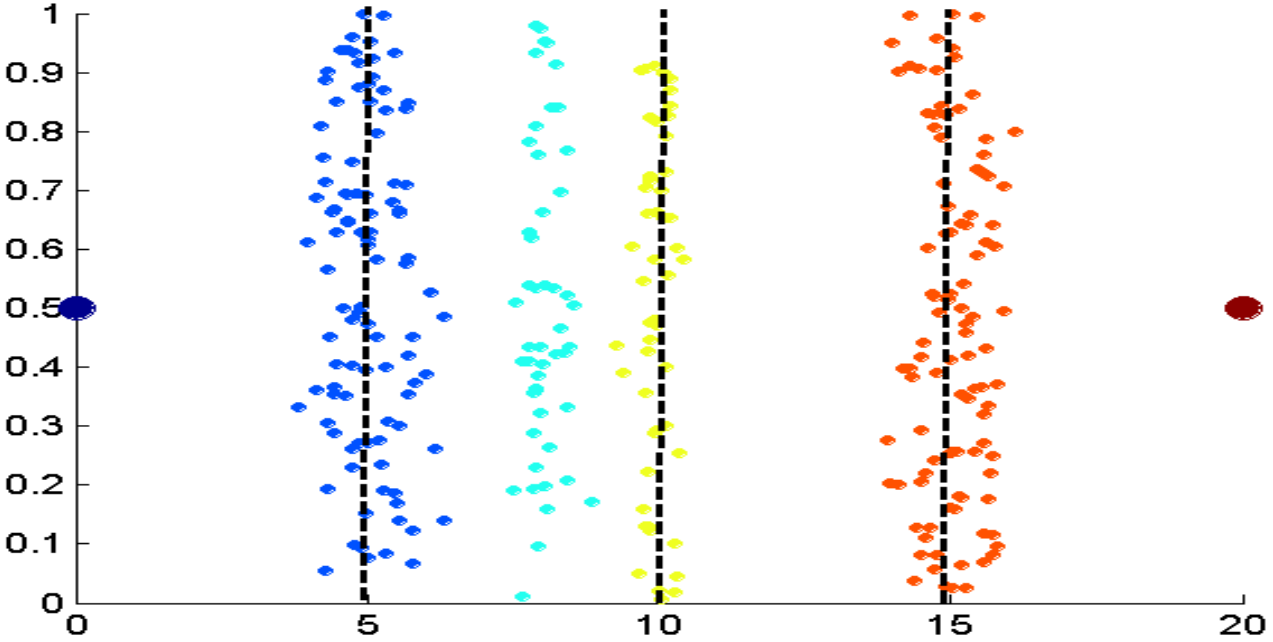
Discretization Without Using Class Labels



Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Discretization Without Using Class Labels

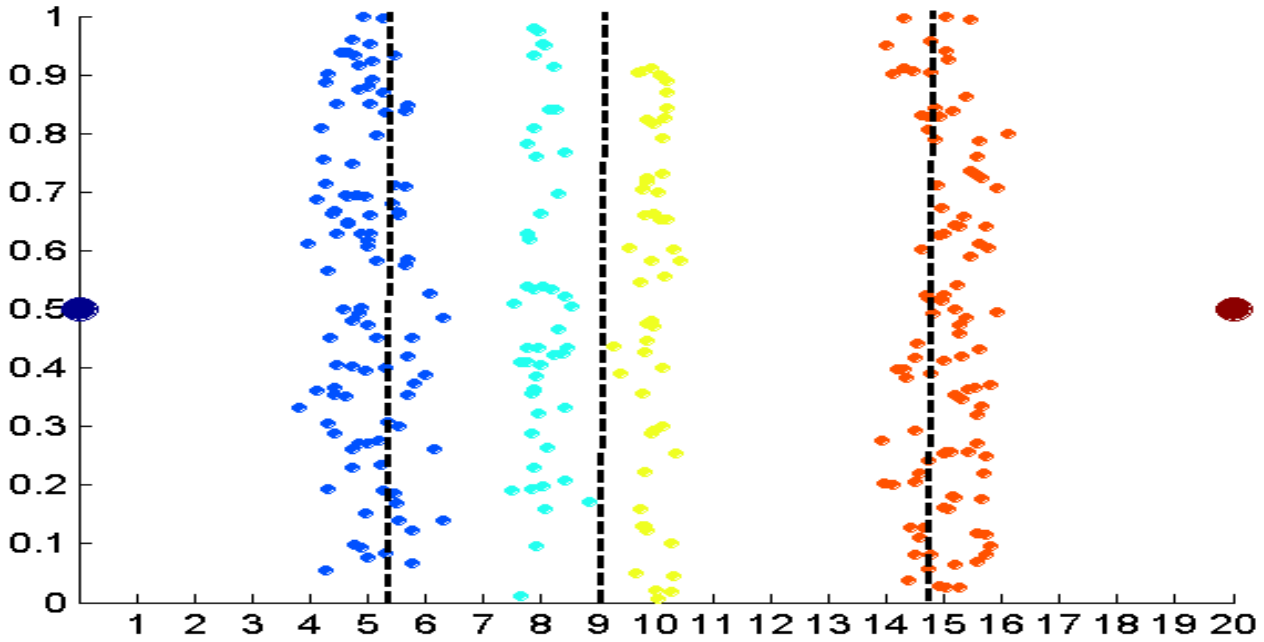
Equal interval width approach used to obtain 4 values



- Equal width (largura igual) divides the range of the attribute into a user-specified number of intervals each having the same width
  - If A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B-A)/N = (20-0)/4 = 5
- Such an approach can be badly affected by outliers

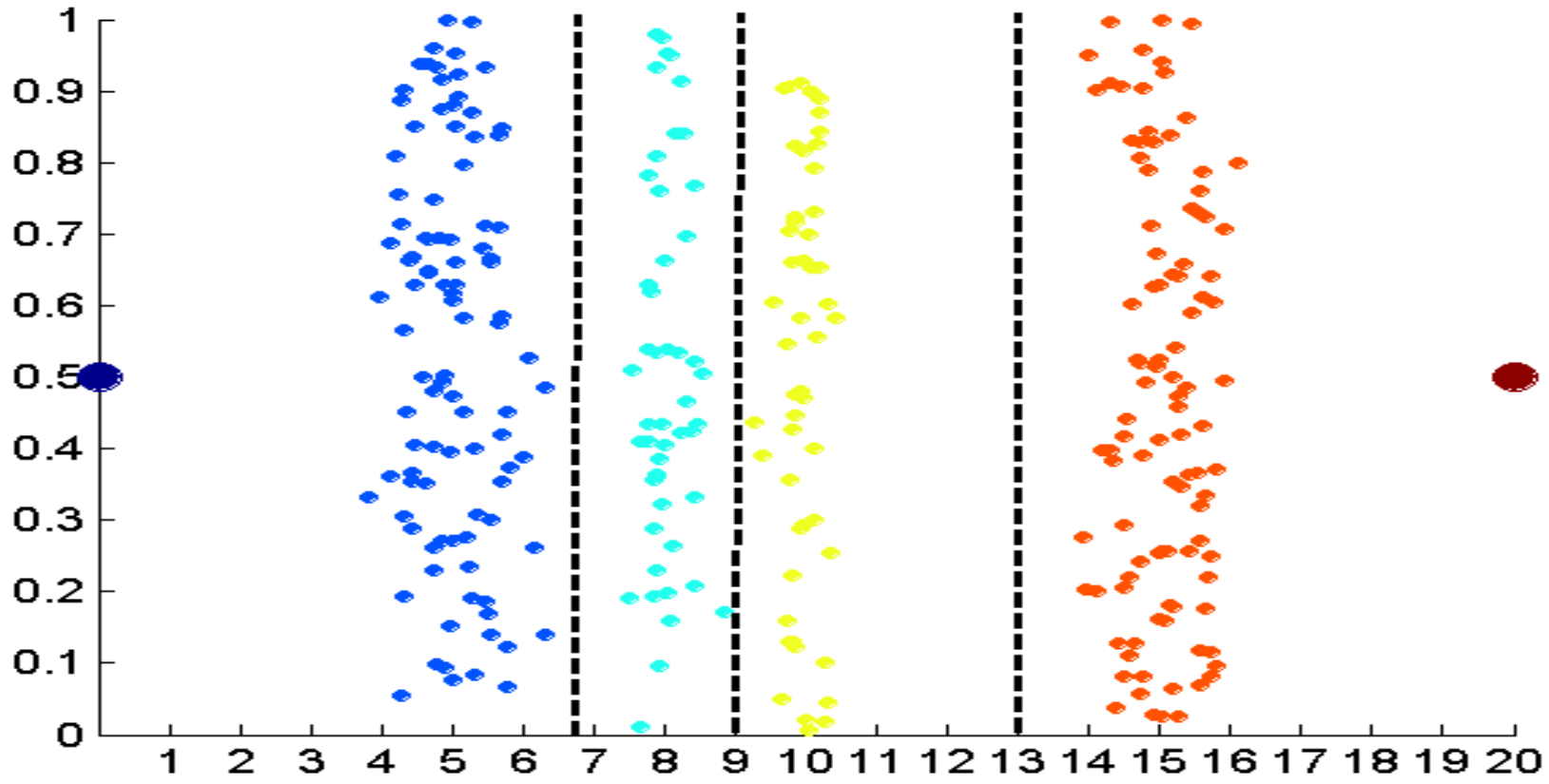
Discretization Without Using Class Labels

Equal frequency approach used to obtain 4 values



- Equal frequency (frequência igual) (equal depth) approach tries to put the same number of objects into each interval – often preferred

Unsupervised Discretization



K-means approach to obtain 4 values

Discretization in Supervised Settings

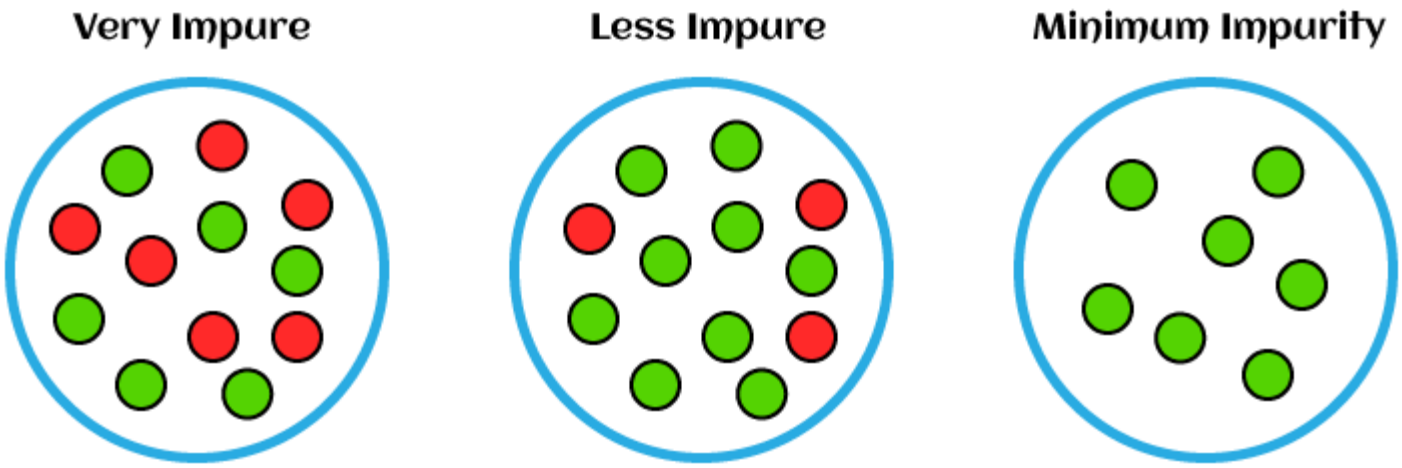
- If classification is our application and class labels are known for some data objects, then discretization approaches that use class labels often produce better classification
- There are many approaches
- Some statistically based approaches:
  - Bottom-up approaches: start with each attribute value in a separate interval and create larger intervals by merging adjacent intervals that are similar according to a statistical test
  - Top-down approaches: start by bisecting the initial values so that the resulting two intervals give minimum entropy

Discretization in Supervised Settings

- Some statistically based approaches:
  - Top-down approaches: start by bisecting the initial values so that the resulting two intervals give minimum entropy
    - This technique only needs to consider each value as a possible split point, because it is assumed that intervals contain ordered sets of values
    - The splitting process is then repeated with another interval, typically choosing the interval with the worst (highest) entropy, until a user-specified number of intervals is reached, or a stopping criterion is satisfied

Discretization in Supervised Settings

- **Entropy** is defined as the randomness or **measuring** the **disorder** of the **information** being processed in Machine Learning
- In other words, we can say that **entropy** is the machine learning metric that **measures** the unpredictability or **impurity** in the system



https://www.javatpoint.com/entropy-in-machine-learning#:~:text=Entropy%20is%20defined%20as%20the,or%20impurity%20in%20the%20system.

Discretization in Supervised Settings

- Entropia

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

O-Ring Failure	
Y	N
7	17

T=24; 7/24=0.29; 17/24=0.71

E (Failure) = E(7, 17) = E(0.29, .71) = -0.29 x $\log_2(0.29)$ - 0.71 x $\log_2(0.71)$ = **0.871**

https://www.saedsayad.com/supervised_binning.htm

Discretization in Supervised Settings

- **Ganho de Informação: redução da entropia dada uma nova informação (o quanto o atributo ajuda organizar os dados)**

$$E(S,A) = \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

		O-Ring Failure	
		Y	N
Temperature	<= 60	3	0
	> 60	4	17

E (Failure,Temperature) = P(<=60) x E(3,0) + P(>60) x E(4,17) = 3/24 x 0 + 21/24 x 0.7= **0.615**

Information Gain = $E(S) - E(S,A)$

Information Gain (Failure, Temperature) = **0.256**

https://www.saedsayad.com/supervised_binning.htm

Discretization in Supervised Settings

O-Ring Failure	Temperature
Y	53
Y	56
Y	57
N	63
N	66
N	67
N	67
N	67
N	68
N	69
N	70
Y	70
Y	70
Y	70
N	72
N	73
N	75
Y	75
N	76
N	76
N	78
N	79
N	80
N	81

Gain = 0.256

		O-Ring Failure	
		Y	**N**
Temperature	**<= 60**	3	0
	> 60	4	17

Gain = 0.101

		O-Ring Failure	
		Y	**N**
Temperature	**<= 70**	6	8
	> 70	1	9

Gain = 0.148

		O-Ring Failure	
		Y	**N**
Temperature	**<= 75**	7	11
	> 75	0	6

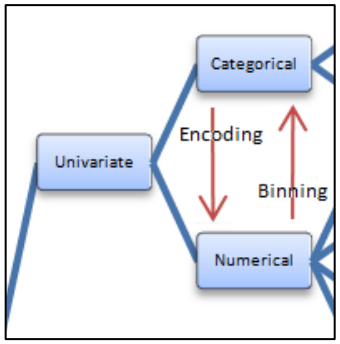
https://www.saedsayad.com/supervised_binning.htm

Binarization/Encoding

- Most machine learning algorithms cannot handle categorical variables unless we convert them to numerical values
- Encoding maps categorical variables to numerical values
- Binarization maps a categorical attribute into one or more binary variables

https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html

https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/



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

Encoding

https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

- **Label Encoding or Ordinal Encoding**
- **One hot Encoding**
- **Dummy Encoding**
- Effect Encoding
- **Binary Encoding [Binarization]**
- BaseN Encoding
- Hash Encoding
- Target Encoding
- ...

Encoding

https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

Label Encoding or Ordinal Encoding

- We use this categorical data encoding technique when the categorical feature is **ordinal**

	Degree
0	High school
1	Masters
2	Diploma
3	Bachelors
4	Bachelors
5	Masters
6	Phd
7	High school
8	High school

	Degree
0	1
1	4
2	2
3	3
4	3
5	4
6	5
7	1
8	1

Encoding

https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

- **One Hot Encoding**
  - We use this categorical data encoding technique when the features are **nominal** (do not have any order)
  - For each level of a categorical feature, we create a new variable

These newly created binary features are known as **Dummy variables**

Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

Encoding

https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

One Hot Encoding

	City
0	Delhi
1	Mumbai
2	Hydrabad
3	Chennai
4	Bangalore
5	Delhi
6	Hydrabad
7	Bangalore
8	Delhi

	City_Delhi	City_Mumbai	City_Hydrabad	City_Chennai	City_Bangalore
0	1.0	0.0	0.0	0.0	0.0
1	0.0	1.0	0.0	0.0	0.0
2	0.0	0.0	1.0	0.0	0.0
3	0.0	0.0	0.0	1.0	0.0
4	0.0	0.0	0.0	0.0	1.0
5	1.0	0.0	0.0	0.0	0.0
6	0.0	0.0	1.0	0.0	0.0
7	0.0	0.0	0.0	0.0	1.0
8	1.0	0.0	0.0	0.0	0.0

Encoding

https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

- **Dummy Encoding**
  - Scheme similar to one-hot encoding, i.e., transforms the categorical variable into a set of binary variables (also known as dummy variables)
  - In the case of one-hot encoding, for N categories in a variable, it uses N binary variables
  - The dummy encoding is a small improvement over one-hot-encoding: it uses N-1 features to represent N labels/categories

Encoding

https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

- Dummy Encoding

Column	Code
A	100
B	010
C	001

One- Hot Coding

Column	Code
A	10
B	01
C	00

Dummy Code

Encoding

https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

- Dummy Encoding

	City
0	Delhi
1	Mumbai
2	Hyderabad
3	Chennai
4	Bangalore
5	Delhi
6	Hyderabad

	City_Chennai	City_Delhi	City_Hyderabad	City_Mumbai
0	0	1	0	0
1	0	0	0	1
2	0	0	1	0
3	1	0	0	0
4	0	0	0	0
5	0	1	0	0
6	0	0	1	0

Encoding

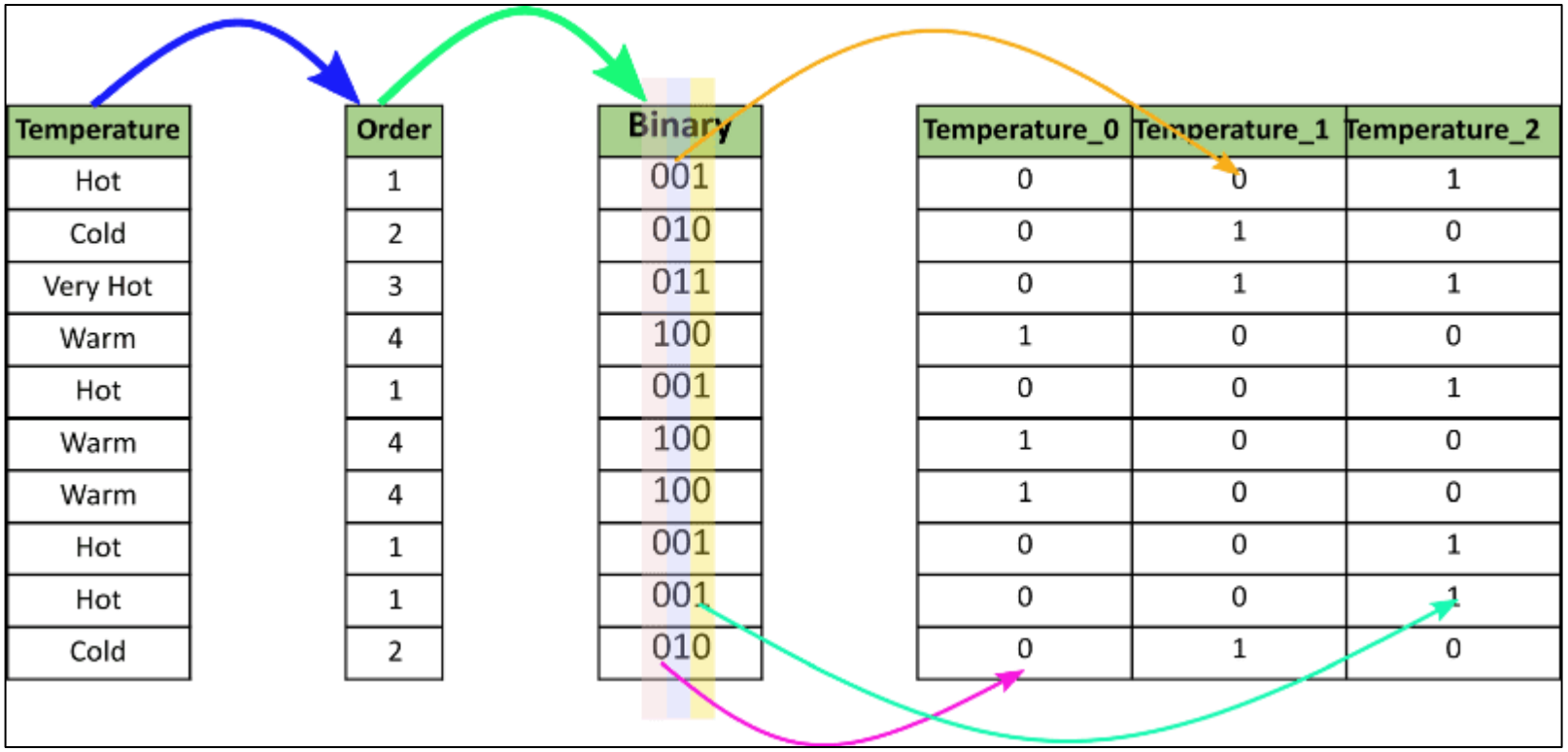
https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

- **Drawbacks of One-Hot and Dummy Encoding**
  - If there are multiple categories in a feature variable in such a case we need a similar number of dummy variables to encode the data. For example, a column with 30 different values will require 30 new variables for coding
  - These two encoding schemes introduce sparsity in the dataset i.e several columns having 0s and a few of them having 1s

Encoding

https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html

- Binary Encoding [Binarization]



Temperature	Order	Binary
Hot	1	001
Cold	2	010
Very Hot	3	011
Warm	4	100
Hot	1	001
Warm	4	100
Warm	4	100
Hot	1	001
Hot	1	001
Cold	2	010

Temperature_0	Temperature_1	Temperature_2
0	0	1
0	1	0
0	1	1
1	0	0
0	0	1
1	0	0
1	0	0
0	0	1
0	0	1
0	1	0

https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02

Encoding

https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

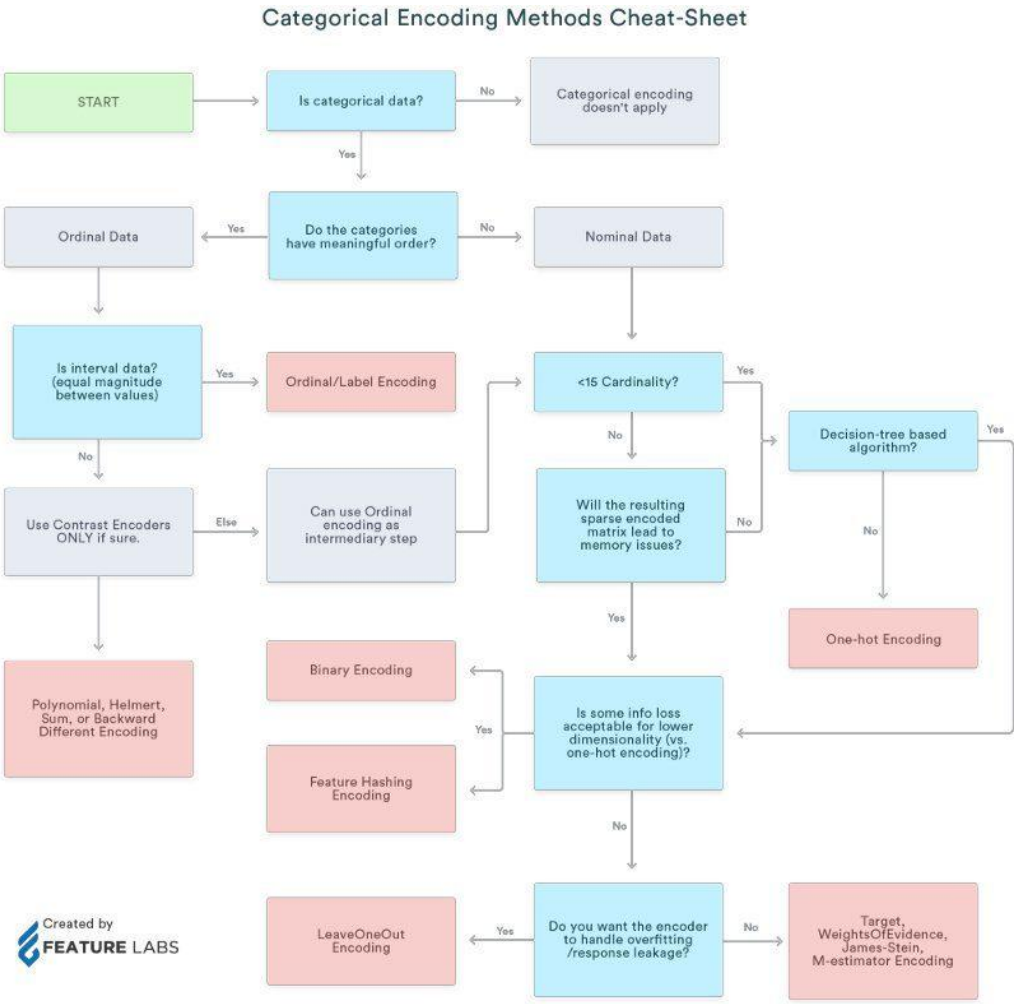
- **Binary Encoding [Binarization]**
  - Binary encoding works really well when there are a high number of categories
  - Binary encoding is a memory-efficient encoding scheme as it uses fewer features than one-hot encoding
  - Further, It reduces the curse of dimensionality for data with high cardinality

	City
0	Delhi
1	Mumbai
2	Hyderabad
3	Chennai
4	Bangalore
5	Delhi
6	Hyderabad
7	Mumbai
8	Agra

	City_0	City_1	City_2	City_3
0	0	0	0	1
1	0	0	1	0
2	0	0	1	1
3	0	1	0	0
4	0	1	0	1
5	0	0	0	1
6	0	0	1	1
7	0	0	1	0
8	0	1	1	0

Encoding

https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html



https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02

Attribute Transformation

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - In other words, for each object, the transformation is applied to the value of the variable for that object
    - **Simple functions**
    - **Normalization or Standardization [Feature Scaling]**

Attribute Transformation

https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-transforming-variables-acea03472e24

- **Simple functions**: x^k, log(x), e^x, |x|
  - In statistics, especially sqrt, log, and 1/*x*, are often used to transform data that does not have a Gaussian (normal) distribution into data that does
  - Variable transformations should be applied with caution because they change the nature of the data
    - For instance, the transformation 1/*x* reduces the magnitude of values that are 1 or larger, but increases the magnitude of values between 0 and 1
    - Thus, for all sets of values, the transformation 1/*x* reverses the order

$\{1, 2, 3\}$ go to $\{1, \frac{1}{2}, \frac{1}{3}\}$, but the values $\{1, \frac{1}{2}, \frac{1}{3}\}$ go to $\{1, 2, 3\}$

Attribute Transformation

https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-feature-scaling-3e79738f7d35

- **Normalization or Standardization [Feature Scaling]**
  - Refers to the methods used to normalize the range of values of independent variables
  - In other words, the ways to set the feature value range within a similar scale
    - Renda, Idade (renda domina a diferença entre os objetos) (colocar na mesma escala))

Attribute Transformation

https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-feature-scaling-3e79738f7d35

- **Feature Scaling**
  - Feature magnitude matters for several reasons:
    - The scale of the variable directly influences the regression coefficient
    - Variables with a more significant magnitude dominate over the ones with a smaller magnitude range
    - Gradient descent converges faster when features are on similar scales
    - Feature scaling helps decrease the time to find support vectors for SVMs
    - Euclidean distances are sensitive to feature magnitude

Attribute Transformation

https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-feature-scaling-3e79738f7d35

- **Feature Scaling**
  - **Algorithms sensitive to feature magnitude**
    - Linear and logistic regression
    - Neural networks
    - Support vector machines
    - KNN
    - K-means clustering
    - Principal component analysis (PCA)
  - **Algorithms not sensitive to feature magnitude**
    - Classification and regression trees
    - Random forests
    - Gradient boosted trees

Attribute Transformation

https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-feature-scaling-3e79738f7d35

- **Feature Scaling Methods**
  - Mean normalization
  - **Standardization (Z-score Normalization)**
  - Robust scaling (scaling to median and IQR)
  - **Scale to maximum and minimum [Normalização]**
  - Scale to the absolute maximum
  - Scale to unit norm
  - …

https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

https://www.analyticsvidhya.com/blog/2020/07/types-of-feature-transformation-and-scaling/

Attribute Transformation

https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-feature-scaling-3e79738f7d35

- **Standardization (Z-score)**
  - Rescales the distribution of values so that the mean of observed values is 0 and the standard deviation is 1
    - Subtracting the mean from the data is called **centering**, whereas dividing by the standard deviation is called **scaling**. As such, the method is sometime called "**center scaling**". [https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/]

Standardization assumes that your observations fit a Gaussian distribution (bell curve) with a well-behaved mean and standard deviation. You can still standardize your data if this expectation is not met, but you may not get reliable results [https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/]

A **z-score** = V means that your data point is V standard deviations above/below the mean

$$\bar{x} = \frac{X - mean(X)}{std(X)}$$

Attribute Transformation

https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-feature-scaling-3e79738f7d35

- **Standardization (Z-score)**
  - It scales the variance at 1
  - It centers the mean at 0
  - It preserves the shape of the original distribution
  - It preserves outliers if they exist
  - Minimum and maximum values vary

A **z-score** = V means that your data point is V standard deviations above/below the mean.

$$\bar{x} = \frac{X - mean(X)}{std(X)}$$

Attribute Transformation

https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-feature-scaling-3e79738f7d35

- Standardization (Z-score)

Age	Income
44	72000
27	48000
30	54000
38	61000
40	50000
35	58000
27	52000
48	79000
50	83000
37	67000

Age	Income
0.827	0.818
-1.370	-1.228
-0.982	-0.716
0.051	-0.119
0.310	-1.057
-0.336	-0.375
-1.370	-0.887
1.344	1.415
1.602	1.757
-0.077	0.392

Standardization

Attribute Transformation

https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-feature-scaling-3e79738f7d35

- **Min-Max Scaling**
  - It subtracts the minimum value from all the variable observations and then divides it by the variable's value range - compress the values between 0 and 1
    - It may not maintain the shape of the original distribution
    - The minimum and maximum values are in the range of [0,1]
    - This method is very sensitive to outliers

$$\bar{x} = \frac{X - min(X)}{max(X) - min(X)}$$

Attribute Transformation

https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-feature-scaling-3e79738f7d35

- Min-Max Scaling

Age	Income
44	72000
27	48000
30	54000
38	61000
40	50000
35	58000
27	52000
48	79000
50	83000
37	67000

Age	Income
0.739	0.685
0	0
0.130	0.171
0.478	0.371
0.565	0.057
0.347	0.285
0	0.114
0.913	0.885
1	1
0.434	0.542

Min-Max Scaling

Observações Finais

- Vide testes sobre distribuição normal