02450 Introduction to Machine Learning and Data Mining

Week 2: Summary statistics, similarity and visualization

Bjørn Sand Jensen

11 February 2025

DTU Compute, Technical University of Denmark

Today

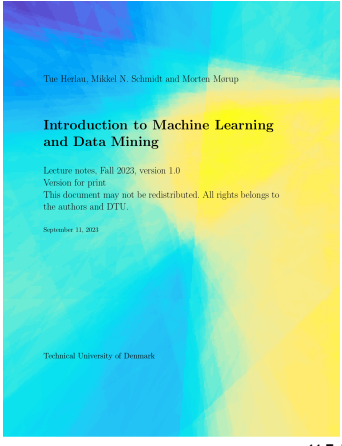
Feedback Groups of the day:

Benjamin Kauffmann Vener, Jonas Kristian Christiansen, Alexander Valdemar Vaaben, Ástríður ., Linus Casimir Adelswärd, Jens Oliver Klinke Jæger, Hallgrimur Thorsteinsson, Jonathan Schnack Ahrenkilde, Anwara Begum Sonia, Vitus Brandt Thaulow, Kei Ching Chan, Povilas Janusauskas, Kepeng Hong, Omid Ghaiby, Jonas Wentzel Sørensen, Panagiota Emmanouilidi, Victor Stubgaard, Carolina María Rodríguez Sánchez, Michael Rezaei, Guillermo Moya Fernández, Eleftheria Gitsouli, Jóhan Schade, Chrestine Hedeager Nielsen, Frederik Benjamin Kliim, Astrid Linnea Damgaard, Boon Keng Leck, Markus Kofod Thomasson, Daniel Amrhein, Florencia Sofía Illanes Vergara, Maja Møller Tranholm, Christian Nkya, Gustav Emil Christensen, Chayanee Suwannachat Christensen, Louise Høybye Sønder, Elisabeth Toft Nielsen, Frederik Wright Olsson, John Vinh Quang Tran, Ali Reza Yaghoubi, Theodora Georgakopoulou, Qinjiang Yang, Marcus Alexander Damgaard Sørensen, Clément Ravet, Hao Liu

Reading/homework material:

Chapter 4, Chapter 7

P4.2, P4.3, P4.5, P7.1



Lecture Schedule

1. Introduction
   4 February: C1,C2

Data: Feature extraction, and visualization

2. **Summary statistics, similarity and visualization**
   11 February: C4,C7
3. Computational linear algebra and PCA
   18 February: C3
4. Probability and probability densities
   25 February: C5, C6

Supervised learning: Classification and regression

5. Decision trees and linear regression
   4 March: C8, C9 **(Project 1 due 6 March at 17:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   11 March: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   18 March: C11, C13
8. Artificial Neural Networks and Bias/Variance
   25 March: C14, C15
9. AUC and ensemble methods
   1 April: C16, C17

Unsupervised learning: Clustering and density estimation

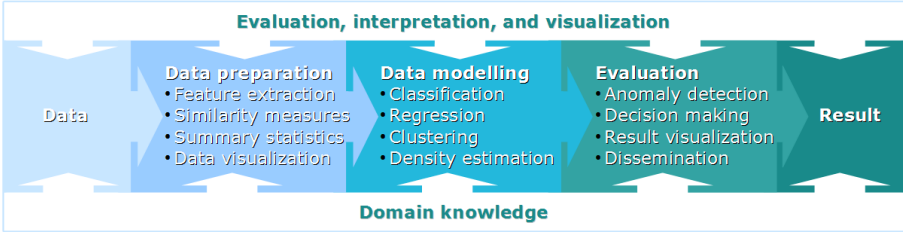
10. K-means and hierarchical clustering
    8 April: C18 **(Project 2 due 10 April at 17:00)**
11. Mixture models and density estimation
    22 April: C19, C20
12. Association mining
    29 April: C21

Recap

13. Recap and discussion of the exam
    6 May: C1-C21

Online help: Piazza
Videos of lectures: `https://panopto.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)

Learning Objectives



Learning Objectives

- Understand and apply a range of basic statistics for summarizing a dataset
- Compute measures of similarity/dissimilarity (Lp distance, cosine distance, etc.)
- Understand and apply a wide range of data visualization approaches
- Understand good practice in plotting including Tufte's guidelines

Plan for today:

- Lecture 2 (13:00 – ∼15:00)
  - **Practicalities and announcements**
  - **Data revisited**
  - **Summary statistics**
  - **Measures of similarity and dissimilarity**
  - Break
  - **Visualizaiton**
  - **A note on exercises**
- Exercises (15:00–17:00)

Practicalities and announcements

- AV/streaming issues.
- Exam: No printed notes! You can not make the notes on your iPad / computer / touchscreen and print them and bring them to the exam.
- Exam: What is a non-programmable calculator?
- Exam date (currently) 28th May 2025. See `https://www.inside.dtu.dk/en/undervisning/regler/regler-for-eksamen/eksamensdatoer`

Data

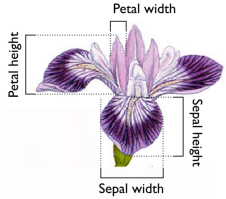
- **Dataset types**: Tabular, relational, sequence
- **Data issues**: Missing, noisy, quality.
- **Attribute types**: Nominal, Ordinal, Interval, Ratio
- **Transformations** of attributes

- **Discrete**
  - Finite (or countably infinite) set of values
- **Continuous**
  - Real number

- **Nominal** (Equal / Not equal)
  - Objects belong to a category
- **Ordinal** (Greater than / Less than)
  - Objects can be ranked
- **Interval** (Addition / Subtraction)
  - Distance between objects can be measured
- **Ratio** (Multiplication / Division)
  - Zero means absence of what is measured

How do we represent, summarise, validate, and understand data before we can carry out machine learning and modelling?

Data points and matrices



A data point is a set of **real** numbers represented as a vector

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \in \mathbb{R}^M \quad \text{e.g. } \boldsymbol{x} = \begin{bmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{bmatrix}$$

A matrix is used to hold multiple data points (as rows)

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \dots & x_{1,M} \\ \vdots & & \vdots \\ x_{N,1} & \dots & x_{N,M} \end{bmatrix} \in \mathbb{R}^{N \times M}$$

$$\mathbf{X} = \begin{array}{c} \begin{matrix} \text{Sepal Length} & \text{Sepal Width} & \text{Petal Length} & \text{Petal Width} \end{matrix} \\ \begin{bmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3.0 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.2 \\ 4.6 & 3.1 & 1.5 & 0.2 \\ 5.8 & 2.7 & 5.1 & 1.9 \\ \vdots & \vdots & \vdots & \vdots \\ 5.7 & 2.8 & 4.1 & 1.3 \end{bmatrix} \end{array}$$

Matrix basics

- Common matrix notation

$$\boldsymbol{A}, A, \overline{\overline{A}}$$

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,M} \\ \vdots & & \vdots \\ a_{N,1} & \cdots & a_{N,M} \end{bmatrix} \in \mathbb{R}^{N \times M}$$

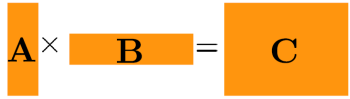
- Common vector notation

$$\boldsymbol{x}, x, \overline{x}, \vec{x}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \in \mathbb{R}^{M}$$

Matrix multiplication

- Two matrices can be multiplied $\boldsymbol{AB} = \boldsymbol{C}$
  - if the number of columns in the first equals the number of rows in the second



$$L \times M \qquad M \times N \qquad L \times N$$

$$\underset{3\times 4 \text{ matrix}}{\begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 2 & 3 & 4 \end{bmatrix}} \underset{4\times 5 \text{ matrix}}{\begin{bmatrix} \cdot & \cdot & \cdot & a & \cdot \\ \cdot & \cdot & \cdot & b & \cdot \\ \cdot & \cdot & \cdot & c & \cdot \\ \cdot & \cdot & \cdot & d & \cdot \end{bmatrix}} = \underset{3\times 5 \text{ matrix}}{\begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & x_{3,4} & \cdot \end{bmatrix}}$$

$$x_{3,4} = 1 \cdot a + 2 \cdot b + 3 \cdot c + 4 \cdot d$$

Example: $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$

Matrix transpose

- The transpose of a matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad \boldsymbol{A}^\top = \begin{bmatrix} 1 & 3 & 7 \\ 2 & 4 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

- Transpose of a sum

$$(\boldsymbol{A} + \boldsymbol{B})^\top = \boldsymbol{A}^\top + \boldsymbol{B}^\top$$

- Transpose of a product

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$$

$$(\boldsymbol{Ax})^\top \boldsymbol{y} = \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{y} = \boldsymbol{x}^\top (\boldsymbol{A}^\top \boldsymbol{y})$$

Matrix inverse

- Ones on the diagonal and zeros everywhere else

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \qquad \boldsymbol{I}^\top = \boldsymbol{I}$$

- Multiplying by the identity does not change anything

$$\boldsymbol{I}\boldsymbol{A} = \boldsymbol{A}$$

$$\boldsymbol{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\boldsymbol{I}_2\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

- For a square matrix, the inverse satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$$

Summary Statistics

Empirical statistics

Given two samples $x_1, x_2, \ldots, x_N \in \mathbb{R}$ and $y_1, y_2, \ldots, y_N \in \mathbb{R}$:

- Empirical mean
  $\hat{\mu} = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- Empirical variance
  $\hat{s} = \hat{\mathrm{var}}[x] = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$
- Empirical covariance
  $\hat{\mathrm{cov}}[x, y] = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$
- Empirical standard deviation
  $\hat{\sigma} = \hat{\mathrm{std}}[x] = \sqrt{\hat{s}}$

$$\mathbf{X} = \begin{bmatrix} \text{Sepal Length} & \text{Sepal Width} & \text{Petal Length} & \text{Petal Width} \\ 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3.0 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.2 \\ 4.6 & 3.1 & 1.5 & 0.2 \\ 5.8 & 2.7 & 5.1 & 1.9 \\ \vdots & \vdots & \vdots & \vdots \\ 5.7 & 2.8 & 4.1 & 1.3 \end{bmatrix}$$

Quantiles and percentiles

Given N observations of an attribute $x_1, x_2, \ldots, x_N \in \mathbb{R}$.

Quantiles describe the *points* that divide the underlying distribution into intervals that are equally probable:

- The one 2-quantile (**median**) divides the distribution in two intervals.
- The three 4-quantiles (**quartiles**) divides the distribution in four intervals.
- The 99 100-quantiles (**percentiles**) divides the distribution in 100 intervals.

The **median** is the same as the 2nd quartile or the 50th percentile.

E.g., we can (approximately) find the **median** by

- Sort the observations in ascending order $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$

$$\text{median}[x] = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{if } N \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right) & \text{if } N \text{ is even.} \end{cases}$$

Covariance

- Consider the case where we have more than two attributes and we want to estimate the (linear) releationship between all of them (pairwise).
- Empirical covariance

$$\hat{\mathrm{cov}}[\boldsymbol{x}^{(k)}, \boldsymbol{x}^{(m)}] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i^{(k)} - \hat{\mu}_{x^{(k)}})(x_i^{(m)} - \hat{\mu}_{x^{(m)}})$$

- Empirical covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\mathrm{cov}}[x^{(1)}, x^{(1)}] & \dots & \hat{\mathrm{cov}}[x^{(1)}, x^{(M)}] \\ \vdots & & \vdots \\ \hat{\mathrm{cov}}[x^{(M)}, x^{(1)}] & \dots & \hat{\mathrm{cov}}[x^{(M)}, x^{(M)}] \end{bmatrix}$$

$$\mathbf{X} = \begin{array}{c} \begin{matrix} \text{Sepal Length} & \text{Sepal Width} & \text{Petal Length} & \text{Petal Width} \end{matrix} \\ \begin{bmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3.0 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.2 \\ 4.6 & 3.1 & 1.5 & 0.2 \\ 5.8 & 2.7 & 5.1 & 1.9 \\ \vdots & \vdots & \vdots & \vdots \\ 5.7 & 2.8 & 4.1 & 1.3 \end{bmatrix} \end{array}$$

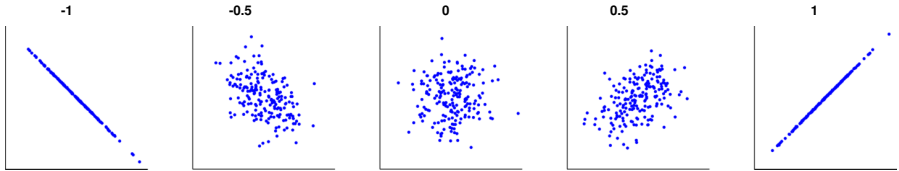
$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.7 & 0.0 & 1.3 & 0.5 \\ 0.0 & 0.2 & -0.3 & -0.1 \\ 1.3 & -0.3 & 3.1 & 1.3 \\ 0.5 & -0.1 & 1.3 & 0.6 \end{bmatrix}$$

Correlation

- Measure of degree of linear relationship

$$\hat{\text{cor}}[x, y] = \frac{\hat{\text{cov}}[x, y]}{\hat{\sigma}_x \hat{\sigma}_y}$$

A correlation of 1 or -1 means that there is a perfect linear relation



Dissimilarity and Similarity

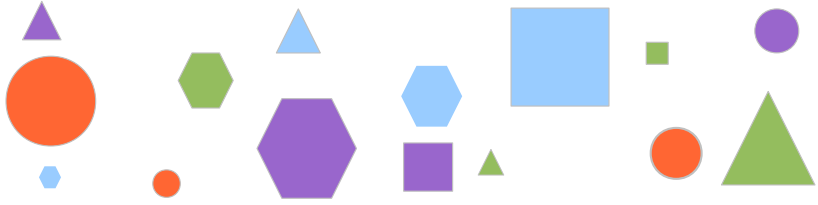
Similarity / Dissimilarity measures

Similarity $s(\boldsymbol{x}, \boldsymbol{y})$ Often between 0 and 1. Higher means more similar

Dissimilarity $d(\boldsymbol{x}, \boldsymbol{y})$ Greater than 0. Lower means more similar.

Classification Classify a document as having the same topic y as the document it is **most similar**/**least dissimilar** to.

Outlier detection The observation most **dissimilar** to all other observations is an outlier



Vector spaces

Any vector of given dimension, M, lies in a vector space, called $\mathbb{R}^M$. It has the operations of:

- **scalar multiplication**: $a\boldsymbol{x}$ is defined for any scalar a.
  For real vectors, $a\boldsymbol{x} = [ax_1, ax_2, \ldots ax_M]$, i.e. element-wise scaling. $(\mathbb{R}, \mathbb{R}^M) \rightarrow \mathbb{R}^n$
- **vector addition**: $\boldsymbol{x} + \boldsymbol{y}$ vectors $\boldsymbol{x}, y$ of equal dimension. For real vectors, $\boldsymbol{x} + \boldsymbol{y} = [x_1 + y_1, x_2 + y_2, \ldots x_d + y_d]$ the element-wise sum. $(\mathbb{R}^M, \mathbb{R}^M) \rightarrow \mathbb{R}^M$
- **linear combinations** results in a vector in the same vector space (i.e. closed under linear combinations):
  $a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \cdots + a_N\boldsymbol{x}_N$ for $a_k \in \mathbb{R}$

We will consider vector spaces which are equipped with two additional operations:

- **a norm** $||\boldsymbol{x}||$ which allows the length of vectors to be measured. $\mathbb{R}^M \rightarrow \mathbb{R}_{\geq 0}$
- **an inner product** $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, $\boldsymbol{x}^\top \boldsymbol{y}$ or $\boldsymbol{x} \bullet y$ allows the angles of two vectors to be compared. The inner product of two orthogonal vectors is 0. For real vectors
  $\boldsymbol{x} \bullet \boldsymbol{y} = x_1y_1 + x_2y_2 + x_3y_3 \ldots x_dy_d$. $(\mathbb{R}^M, \mathbb{R}^M) \rightarrow \mathbb{R}$

Norms: How big is that vector / matrix?

- The (Euclidian) norm of a vector measures it's length (magnitude):

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{n=1}^{N} |x_n|^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$$

- The Frobenius norm of a matrix measures it's magnitude:

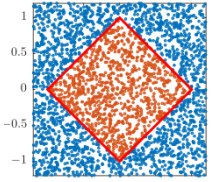
$$\|\boldsymbol{X}\|_F^2 = \sum_{i,j} x_{i,j}^2 = \mathrm{trace}(\boldsymbol{X}\boldsymbol{X}^T) = \mathrm{trace}(\boldsymbol{X}^T\boldsymbol{X})$$

Where trace takes the sum of the diagonal elements, i.e. $\mathrm{trace}(\boldsymbol{A}) = \sum_{i=1}^{N} a_{i,i}$

Dissimilarity measures

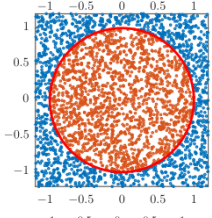
- General Minkowsky distance (p-distance)

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{M} |x_j - y_j|^p \right)^{\frac{1}{p}} = ||\boldsymbol{x} - \boldsymbol{y}||_p$$



- One-norm ($p = 1$)

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{M} |x_j - y_j|$$

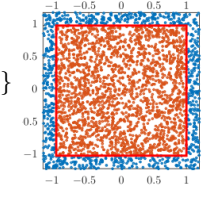


- Euclidean ($p = 2$)

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{M} (x_j - y_j)^2}$$

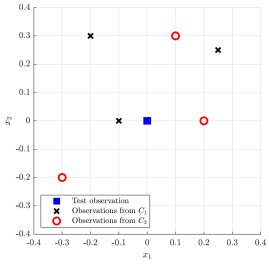
- Max-norm distance ($p = \infty$)

$$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max \{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_M - y_M|\}$$



Usage: Regularization and alternative optimization targets.
For instance, $p = \infty$ is very affected by outliers, $p = 1$ much less so.

Quiz 1: Norms and distances



	x_1	x_2
$\boldsymbol{o}_1$	0.25	0.25
$\boldsymbol{o}_2$	−0.2	0.3
$\boldsymbol{o}_3$	−0.1	0.0
$\boldsymbol{o}_4$	0.1	0.3
$\boldsymbol{o}_5$	−0.3	−0.2
$\boldsymbol{o}_6$	0.2	0.0

Figure 8: The synthetic dataset, where the colors indicate the classes such that the black observations $\{o_1, o_2, o_3\}$ belong to class C_1, and the red observations $\{o_4, o_5, o_6\}$ belong to class C_2.

Consider a small synthetic dataset consisting of six observations and two attributes. The subset is presented in the figure on the left. We consider the p-distances d_1, d_2, d_∞. A test point, $\boldsymbol{o}^*$, is located at $[0, 0]^\top$.

Determine which one of the following statements is correct?

A $d_1(\boldsymbol{o}^*, \boldsymbol{o}_3) > d_2(\boldsymbol{o}^*, \boldsymbol{o}_6)$

B $d_\infty(\boldsymbol{o}_2, \boldsymbol{o}_5) = 1$

C $\left\| \boldsymbol{o}^* - \begin{bmatrix} -0.2 \\ 0.3 \end{bmatrix} \right\|_2^2 = 0.13$

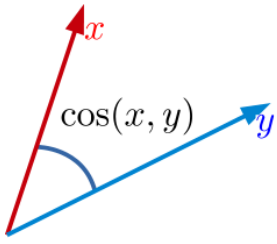
D $d_p(\boldsymbol{o}_4, \boldsymbol{o}_k) > 0.1 \ \forall k \in [1, 2, 3, 4, 5, 6]$ $\forall p \in \{1, 2, \infty\}$

E Don't know

Similarity measures

Cosine similarity

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$



Similarity measures
— mostly for binary vectors

$$\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

K : Total number of attributes

f_{00} : Number of attributes where $x_k = y_k = 0$

f_{11} : Number of attributes where $x_k = y_k = 1$

Simple Matching Coefficient (SMC)

$$\mathrm{SMC}(\boldsymbol{x}, \boldsymbol{y}) = \frac{f_{00} + f_{11}}{K}$$

Jaccard Coefficient

$$\mathrm{J}(\boldsymbol{x}, \boldsymbol{y}) = \frac{f_{11}}{K - f_{00}}$$

Extended Jaccard coefficient

$$\mathrm{EJ}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \boldsymbol{x}^\top \boldsymbol{y}}$$

Quiz 2: Similarity measures

Calculate the simple matching coefficient, Jaccard, cosine and extended jaccard similarity between customer 1 and customer 2 in the market basket data below.

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	0	1	1	0	1

Which of the following statements are true?

A. $\mathrm{SMC}(o_1, o_2) = \frac{3}{5}$ $J(o_1, o_2) = \frac{1}{2}$, $\cos(o_1, o_2) = \frac{2}{3}$,

B. $\mathrm{SMC}(o_1, o_2) = \frac{3}{5}$ $J(o_1, o_2) = \frac{3}{4}$, $\cos(o_1, o_2) = \sqrt{\frac{2}{3}}$,

C. $\mathrm{SMC}(o_1, o_2) = \frac{2}{5}$ $J(o_1, o_2) = \frac{1}{3}$, $\cos(o_1, o_2) = \frac{2}{3}$,

D. $\mathrm{SMC}(o_1, o_2) = \frac{2}{5}$ $J(o_1, o_2) = \frac{1}{3}$, $\cos(o_1, o_2) = \sqrt{\frac{2}{3}}$,

E. Don't know.

K : Total number of attributes

f_{00} : Number of attributes where $x_k = y_k = 0$

f_{11} : Number of attributes where $x_k = y_k = 1$

$$\mathrm{SMC}(x, y) = \frac{f_{00} + f_{11}}{K}$$

$$\mathrm{J}(x, y) = \frac{f_{11}}{K - f_{00}}$$

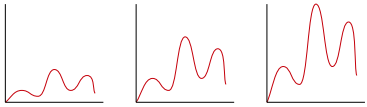
$$\cos(x, y) = \frac{x^\top y}{\|x\|_2 \|y\|_2}$$

$$\mathrm{EJ}(x, y) = \frac{x^\top y}{\|x\|_2^2 + \|y\|_2^2 - x^\top y}$$

Invariance

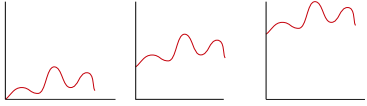
Scale invariance

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\alpha \boldsymbol{x}, \boldsymbol{y})$$



Translation invariance

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\alpha + \boldsymbol{x}, \boldsymbol{y})$$



General invariances

$$d\left(\boxed{6}, \boxed{2}\right) = d\left(\boxed{6}, \boxed{2}\right)$$

Transformations

Standardization: Ensure a single attribute will not dominate:

$$\tilde{x}_i^{(k)} = \frac{x_i^{(k)} - \hat{\mu}_k}{\hat{\sigma}_k}$$

Example:
- **Number of children** 0-5
- **Age** 0-100 years
- **Annual income** 0-50.000 €

Combining heterogeneous attributes Transform measures and combine

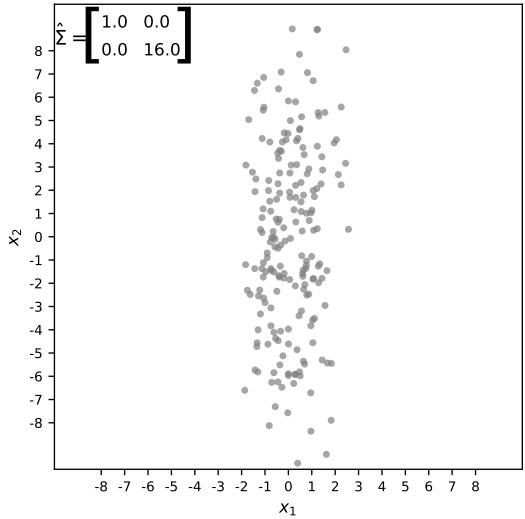
$$s_{\text{Edu.}} = \text{SMC}(x_{\text{Edu.}}, y_{\text{Edu.}})$$
$$s_{\text{Age.}} = a\left(a + d_1(x_{\text{Age.}}, y_{\text{Age.}})\right)^{-1},\ a = 1$$
$$s(x,y) = \frac{1}{2}\left(s_{\text{Edu.}} + s_{\text{Age.}}\right)$$

Example:
- **Age:** Continuous
- **Education:** Binary
  - Primary (yes/no)
  - Secondary (yes/no)
  - Tertiary (yes/no)

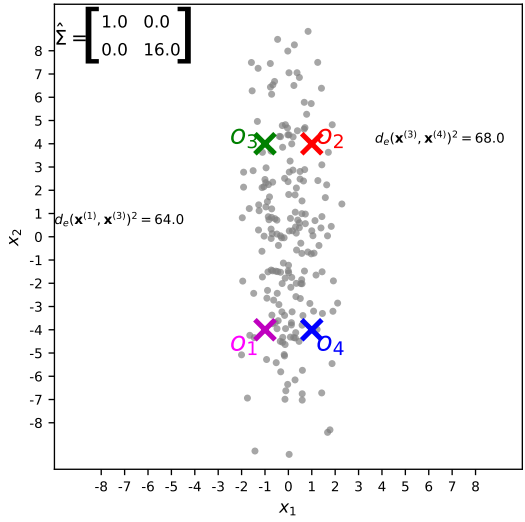
Weighting Attributes have different importance

$$s(x,y) = 0.99 s_{\text{Edu.}} + 0.01 s_{\text{Age.}}$$

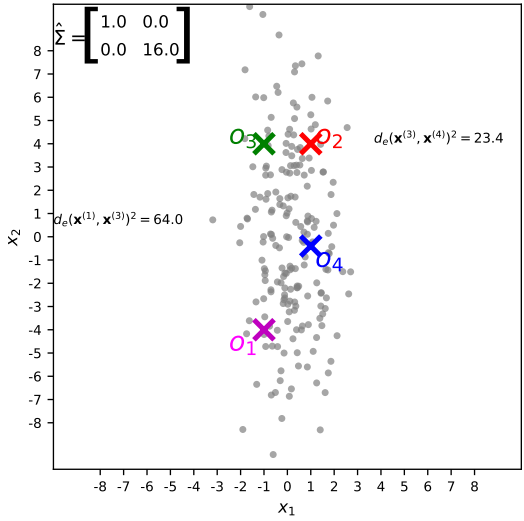
Scale and Standardization



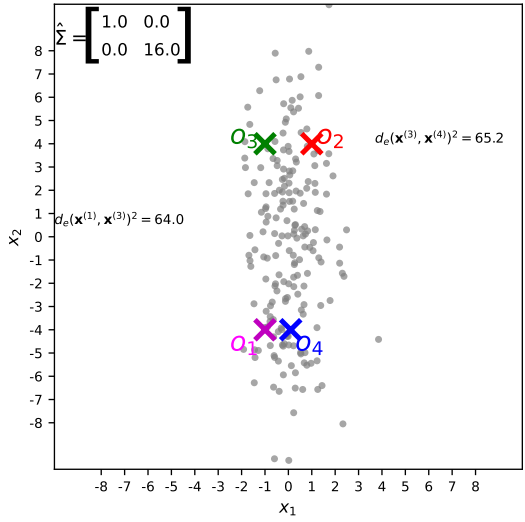
Scale and Standardization



Scale and Standardization

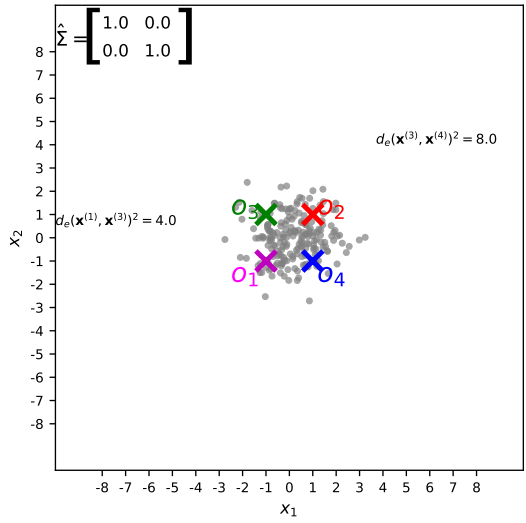


Scale and Standardization

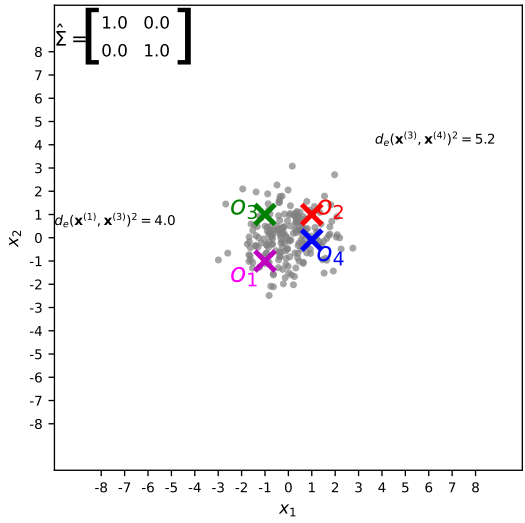


Scale and Standardization

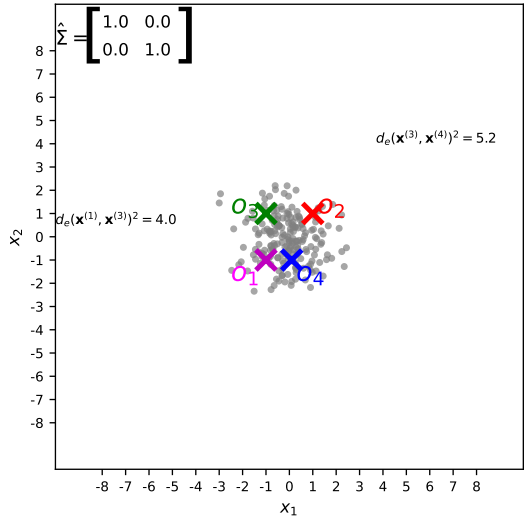
$$\tilde{x}_i^{(k)} = \frac{x_i^{(k)} - \hat{\mu}_k}{\hat{\sigma}_k}$$



Scale and Standardization



Scale and Standardization

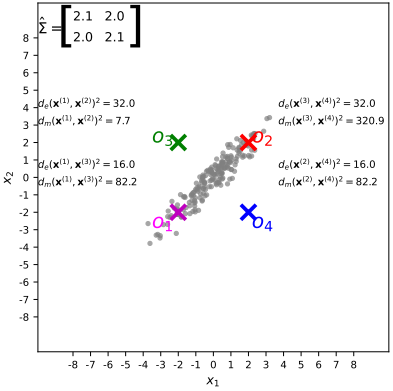


Mahalanobis distance

Define a (squared) distance that takes into account variance and covariance.

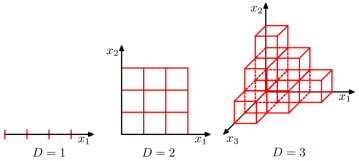
$$d_{euclidian}(\boldsymbol{x}^{(k)}, \boldsymbol{x}^{(m)})^2 = (\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(m)})^\top \boldsymbol{I}^{-1} (\boldsymbol{x}^{(k} - \boldsymbol{x}^{(m)})$$

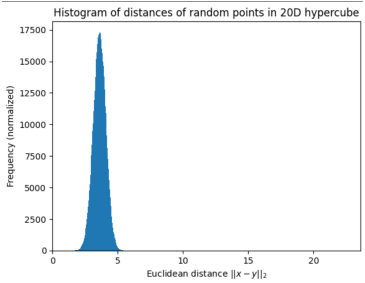
$$d_{mahanalobis}(\boldsymbol{x}^{(k)}, \boldsymbol{x}^{(m)})^2 = (\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(m)})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(m)})$$



What happens in (very) high dimensions...

- The volume of a unit-cube in high dimesnions is 1, but the number of points needed to densely cover the space grows exponentially.
- Imagine a high dimensional box (e.g. 20D), fill it with random points. For any given point, in high enough dimension, the boundaries of the box will be closer than any other point in the box.
- In high dimensions, the ratio of the smallest distance to the largest distance between points approaches 1. This means that all points become nearly "equidistant" from each other.
- ... many more paradoxes to be explored





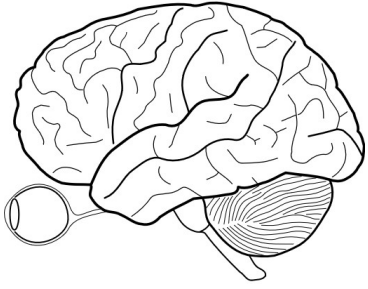
Break

Next up: *Visualization*

Visualization

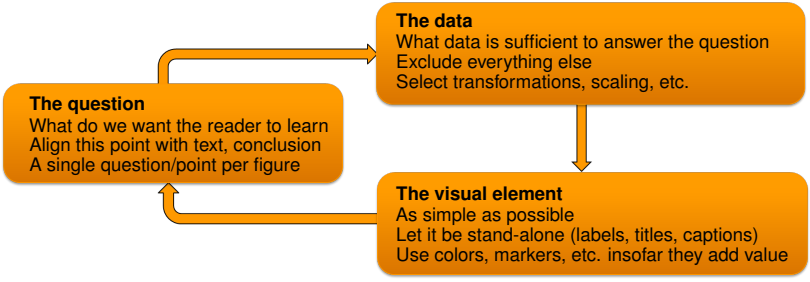
The drawing shows me at one glance what might be spread over ten pages in a book."
- Ivan S. Turgenev's novel Fathers and Sons, 1862.
Use a picture. It's worth a thousand words."
- Arthur Brisbane to the Syracuse Advertising Men's Club, in March 1911

- A main function of the brain is to process visual information
- We can exploit this capacity using visualization of the data in order to:
  - Detect new patterns, i.e. exploratory data analysis)
  - **Dissiminate results, i.e. visualizations/plots in written work (today)**
- We should take into account how the brains visual system works



Illustrations as technical writing

- The purpose of the text is to communicate an idea **(vs. plots has a purpose)**
- Be grammatically correct **(vs. elementary "rules" of good plotting)**
- Ensure the text is readable **(vs. labels, legends or lines nobody can read)**
- Avoid long/complicated paragraph **(vs. plots that are overly complicated)**
- Don't lie or exagerate **(vs. distort data in a plot)**



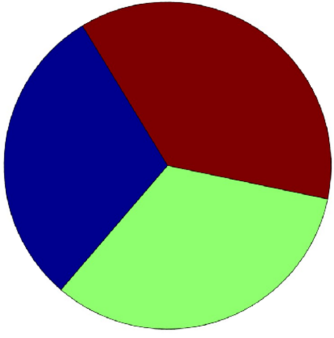
Important choices for visualizations

- **Representation:** How will you map objects, attributes, and relations to visual elements?
  - Positions, lengths, colors, areas, orientation
- **Arrangement** How will you display the visual elements?
  - Viewpoint, transparency, separation, grouping
- **Selection:** How will you handle a large number of attributes and data objects?
  - Display a subset, focus on a region of interest, show summaries

Representation

Area represents proportion

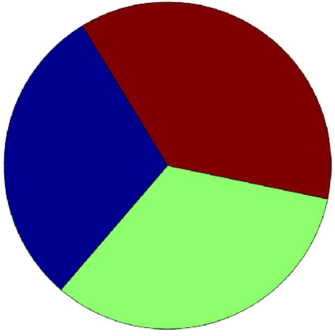
- Which is smallest, middle, and largest?
- What are the proportions approximately?



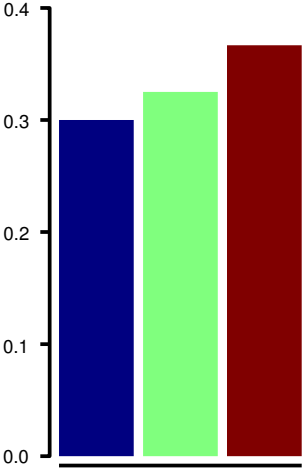
Representation

Area represents proportion

- Which is smallest, middle, and largest?
- What are the proportions approximately?



- **Height represents proportion**



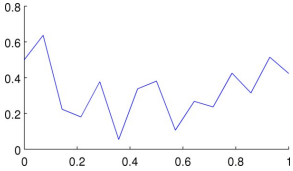
Selection

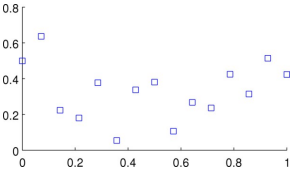
- Elimination or de-emphasis of certain objects or attributes
- A subset of **attributes**
  - **Why?** A graph can only show so many attributes – focus on the relevant
  - How?
    - Dimensionality reduction
    - Plot pairs of attributes
- A subset of **objects**
  - **Why?** A graph can only show so many objects – focus on the relevant
    - Random sampling
    - Display of region of interest
    - Use density estimation

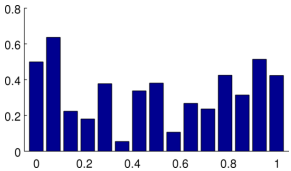
Types of plots

- **Distribution of a single attribute**
  - Histogram
  - Empirical cumulative distribution
  - Percentile plots
  - Box plot
- **Relation among attributes**
  - 2D histogram
  - Heat maps and contour plots
  - Scatter plots
- **Visualization of high-dimensional objects**
  - Matrix plots
  - Parallel coordinates
  - Star plots

Basic plots


0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1


0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1


0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

The Iris Dataset

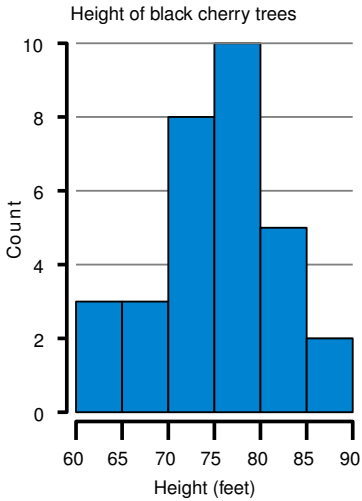
- **Three types of Iris flowers**
  - 50 instances of each class, 150 in total
- **Attributes**
  - Sepal (outermost leaves)
    - length in cm
    - width in cm
  - Petal (innermost leaves)
    - length in cm
    - width in cm
  - Class of flower
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

Fower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.	.	.	.	.
.	.	.	.	.
150	5.9	3.0	5.1	1.8

Distribution of a single attribute

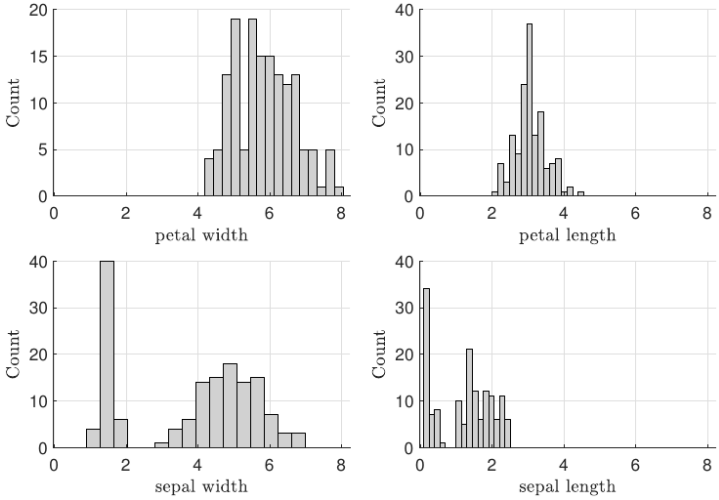
Histograms

- Shows distribution of a single variable
  - Divide the values into bins
  - Bar plot of the number of values in bin
  - Height indicates count of values
  - Shape determined by
    - Distribution of data
    - Number of bins / bin width

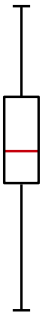


$$\mathbf{x}^\top = [60, 64, 64, 66, 67, 69, 71, 72, 72, 72, 72, 73, 74, 74, 75, 75, 76, 76, 76, 77, 77, 78, 78, 79, 80, 80, 81, 82, 84, 85, 85, 89]$$

Histograms of the Iris data attributes



Box plots



min (max(data), $q_{75} + 1.5(q_{75} - q_{25})$)

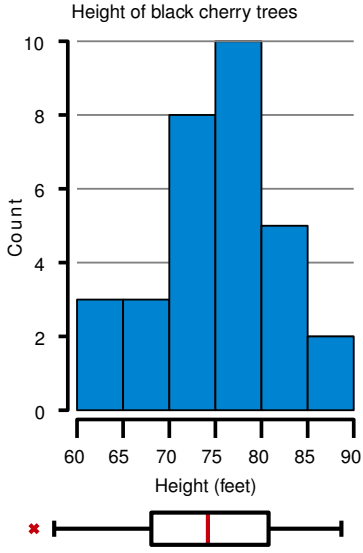
75th percentile (q_{75})

Median (q_{50})

25th percentile (q_{25})

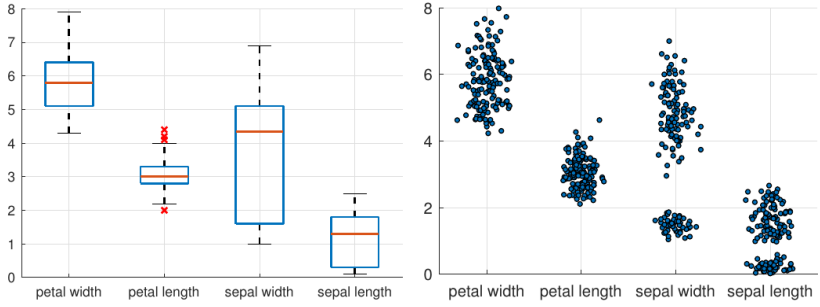
max (min (data), $q_{75} - 1.5(q_{75} - q_{25})$)

× Extreme value



The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier.

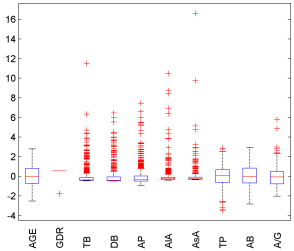
Box plots


8
7
6
5
4
3
2
1
0
petal width
petal length
sepal width
sepal length
8
6
4
2
0
petal width
petal length
sepal width
sepal length

Quiz 3: Boxplots

No.	Attribute description	Abbrev.
x_1	Age (in years)	AGE
x_2	Gender (Female=0, Male=1)	GDR
x_3	Total Bilirubin	TB
x_4	Direct Bilirubin	DB
x_5	Alkaline Phosphotase	AP
x_6	Alamine Aminotransferase	AlA
x_7	Aspartate Aminotransferase	AsA
x_8	Total Protiens	TP
x_9	Albumin	AB
x_{10}	Albumin to Globulin ratio	A/G
y	0=No liver disease, 1=Liver disease	LD

Table 1: Liver disease dataset.



The attributes x_1–x_{10} are standardized (i.e., the mean has been subtracted each attribute and the attributes divided by their standard deviations). The figure shows a boxplot fo the standardized data. Which of the following statements is *correct*?

A. The value of the 50th and 75th percentiles of the attribute DB coincides.

B. Even though the distribution of AlA and AsA may have a similar shape this does not imply that the two attributes are correlated.

C. The attribute TB is likely to be normal distributed.

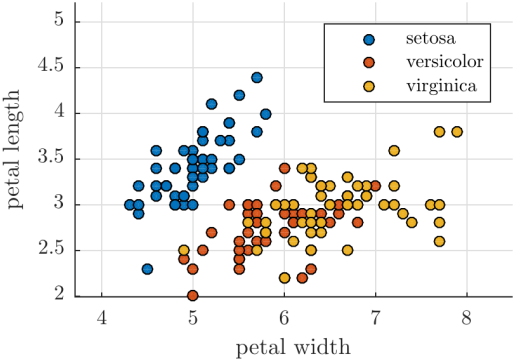
D. The attribute GDR has a clear outlier that should be removed.

E. Don't know.

Relation between attributes

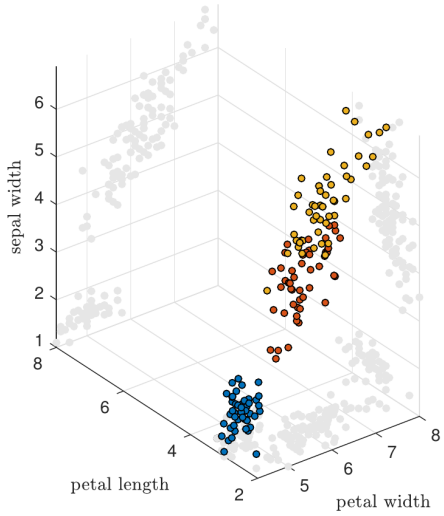
Scatter plots

- Shows **relations** between attributes
  - Assess dependence between attributes
  - Used with classes to assess separability



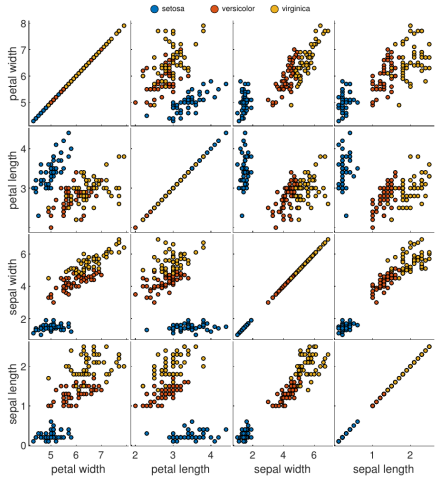
Scatter plots

- Shows **relations** between attributes
  - Assess dependence between attributes
  - Used with classes to assess separability
  - 3D plots are often confusing; avoid if possible



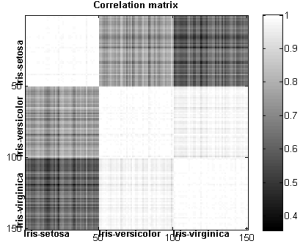
Scatter plots

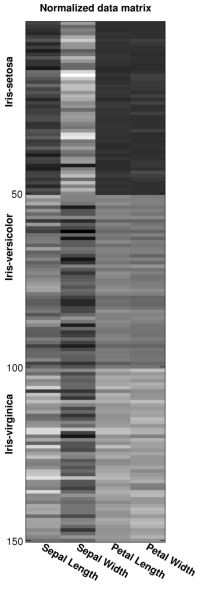
- Scatter plot matrix — all pairs of attributes



Matrix plots

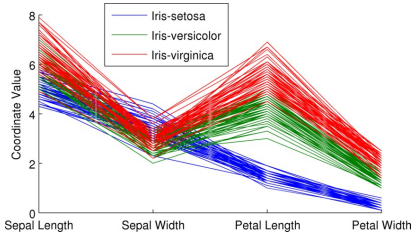
- Plot of raw data matrix
  - Useful when objects are sorted according to class
  - Typically, attributes are normalized
- Plots of similarity matrices
  - Useful for visualizing the relation between objects

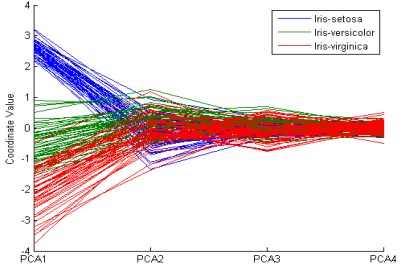




Parallel coordinates

- **Plot high-dimensional data**
- Instead of perpendicular axes
  - Use parallel axes
- Attribute values are plotted as a point
  - and the points are connected by a line
- Each object is represented as a line
- Lines representing a group of objects
  - Are similar in some sense
  - Ordering of attributes is important in seeing such groupings





ACCENT

- **Apprehension**
  - Is it easy to see what is important in the graph?
- **Clarity**
  - Are the most important elements visually most prominent?
- **Consistency**
  - Have you used the same colors, shapes, etc. as in other graphs?
- **Efficiency**
  - Does it convey its information in the most simple and efficient way?
- **Necessity**
  - Are all elements of the graph necessary to represent data?
- **Truthfulness**
  - Does the graph represent the data correctly?

Tufte's guidelines

Graphical excellence

- Well-designed presentation of interesting data – a matter of
  - substance, statistics, and design
- Complex ideas communicated with
  - clarity, precision, and efficiency
- Gives the viewer
  - the greatest number of ideas
  - in the shortest time
  - with the least ink
  - in the smallest place.
- Nearly always multivariate
- Requires telling the truth about the data
- Maximise Data-ink ratio:

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used}}$$

Summary

Summary

- Data points lies in a vector space with certain properties (norms, inner products, addition, scaling)
- Similarity / dissimilarity can be measureed (for both continuous and binary vectors)
- Visualization for dissimination and exploration
  - ACCENT

A note on exercises in VScode and Jupyter (Python) (only if time permits)

- Demo

Resources

https://www.3blue1brown.com An great, animated recap of linear algebra (https://www.3blue1brown.com/essence-of-linear-algebra-page/)

https://vita.had.co.nz An attempt at formalizeing the elements of a graphic (e.g. a plot) (https://vita.had.co.nz/papers/layered-grammar.html)

https://junkcharts.typepad.com Excellent resource on creating good visualizations (https://junkcharts.typepad.com/junk_charts/)

http://www2.imm.dtu.dk Our demo of the multivariate normal distribution which illustrates the effect of the covariance matrix" (http://www2.imm.dtu.dk/courses/02450/DemoNormal.html)