Machine Learning for Networking
ML4N

Luca Vassio
Gabriele Ciravegna
Zhihao Wang
Tailai Song

Recap – key concepts

- Big Data
- Data science
- Artificial intelligence
- Machine Learning

What ML is all About ?

Fit models to data to make predictions or forecasts !

Notation

$a \in \mathcal{A}$	This statement indicates that the object a is an element of the set $\mathcal{A}$.
$\lvert\mathcal{A}\rvert$	The cardinality (number of elements) of a finite set $\mathcal{A}$.
$\mathcal{A} \subseteq \mathcal{B}$	$\mathcal{A}$ is a subset of $\mathcal{B}$.
$\mathcal{A} \subset \mathcal{B}$	$\mathcal{A}$ is a strict subset of $\mathcal{B}$.
$\mathbb{N}$	The set of natural numbers $1, 2, \ldots$.
$\mathbb{R}$	The set of real numbers x [120].
$\{0, 1\}$	The set consisting of two real-number 0 and 1.
$[0, 1]$	The closed interval of real numbers x with $0 \leq x \leq 1$.

Notation

$\mathbf{x} = (x_1, \ldots, x_n)^T$	A vector of length n. The jth entry of the vector is denoted x_j.
$\|\mathbf{x}\|$	Some norm of the vector $\mathbf{x} \in \mathbb{R}^n$ [46]. Unless specified otherwise, we mean the Euclidean norm $\|\mathbf{x}\|_2$.

Notation

Data exploration and visualization

Dataset

- A dataset is a collection of data
  - e.g., a tabular representation of data includes rows and columns
    - Rows correspond to objects, records, points, cases, samples, entities, or instances (notation: m rows)
    - columns are the attributes (notation: n columns)

Dataset

- The size of the dataset has an impact on the choice of the analyses
  - Some algorithms require considerable hardware resources when applied to large datasets, in some cases it is not possible to execute them at all
  - There are solutions to reduce the size of the dataset preserving the completeness of the data
    - data sampling can reduce the dataset size in terms of number of rows (m’<m)
    - feature selection can reduce the number of attributes (n’<n)

Feature/attribute

- Each column of the dataset represents one attribute/feature
  - Data exploration can be performed in a univariate or multivariate fashions
- For further analysis consider the following basic information for each attribute
  - Unit of measurement
  - Attribute Type
    - Categorical (not numerical or fixed number of possible values)
    - Numerical (discrete or continuous)
  - Attribute Domain
    - It is a good practice to verify if the attribute values satisfy the domain-driven constraints

Characterizing Distributions

Theoretical distribution. Random variable X

- Domain bounds (minimum possible value, maximum possible value)
- Mean value = expected value
- Standard deviation
- Mathematical functions
  - A probability density distribution
  - A cumulative density distribution
  - Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean
  - Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable
  - ...

Univariate analysis: Distribution

- Attribute values/sample $(x_1, \ldots, x_m)$

  E.g., [0, 1, 4, 0 ,3 , 2.4, 10, 0, -1]

- This numbers represent an **empirical distribution**
  - samples from a (not accessible) theoretical distribution of a random variable X

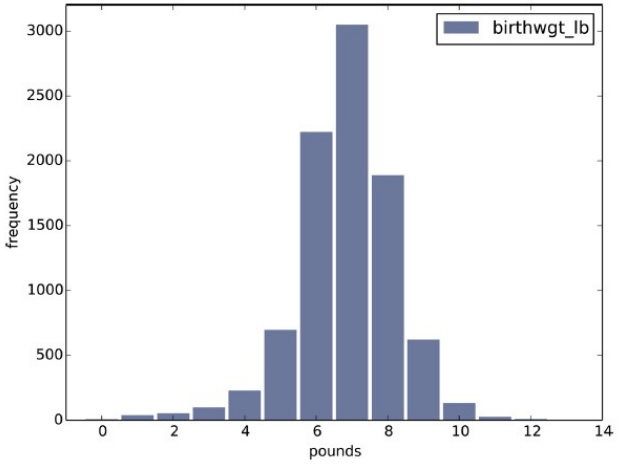
Characterizing Distributions

Distribution of samples $(x_1, \ldots, x_m)$

- Minimum value
- Maximum value
- Sample mean
- Sample standard deviation
- Frequency plot (histogram)
- Mathematical functions
  - Empirical probability density distribution
  - Empirical cumulative density distribution
  - Sample Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean
  - Sample Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable
  - ...

Univariate analysis: Distribution

- Distribution of the attribute
  - A possible representation of a distribution is a histogram
  - The range of values are divided into bins (series of intervals)
  - Count how many values in the dataset fall into each interval (frequency)

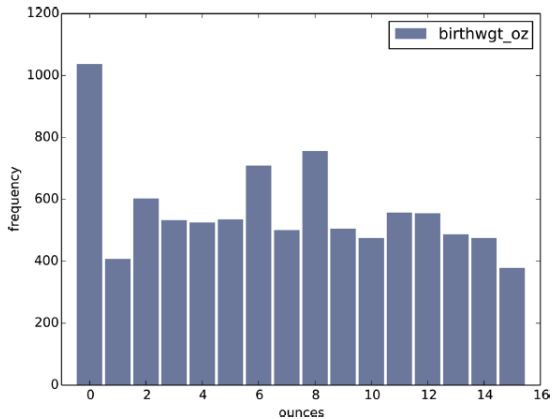


Example of a histogram that shows the distribution of pound part of birth weight

The distribution is approximately bell-shaped, which is the shape of the normal distribution

Univariate analysis: Distribution

- Distribution of the attribute
  - A possible representation of a distribution is a histogram
  - The range of values are divided into bins (series of intervals)
  - Count how many values in the dataset fall into each interval (frequency)



Example of a histogram that shows the distribution of the ounces part of birth weight

This distribution is not uniform
0 is more common than the other values,
1 and 15 are less common, probably because respondents round off birth weights that are close to an integer value.

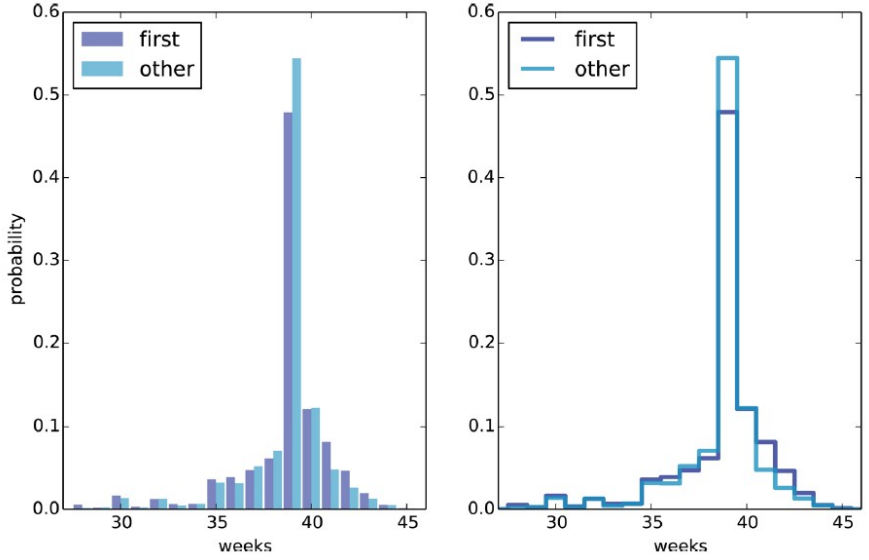
Univariate analysis: Distribution

Probability distributions are usually classified as

- Discrete probability distribution
  - Characterized by a discrete list of the probabilities of the outcomes
  - It is typically described by a probability mass function
- Continuous probability distribution
  - The set of possible outcomes can assume values in a continuous range
  - It is typically described by a probability density functions

Discrete probability

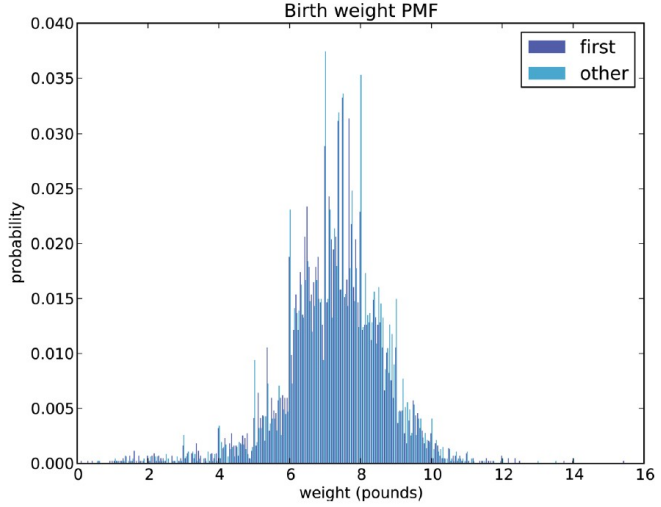
- Example of the (empirical) probability mass function (PMF) of a discrete probability distribution



PMF of pregnancy lengths for first babies and others, using bar graphs (left) and step functions (right)

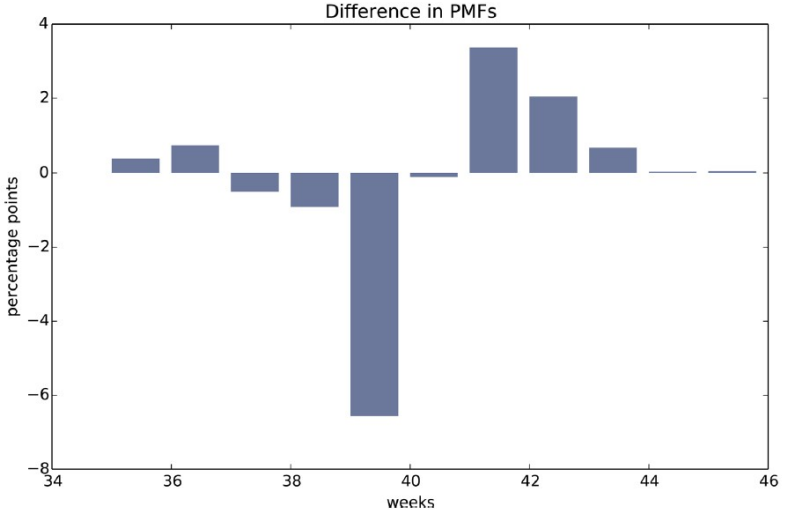
PMF limitations

- In cases of a lot of samples to show it is very hard to read information from a PMF plot
- Possible solutions include
  - Showing the difference of distributions
  - Calculating the cumulative distribution function (CDF)



Difference of distributions

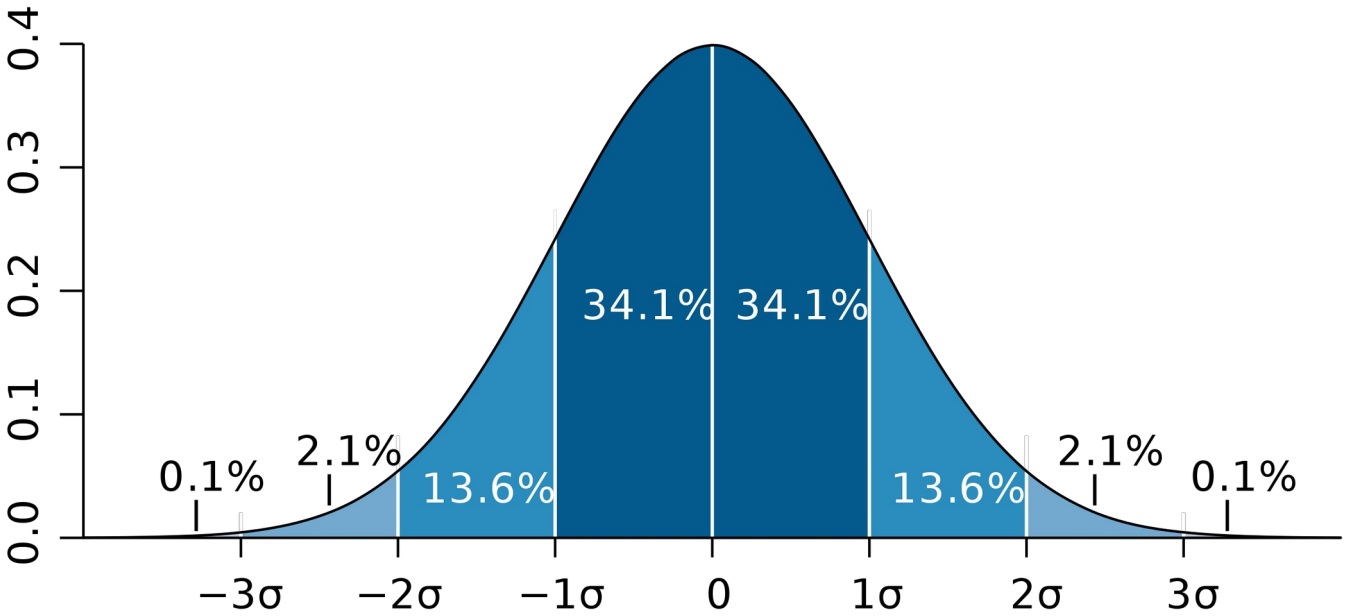
- Example of difference between two PMFs (probability mass functions)



Difference, in percentage points, by week

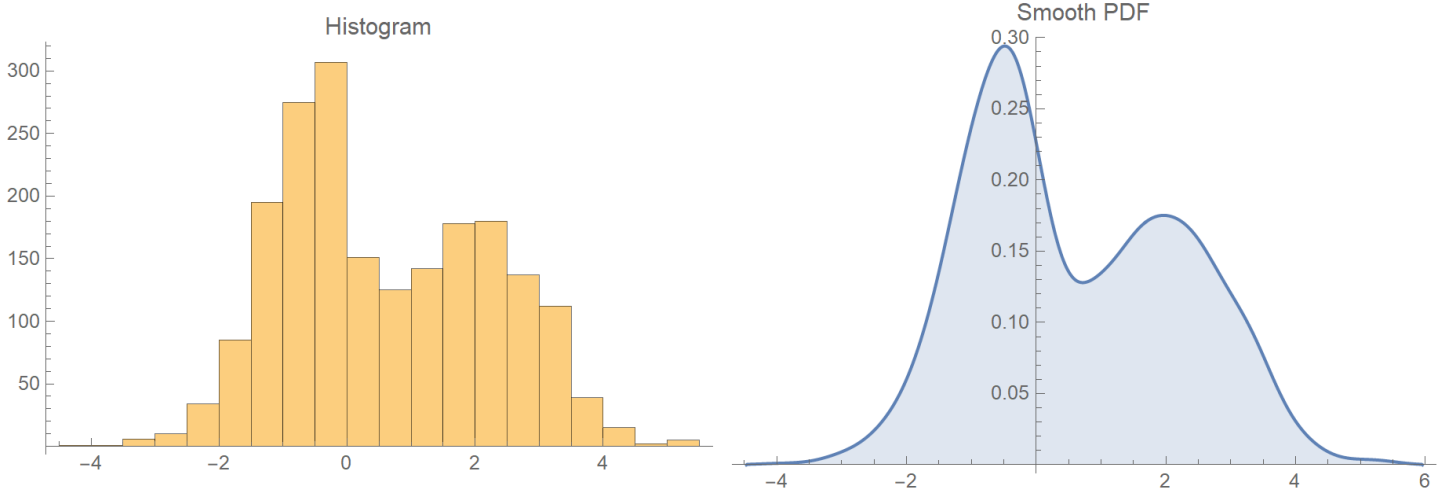
Continuous probability

- Example of the probability density function (PDF) of the normal distribution



Kernel density estimation

- Estimate continuous probability density function based on samples



Kernel density estimation

- Estimate f as the **weighted average of neighboring observed data**
- The weight is defined by the **kernel** function, such that closer points are given higher weights
- The estimated function is **continuous** and '**smooth'**
- The level of smoothness is set by parameters

Kernel Density Estimate (KDE)

- Estimating the probability density function of a spatial random variable
- From its observations → from its m samples x_i

$$\widehat{f}_h(x) = \frac{1}{m}\sum_{i=1}^{m} K_h(x - x_i)$$

Kernel Density Estimate (KDE)

- Estimating the probability density function of a spatial random variable
- From its observations → from its m samples x_i

$$\widehat{f}_h(x) = \frac{1}{m}\sum_{i=1}^{m} K_h(x - x_i) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right)$$

Kernel Density Estimate (KDE)

$$\widehat{f}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right)$$

Kernel K

?

Gaussian Kernel

- Gaussian function: $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- Standard normal distribution and kernel K(z):

$$\frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

Kernel Density Estimate (KDE)

$$\widehat{f}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right)$$

Kernel K

Gaussian

Bandwidth h

?

Kernel Density Estimate (KDE)

$$\widehat{f}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right)$$

Kernel K

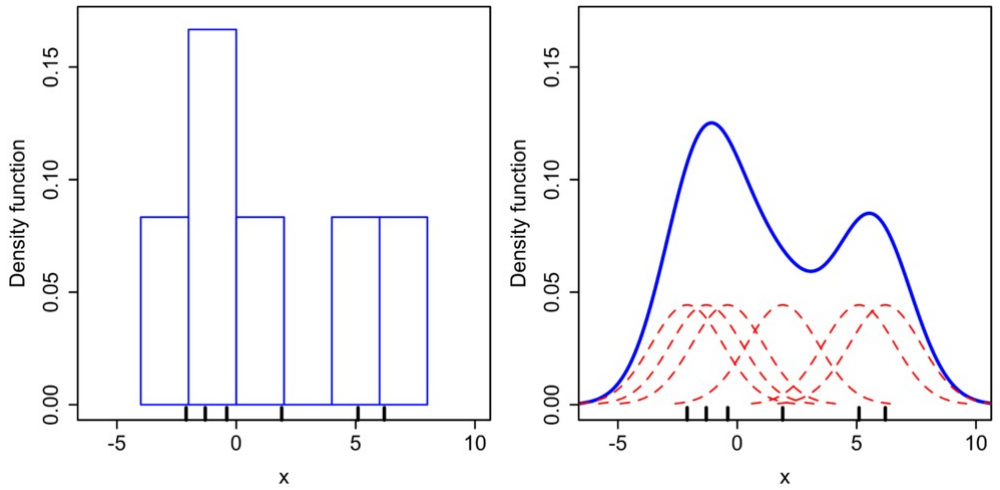
Gaussian

Bandwidth h

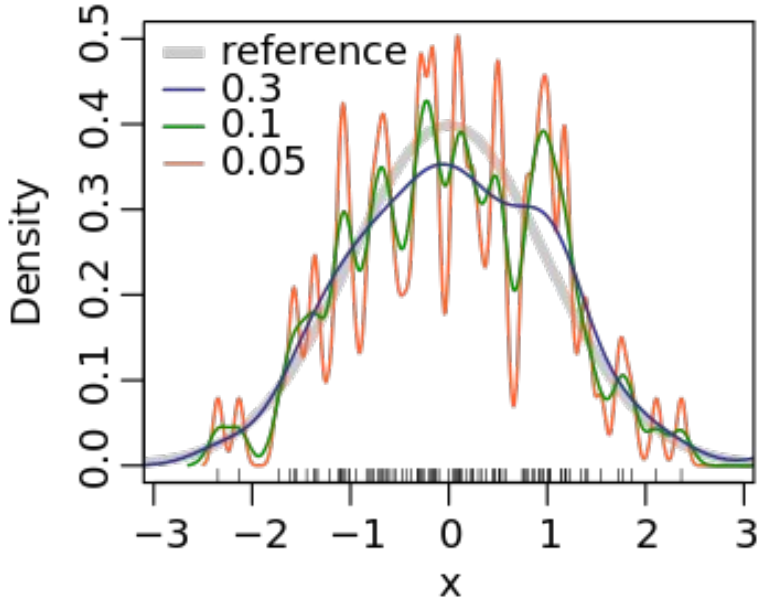
rule of thumb exists
depending on distribution

Kernel Density Estimate (KDE)

$$\widehat{f}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right)$$



Effect of bandwidth



- Misspecification of the bandwidth can produce a distorted representation of the data.
- Like the choice of bin width in a histogram, an over-smoothed curve can erase true features of a distribution
- An under-smoothed curve can create false features out of random variability

Kernel Density Estimate (KDE)

- Relative to a histogram, KDE can produce a plot that is less cluttered and more interpretable
- Potential to introduce distortions if the underlying distribution is bounded or not smooth.
- Like a histogram, the quality of the representation also depends on the selection of good smoothing parameters.

Cumulative Distribution Function (CDF)

- The CDF is a function of *x*, where *x* is any value that might appear in the distribution

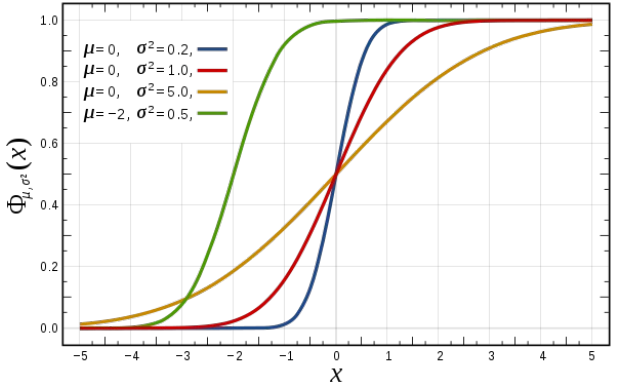
$$F_X(x) := P(X \leq x)$$

$$P(a \leq X \leq b) = F_X(b) - F_X(a)$$

Cumulative Distribution Function (CDF)

$$F_X(x) := P(X \leq x)$$

- The CDF also provides a visual representation of the shape of the distribution
  - probable values/intervals appear as steep or vertical sections of the CDF



Cumulative distribution function for the normal distribution

Cumulative Distribution Function (CDF)

Fundamental Theorem of Calculus:

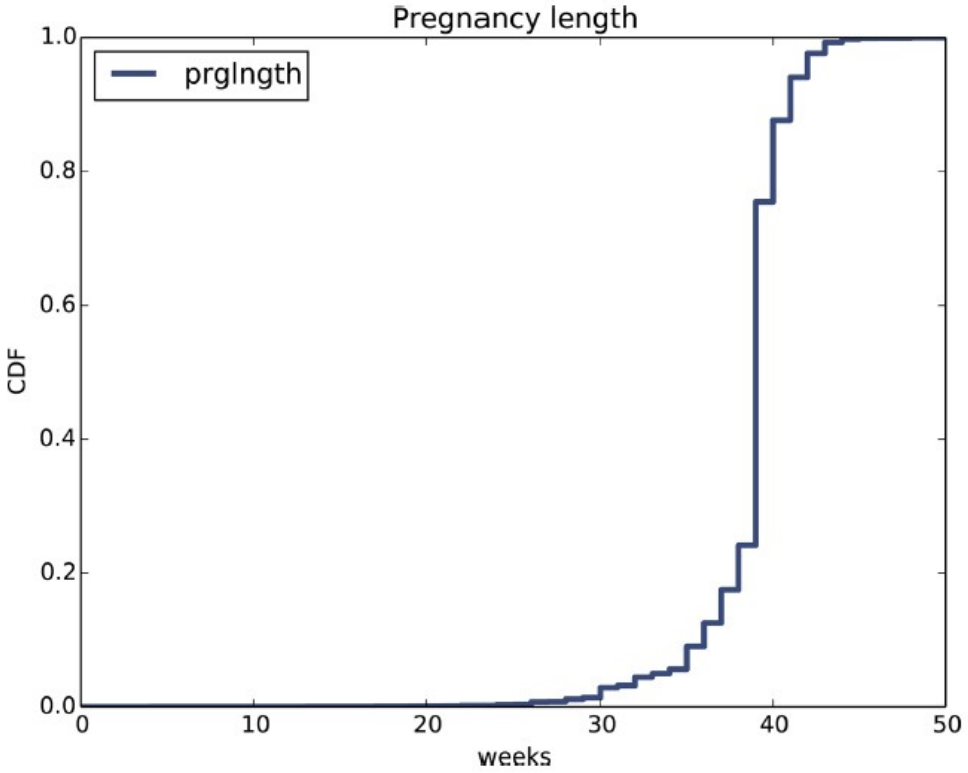
$$f_X(x) = \frac{dF_X(x)}{dx}$$

Empirical Cumulative Distribution Function (ECDF)

- The ECDF is an estimate of the CDF that generated the points in the sample
- To evaluate ECDF(x) for a particular value of x, we compute the fraction of values in the distribution less than or equal to x
- The ECDF is a step function that jumps up by 1/m at each of the m data points

Empirical Cumulative Distribution Function

- Example of (E)CDF



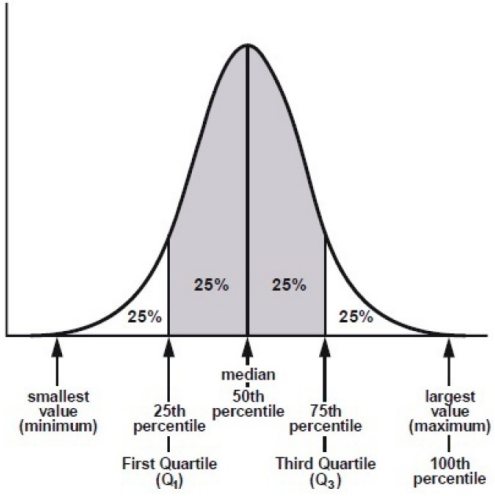
E.g. about 10% of pregnancies are shorter than 36 weeks, and about 90% are shorter than 41 weeks.

The CDF also provides a visual representation of the shape of the distribution. Common values appear as steep or vertical sections of the CDF; in this example, the mode at 39 weeks is apparent.

There are few values below 30 weeks, so the CDF in this range is flat.

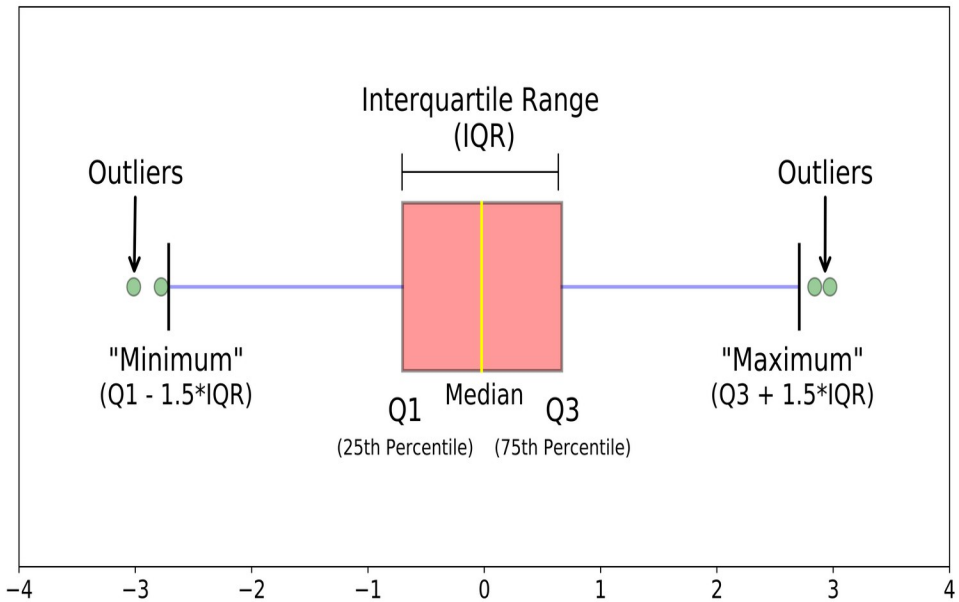
Percentiles

- Percentile indicates the value below which a given percentage of observations in a group of observations falls
- Representing a feature/attribute/variable through percentiles allow representing the entire distribution
  - 25^{th}, 50^{th} , 75^{th} percentile: quartiles
- Outliers
  - e.g., values in the first and last percentile of the distribution



Boxplot

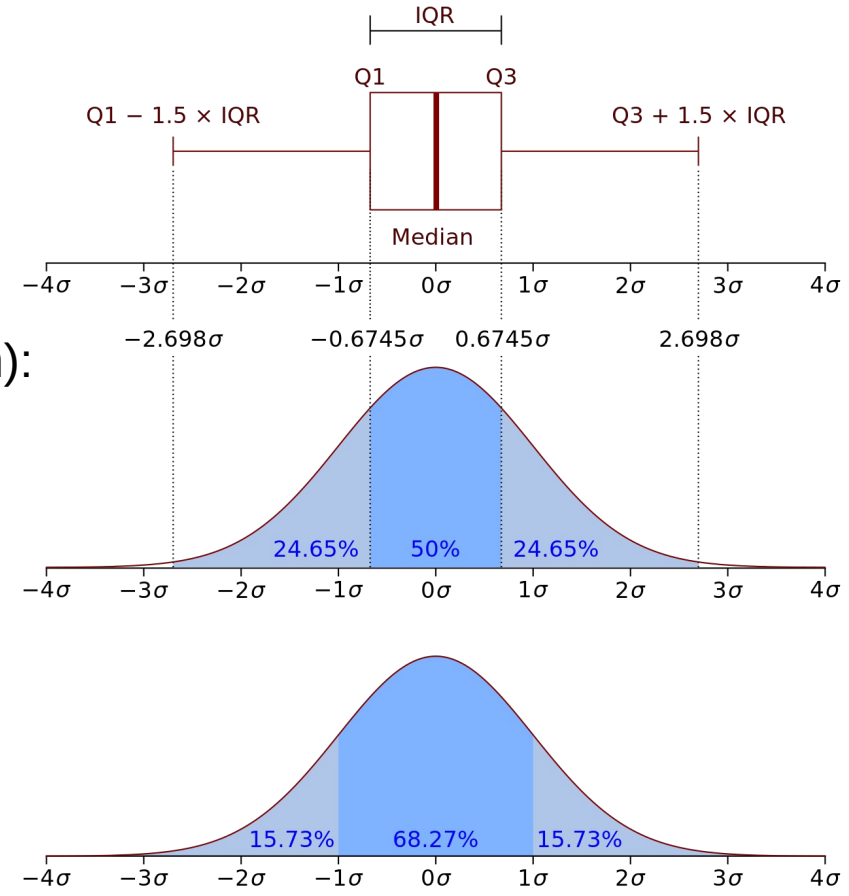
- Boxplots are a standardized way of displaying/summarizing the distribution of data based on a summary values



- **median** (Q2/50th Percentile): the middle value of the attribute.
- **first quartile** (Q1/25th Percentile): the middle number between the smallest number (not the “minimum”) and the median of the attribute.
- **third quartile** (Q3/75th Percentile): the middle value between the median and the highest value (not the “maximum”) of the attribute.
- **interquartile range** (IQR): 25th to the 75th percentile.
- **whiskers** (shown in blue)
- **outliers** (shown as green circles)

Boxplot

If data is normal (Gaussian):



Outliers

- Outliers are extreme values that might be
  - errors in measurement and recording
  - accurate reports of rare events
- The best way to define/handle outliers depends on “domain knowledge”
  - Information about where the data come from and what they mean
  - it depends on what analysis you are planning to address

Outliers

Outliers can be detected through:

- Univariate analysis
  - Boxplot
  - Percentiles
  - Histograms
  - GESD
  - …
- Multivariate analysis
  - DBSCAN
  - …
- More specific techniques

Outlier Detection GESD

Generalized Extreme Studentized Deviate (GESD)

a.k.a. generalized Grubbs test

- It is used to detect one or more outliers in a univariate data set that follows a normal distribution
- The GESD test performs separate tests for 1 outlier up to an upper bound k of outliers
- Test statistic:

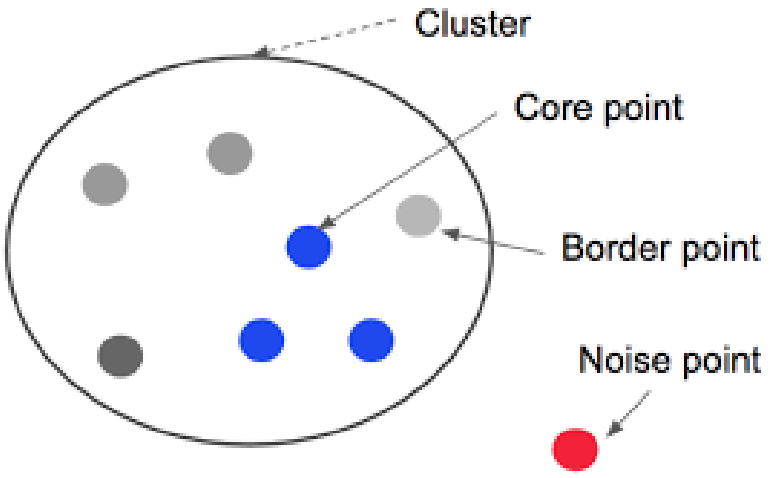
$$R = \frac{\max_{i=1,\ldots,m} |x_i - \bar{x}|}{\sigma}$$

- With $\bar{x}$ and σ denoting the sample mean and sample standard deviation, respectively
- Test statistic R is recomputed by recursively removing value R from the dataset and recomputing the new R up to the k^{th} R value

https://towardsdatascience.com/anomaly-detection-with-generalized-extreme-studentized-deviate-in-python-f350075900e2

Outlier Detection DBSCAN

- DBSCAN is a density-based clustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in **low-density regions** (whose nearest neighbors are too far away)



Correlation analysis

Characterizing multivariate dataset

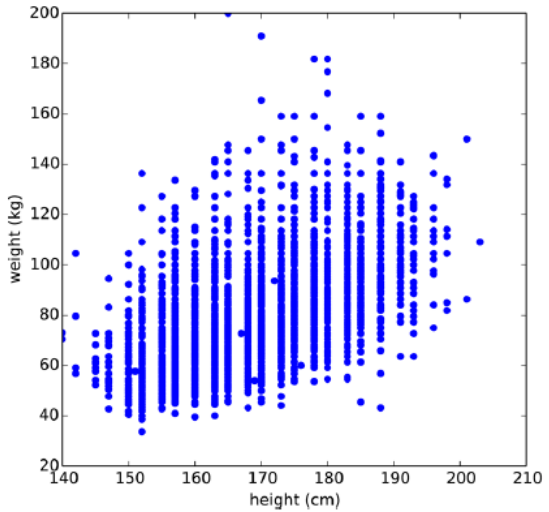
- A dataset usually includes different features/attributes
  - The description of the main relations between attributes assumes a key role
- Statistical descriptions includes
  - Scatter plot
  - Scatter plot percentiles
  - Correlation analysis
  - …

Correlation

- Measure of the relationship between two data objects
- Useful during the data exploration phase
  - To be better aware of data properties
- Analysis of feature correlation
  - Correlated features could be removed simplifying the next analytics steps
  - improving the performance of the data-driven algorithms

Scatter plot

- The simplest way to visually check for a relationship between two variables is a scatter plot
- e.g. plot the height and weight

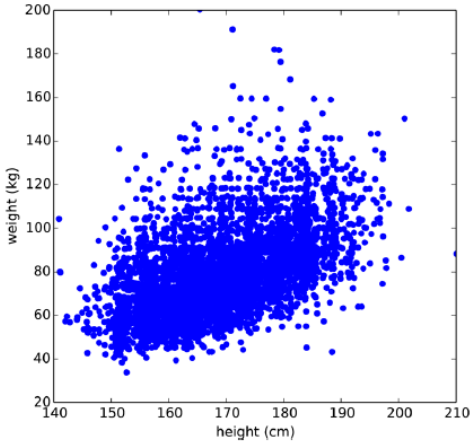


People who are taller tend to be heavier

This example data is affected by rounding and conversion (inch to cm)

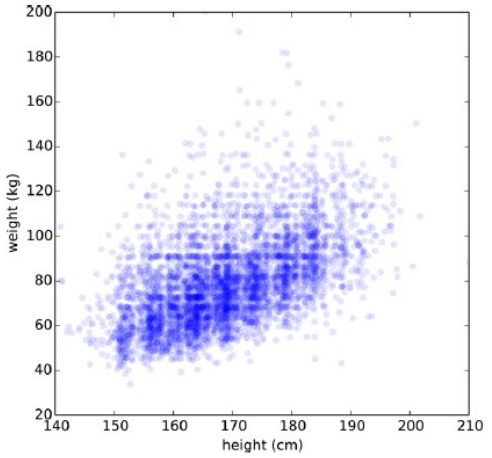
Scatter plot

A possible solution is to **jitter** the data, which means adding random noise to reverse the effect of rounding off



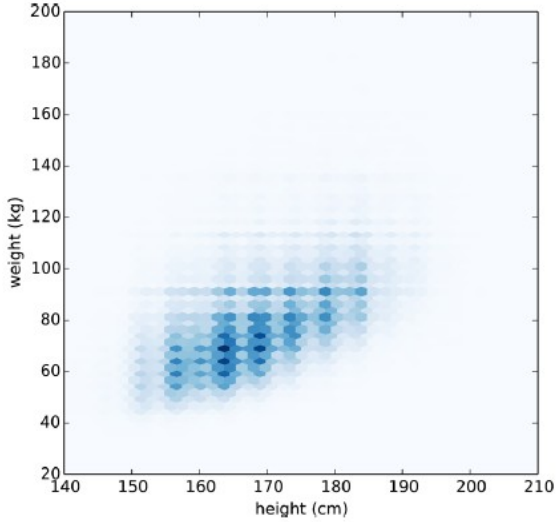
Then adding **alpha** parameter (transparency) to each point in order to retrieve density information

Darker zones correspond to higher density zones



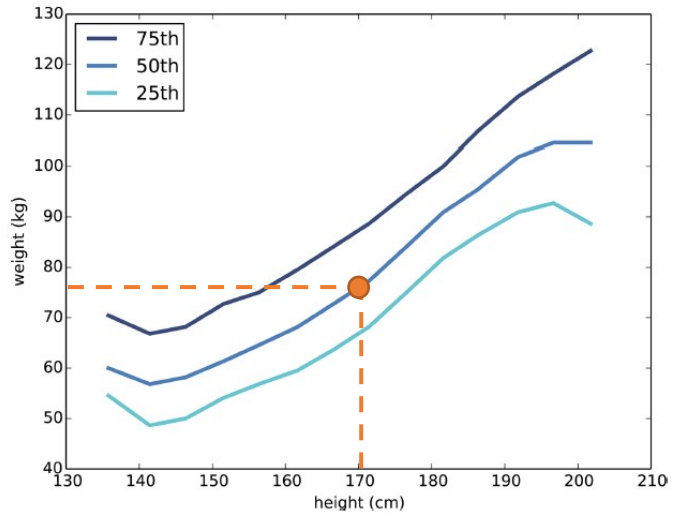
Scatter plot

- Another improvement is to convert scatter plot into hexbin plot
  - The hexbin plot uses hexagonal bins that are colored according to how many data points fall in it
  - This is ~ 2D histogram (density representation)
  - The main issue of the scatter plot is the limitation of representing huge quantity of points



Scatter Plot Percentiles

- This technique requires to bin one variable and plot a curve for the two variables in each bin
  - Each curve is derived for the population of the bin
  - It is an alternative to scatter plot



Line plot where the 25th, 50th and 75th percentiles are shown for weight

e.g. orange intersection means that 50% of people 170 cm tall weigh less than 75kg

Correlation

- A correlation is a statistic intended to quantify the strength of the relationship between two variables.
- Possible way to compute correlations are:
  - Covariance
    - In order to compare two variables, they must have the same unit of measurement
    - alternatively, they must be normalised
  - Pearson
    - Solves the problem of normalization
    - Only detects linear correlation
  - Spearman

Covariance and Pearson

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y},$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Pearson

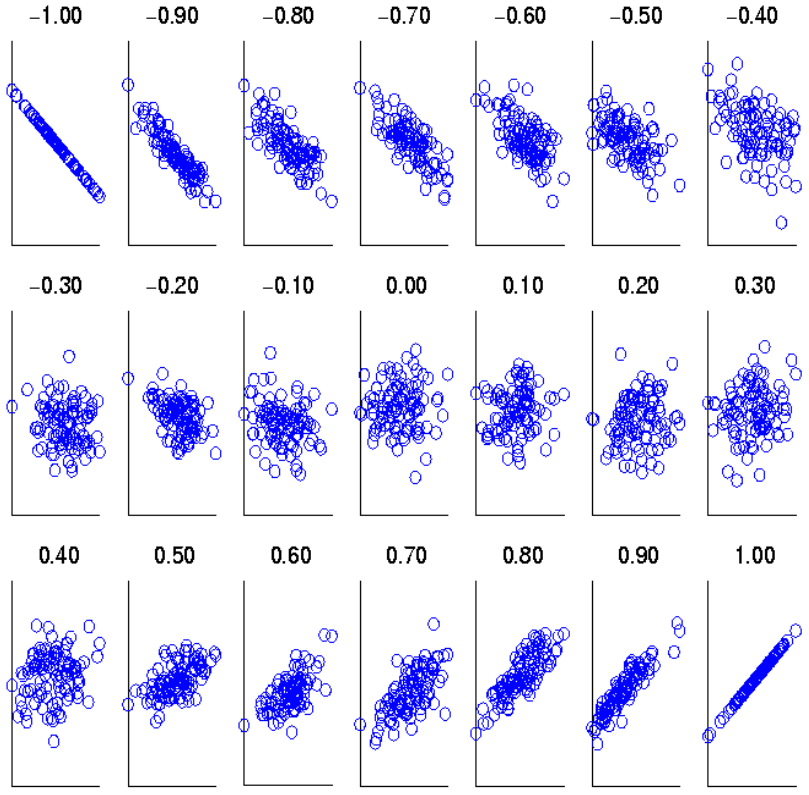


Figure 5.11. Scatter plots illustrating correlations from -1 to 1.

Scatter plots showing the similarity from –1 to 1.

Nonlinear correlation

- **x** = (-3, -2, -1, 0, 1, 2, 3)
- **y** = (9, 4, 1, 0, 1, 4, 9)

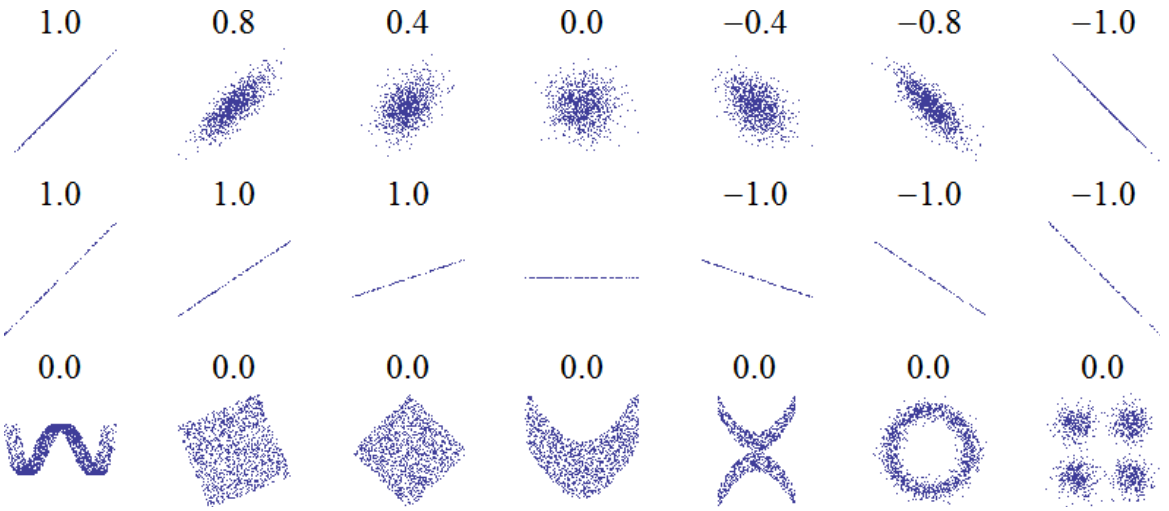
$y_i = x_i^2$

- mean(**x**) = 0, mean(**y**) = 4
- std(**x**) = 2.16, std(**y**) = 3.74

corr = [(-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+(3)(5)] / (6 * 2.16 * 3.74) = 0

Nonlinear correlation

There are some types of nonlinear correlations that Pearson's correlation cannot detect



Spearman’s Rank Correlation

- The Spearman index or Spearman's rank uses the variables rank instead of Pearson that uses the variables themselves
- It assesses how well the relationship between two variables can be described using a monotonic function
- In some cases the Spearman index allows to find a correlation when the Pearson index returns a value close to 0

Spearman's Rank Correlation

For a sample of size m, the m raw scores x_i, y_i are converted to ranks $\mathrm{R}(x_i), \mathrm{R}(y_i)$

$$r_s = \rho_{\mathrm{R}(x),\mathrm{R}(y)} = \frac{\mathrm{cov}(\mathrm{R}(x), \mathrm{R}(y))}{\sigma_{\mathrm{R}(x)} \sigma_{\mathrm{R}(y)}}$$

if all m ranks are distinct integers, it can be computed using the simplified formula

$$r_s = 1 - \frac{6 \sum (\mathrm{R}(X_i) - \mathrm{R}(Y_i))^2}{m(m^2 - 1)}$$

Spearman’s Rank Correlation

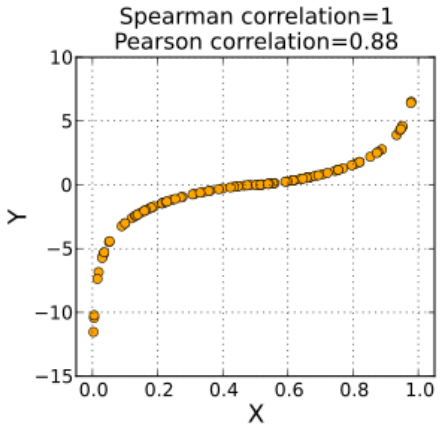
Method - calculating the coefficient

- Rank the two data sets. Ranking is achieved by giving the ranking '1' to the biggest number in a column, '2' to the second biggest value and so on. The smallest value in the column will get the lowest ranking.
- Tied scores are given the mean (average) rank. For example, if there are three tied scores ranked fifth (fifth, sixth and seventh), the mean rank in this case is calculated as (5+6+7) ÷ 3 = 6.
- Find the difference in the ranks: This is the difference between the ranks of the two values on each row.
- … compute the formula below

$$r_s = 1 - \frac{6 \sum (\mathrm{R}(X_i) - \mathrm{R}(Y_i))^2}{m(m^2 - 1)}$$

Spearman’s Rank Correlation

- Example of a monotonic correlation, Spearman index is 1 while Pearson is only 0.88 because the correlation in non linear

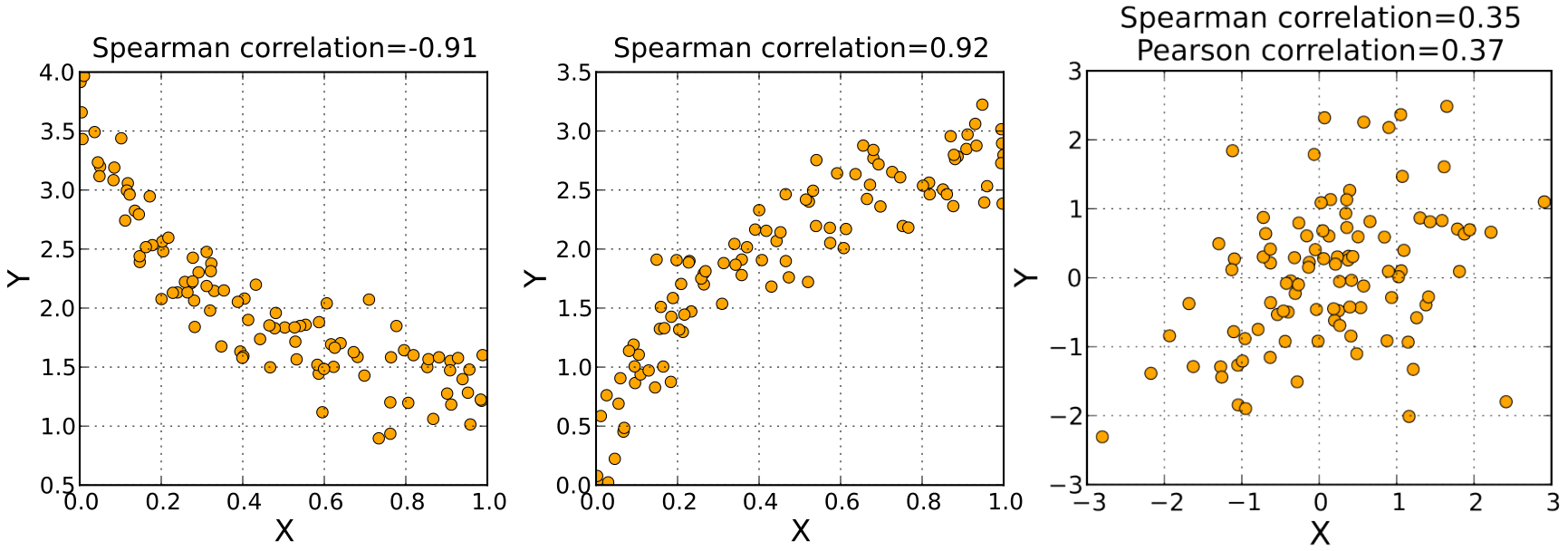


assesses monotonic relationships (whether linear or not)

Spearman’s Rank Correlation

Spearman's rank is

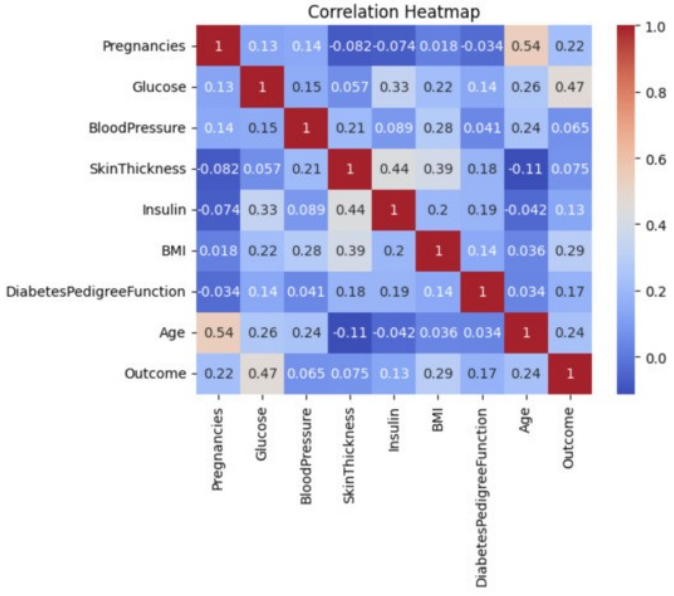
- 1 when two variables are correlated by an increasing monotonic function
- -1 if the function is decreasing monotonic
- 0 if there isn't a monotonic function correlation



Data visualization – more of it

Visualization: heatmap

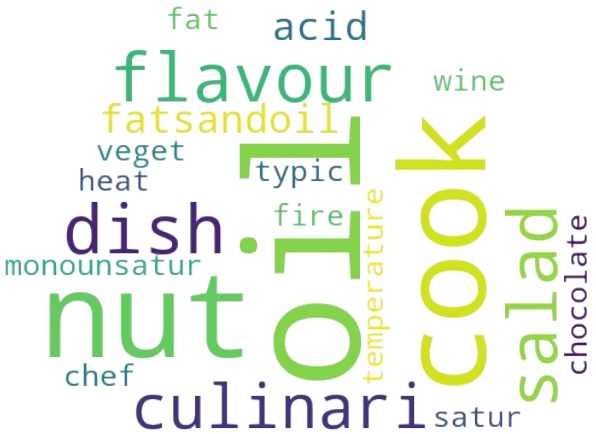
- Correlation matrix can be visualized through heatmaps to represents the correlation between all couples of variables
- Each component of the matrix represent two variables ($\mathbf{x}_i, \mathbf{x}_j$)
- The color of it represent the intensity of correlation between them
- Symmetric by construction



e.g. correlation between health and biometric features
Red dots represent high correlation, while blue dots low correlation.

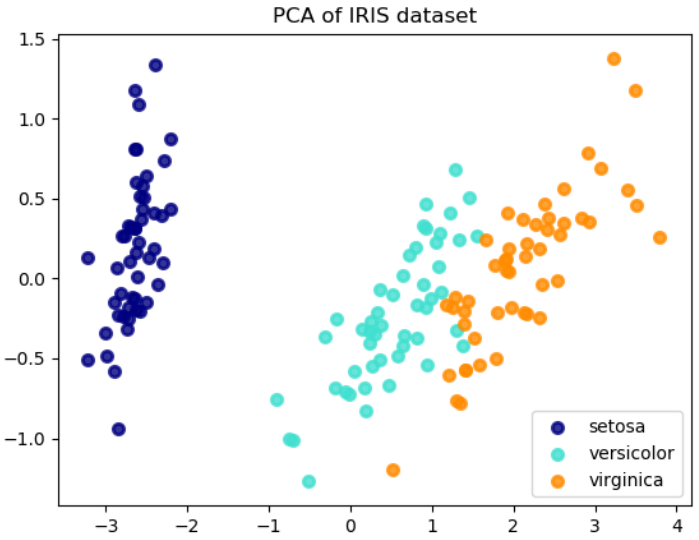
Visualization: word cloud

- In the context of text mining the word cloud can easily represents the topics of a group of similar documents
- Each word cloud contains the most important words characterizing a topic
- Font size/color represent frequency or other importance feature



Visualization: PCA

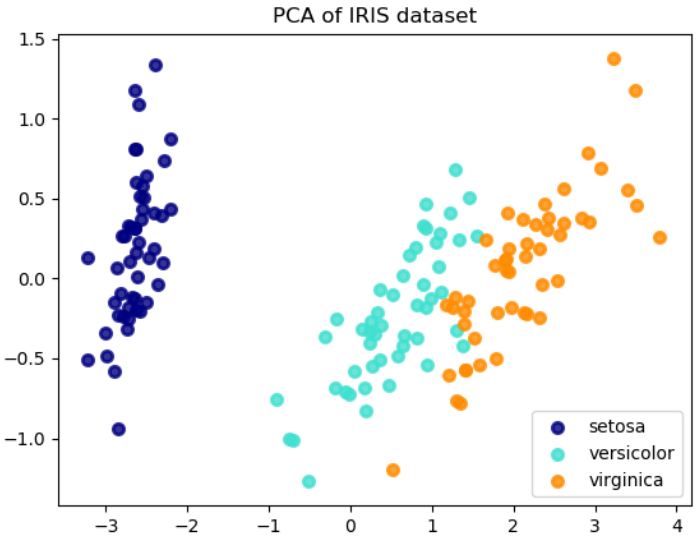
- In some cases it is useful to reduce the number of attributes to show information in plots
- **Principal component analysis (PCA)**



In the example, the **Iris** dataset wad reduced with **PCA** to two features and represented in scatter plot. Each color corresponds to the original labels.
See how the 3 categories are separated in bidimensional space

Visualization: PCA

- In some cases it is useful to reduce the size of attributes to show information in plots
- **Principal component analysis (PCA)**

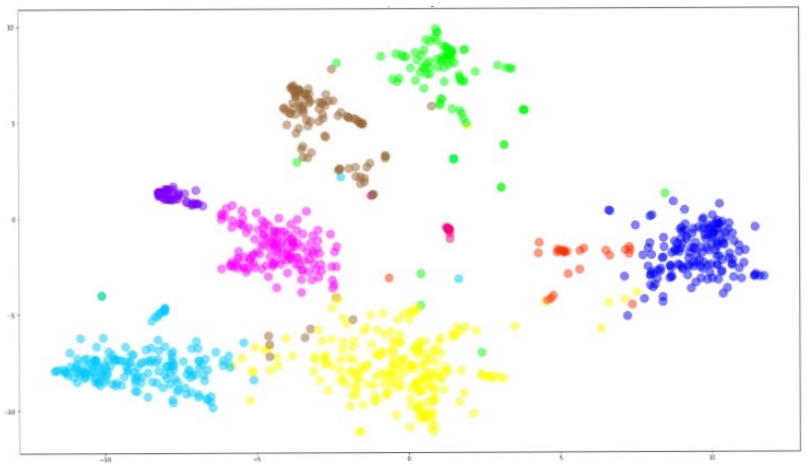


In the example, the **Iris** dataset wad

More on the next lecture

Visualization: t-SNE

- In some cases it is useful to reduce the size of attributes to show information in plots
- **t-distributed stochastic neighbor embedding** (t-SNE)

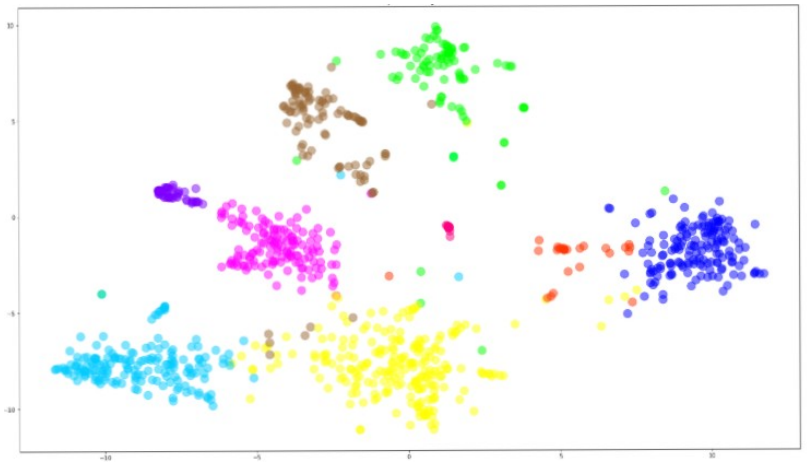


In the example t-SNE shows text document in bidimensional space. Each color corresponds to a cluster label

WARNING: using the t-SNE as dimensionality reduction technique in ML pipelines might not preserve the information of your data.

Visualization: t-SNE

- In some cases it is useful to reduce the size of attributes to show information in plots
- **t-distributed stochastic neighbor embedding** (t-SNE)



More on the next lecture

Data Visualization – take home message

- It is important to visualize your data when possible
  - To explore the raw input data
  - To analyze your output
- Choosing the correct visualization method is not trivial
  - Different kinds of analytics tasks require proper visualization techniques

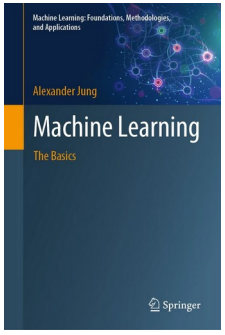
Any questions?

Self-assessment quiz

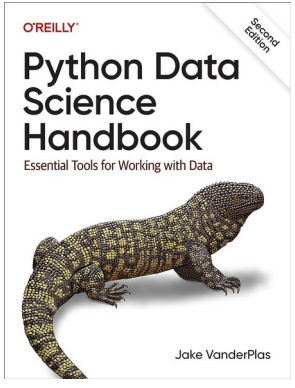
- Consider the following 12 samples
  - **x** = [0, 1, 1, 0 , 5, 4.5, 7, 10, 0, 5 , 6, -1]
  - Plot a histogram of the samples
  - Plot the ECDF
  - Plot the boxplot
  - Compute Pearson and Spearman correlation coefficient with these other 12 corresponding samples of another measurement/attribute:
  - **y** = [5, 1, 7, 4 , 9, 8, 7, 15, 5, 9 , 9, 5]

References: readings

- Chapter 1



- Chapter 9



Slide acknowledgments

- Tania Cequitelli and Elena Maria Baralis – Politecnico di Torino
- Think Stats, Allen B. Downey - Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists
- Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006