Last lecture: Perceptron – Machine Learning

- Computational task classification: Assign one of a discrete number of labels to a (multidimensional) input.
  - A typical supervised machine learning task: Given a set of lablled training data $\{(x_1,y_1), (x_2,y_2) \ldots (x_n,y_n)\}$, discover a function f that approximtaes the true function that generates the labels y, such that $y=f_{TRUE}(x)$
- If two classes can be linearly separated, a perceptron can be iteratively parametrised, and the perceptron learning algorithm will converge towards a solution
  - Doesn‘t slowly converge but jumps around
  - Not necessarily the optimal one – inclusion of optimsiation criteria such as maximising distance to separation function is desirable
  - Convergence not guaranteed if classes are not linearly separable

Review: Which questions can different knowledge representation and reasoning paradigms answer?

Rules

- Represent: How are states/conditions related to each other?
- Questions: Which facts follow from the knowledge base? Is X true? Is the knowledge base consistent?

Object-oriented KR/Logic:

- Represent: How are different types of things (classes, concepts, frames) related to each other?
- Questions: Is the knowledge base consistent? What general (terminological) and assertional axioms (="truths") can be inferred? Is X true?

Graphs

- Represent: How are entities related to each other (weighted or labelled relationships)?
- Questions: How are X and Y related to each other? Which entities are important in the network? If X is important/affected, which other entities are important/affected? (spreading activation)?

Vector representation

- Represents: Complex entities as a collection of numbers (=vector), with informally defined semantics (=documentation of features in the vector)
- Questions: How similar are X and Y to each other – application in IR and RecSys? What category is X (classification)? Which groups can we identify in a larger dataset (clustering)?

Today: Shift away from entities towards variables = features = aspects of interest with respect to (interesting) complex entities or phenomena

Correlation

Introduction to Data Science and Artificial Intelligence

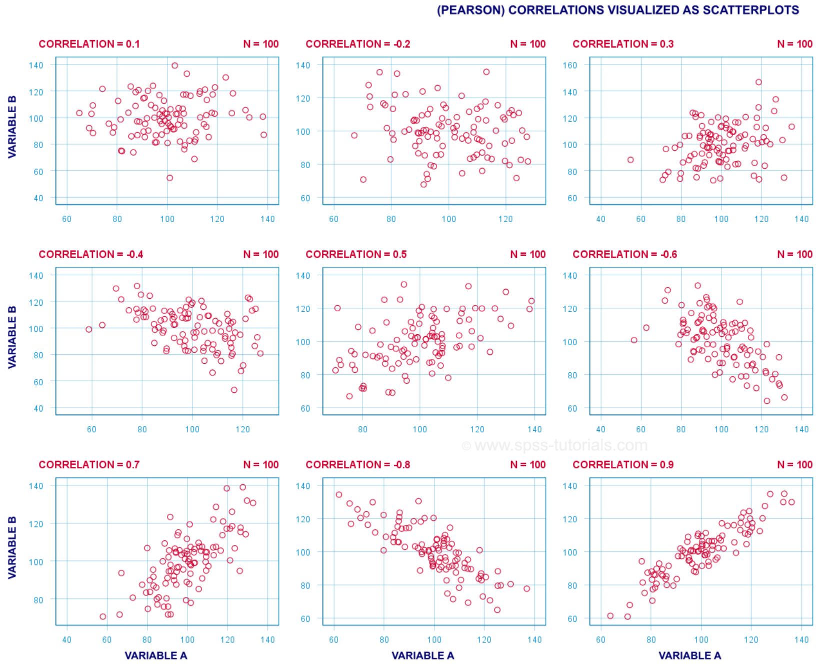
WISSEN
TECHNIK
LEIDENSCHAFT

Correlation

To what extent are two variables related?

Pearson correlation (which we will use): To what extent are two quantitative variables related?

- No statement about **causality**
  - … even if in some literature you read about “independent/explaining variable” and “dependent/outcome” variable!
  - Good practice: Experimental set-up that allows inferring causaility + theory-based explanation of why you would expect causality!



https://www.spss-tutorials.com/spss-correlation-analysis/

ISDS INSTITUTE OF INTERACTIVE SYSTEMS AND DATA SCIENCE

Pearson correlation - formula

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

$\bar{X}, \bar{Y}$ = average of X,Y respectively

Correlation coefficient r describes the strength of the relationship between two variables

-1: Perfect descending linear relationship – if the value of x is higher of a given delta, the value of y will be respectively lower (fixed ratio)

0: Variables aren’t systematically related

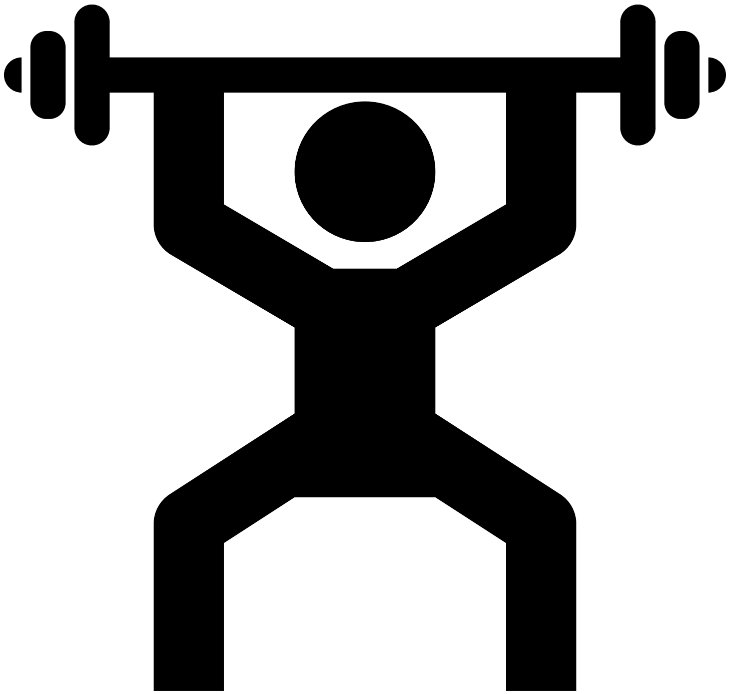
1: Perfect ascending linear relationship – if the value of x is higher of a given delta, the value of y will be respectively higher (fixed ratio)

Looks and sounds familiar? -> Pearson correlation == cosine similarity *IFF* each of two vectors can be understood as representing a (mean-adjusted) variable

Correlation Analysis

- Correlation coefficient *r* describes the strength of the linear relationship between two variables…
- … but in case of a large population, typically only a (hopefully good) sample is available.
  - How large is the likelihood that we are getting a correlation based on our sample even though there isn't really a correlation in the whole population? (null hypothesis)
    - We want this likelihood to be small! Typical threshold values for "small enough" are $p<0.05$, $p<0.01$, $p<0.005$

Exercise 18

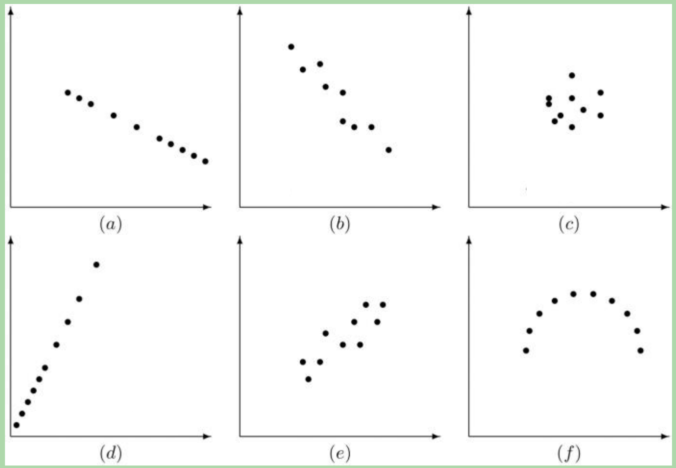


Correlation

Can you make a plausible guess as to the direction of the correlation of the following?

- The weight of an automobile and its fuel consumption?
- The age of an automobile and the distance driven?
- The weight of an automobile and the distance driven?
- The age and weight of a person?
- The number of pages of a book and the expected age of its readers?
- The number of pages of a book and the age of its author?

Is the correlation value in the scatterplots positive, negative, or rather around zero?



Picture from: https://saylordotorg.github.io/text_introductory-statistics/s14-02-the-linear-correlation-coeffic.html

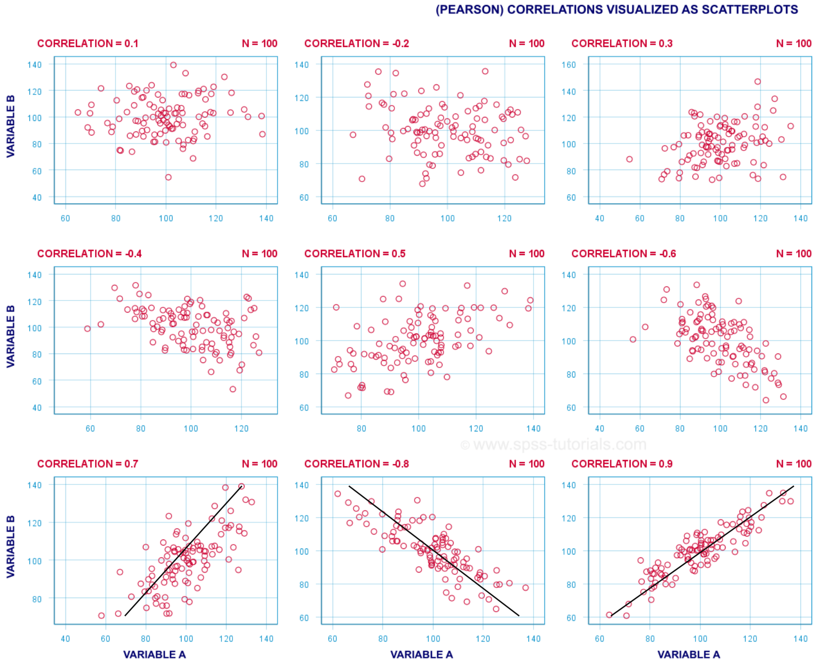
TU Graz

SCIENCE
PASSION
TECHNOLOGY

Linear Regression

ISDS
INSTITUTE OF INTERACTIVE SYSTEMS AND DATA SCIENCE

If we have two correlated variables – let‘s fit the line in! -> **linear regression**



https://www.spss-tutorials.com/spss-correlation-analysis/

Linear regression line via the least squares method

Goal: Identify the line that best fits the linearly correlated data

Regression line: $y = a + bx$

- Underlying assumption: We never know all data, the data that we know are just *training data*.
  - Typical ML approaches to data are taken, such as choosing a part of data as training data, keeping a part of data as test data for evaluating the quality of the model.
- Optimisation criterion: Minimal least squares error –minimal sum of distances (in whichever direction) of points to line.

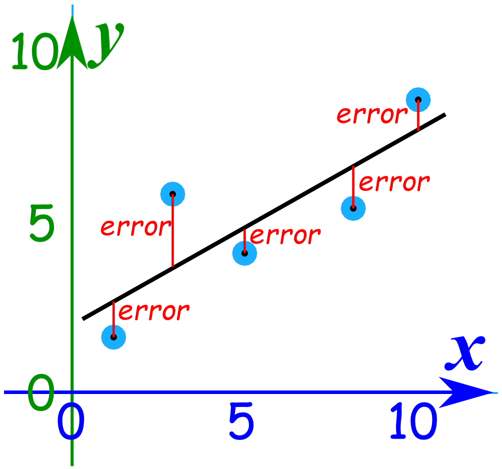
Linear regression line via the least squares method

Regression line: $y = a + bx$

Slope $b = \frac{N \sum xy - \sum x \sum y}{N \sum x^2 - (\sum x)^2}$

OR $b = r\left(\frac{s_y}{s_x}\right)$ with *r* the correlation coefficient

Intercept $a = \frac{\sum y - b \sum x}{N}$



From: https://www.mathsisfun.com/data/least-squares-regression.html

- This method minimizes the sum of the squared errors.
- Error = difference between the estimated y-value (y=a+bx) for a given x value, and the real/measured y-value in the sample data for the same x value.

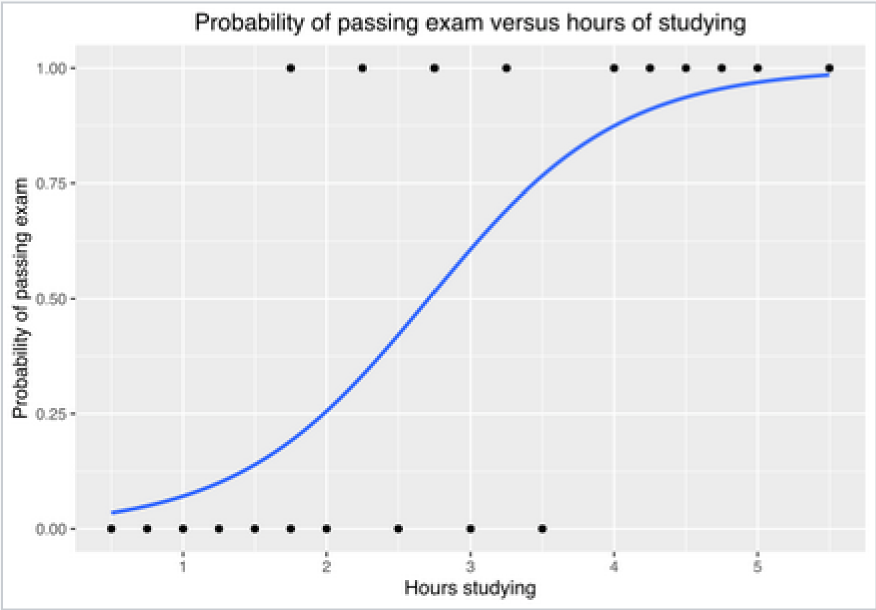
Other types of regression

Non-linear regression – curve fitting: Fitting a non-linear function to data

Logistic regression: Fitting a logarithmic function to continuous independent data and dichotomous outcome data

- Looks familiar?

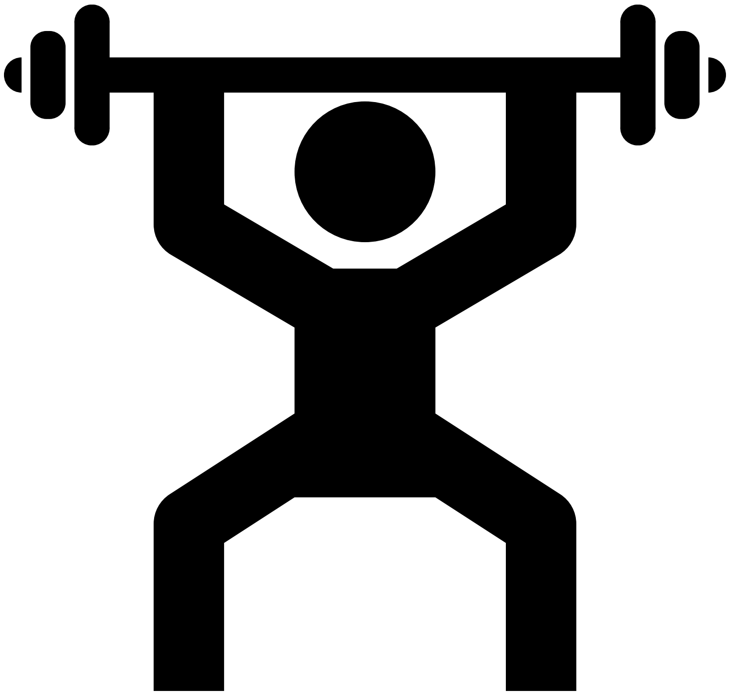
➢ Logistic regression is a method for classification into two classes



Example graph of a logistic regression curve fitted to data. The curve shows the probability of passing an exam (binary dependent variable) versus hours studying (scalar independent variable). See § Example for worked details.

ISDS INSTITUTE OF INTERACTIVE SYSTEMS AND DATA SCIENCE

Exercise 19



Linear Regression

For the following sample data, compute

- The correlation value
- The best fitting regression line via least squares regression

As an exception – do use a calculator, Excel/spreadsheets, R, … to do the calculation!

X	**Y**
6	45
12	47
13	39
17	58
22	68
25	76
27	75
29	74
30	78
32	81

Discussion

WISSEN
TECHNIK
LEIDENSCHAFT

Correlation, linear regression, and machine learning

Prediction:

Estimating the value y_i (y is the dependent, the outcome variable), given x_i (x is the independent, the explaining variable)? – Use the regression line

- What example use cases can you think of for prediction?

Feature engineering:

Sometimes features are desirable that are not highly correlated pairwise

Learning Goals

- Understand correlation
- Be able to compute correlation between two numerical variables
- Understand the computational task of (linear) regression analysis
- Be able to carry out least-squares linear regression
- Understand the computational task of prediction, and regression analysis as a method for prediction.