

Attributes classification

28. oktober 2021 16:58

Spring2020

No.	Attribute description	Abbrev.
x_1	Live birth rate per 1000 population	BirthRt
x_2	Death rate per 1000 population	DeathRt
x_3	Infant deaths per 1000 population under 1 year	InfMort
x_4	Life expectancy at births for males	LExpM
x_5	Life expectancy at births for females	LExpF
x_6	Region encoded as 1, 2, ..., 6	Region
y	Gross National Product, per capita, US\$	GNP

Table 1: Description of the features of the Poverty dataset used in this exam. The dataset consists of population statistics of countries provided by the 1990 United Nations statistical almanacs. x_1, \dots, x_5 respectively provide statistics on birth rates, death rates, infant deaths, and life expectancy by gender and x_6 denotes location of each country in terms of regions such that 1 = Eastern Europe, 2 = South America/Mexico, 3 = Western Europe/US/Canada/Australia/NewZealand/Japan, 4 = Middle East, 5 = Asia and 6 = Africa. The data has been processed such that countries having missing values have been removed. We consider the goal as predicting the gross national product (GNP) pr. capita both as a regression and classification task. For regression tasks, y_r will refer to the continuous value of GNP. For classification tasks the attribute y_b is discrete formed by thresholding y_r at the median value and takes values $y_b = 0$ (corresponding to low GNP level) and $y_b = 1$ (corresponding to a high GNP level). The dataset used has $N = 91$ observations in total.

Question 1. We will consider the Poverty dataset¹ described in Table 1. The dataset consists of 91 countries (observations) and six input attributes x_1, \dots, x_6 as well as the output y_r providing the gross national product pr. capita (denoted GNP). Which one of the following statements regarding the dataset is correct?

- A. All the input attributes x_1, \dots, x_6 are ratio.
- B. One of the six input attributes is nominal.**
- C. All the input attributes x_1, \dots, x_6 are interval.
- D. The output attribute y_r is ordinal.
- E. Don't know.

Solution 1. For the attributes x_1, \dots, x_5 zero means absence of what is being measured and we can naturally

talk about a quantity being say twice as large as another etc. thus these five input attributes are all ratio. x_6 is nominal as this variable categorizes which region each observation belongs to of the six different regions in the dataset. The output y_r is ratio as zero naturally indicates absence of GNP and we again can naturally apply subtraction and addition (required for an interval attribute) but also multiplication (the GNP of one country can be three times larger than of another etc.).

Bayesian Classifier

29. oktober 2021 09:34

Spring 2020

	f_1	f_2	f_3	f_4	f_5
o_1	1	1	1	0	0
o_2	1	1	1	0	0
o_3	1	1	1	0	0
o_4	1	1	1	0	0
o_5	1	1	1	0	0
o_6	0	1	1	0	0
o_7	0	1	0	1	1
o_8	1	1	1	0	0
o_9	1	0	1	0	0
o_{10}	0	0	0	1	1
o_{11}	0	1	0	1	1

Table 5: Binarized version of the Poverty dataset in which the features x_1, \dots, x_5 are binarized. Each of the binarized features f_i are obtained by taking the corresponding feature x_i and letting $f_i = 1$ correspond to a value x_i greater than the median (otherwise $f_i = 0$). As in Table 3 the colors indicate the two classes such that the red observations $\{o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8\}$ belong to class $y_b = 0$ (corresponding to a low GNP). and black observations $\{o_9, o_{10}, o_{11}\}$ belongs to class $y_b = 1$ (corresponding to a high GNP)

What is the probability estimate if $\alpha = 1$?

- A. $p(f_2 = 1, f_3 = 1 | y_b = 1) = \frac{1}{9}$
- B. $p(f_2 = 1, f_3 = 1 | y_b = 1) = \frac{1}{5}$
- C. $p(f_2 = 1, f_3 = 1 | y_b = 1) = \frac{4}{11}$
- D. $p(f_2 = 1, f_3 = 1 | y_b = 1) = \frac{2}{3}$
- E. Don't know.

Solution 16. Of the observations in class $y_b = 1$ zero have simultaneously $f_2 = 1$ and $f_3 = 1$. As this class contains three observations, we see the answer is

$$\frac{0 + \alpha}{3 + 2\alpha} = \frac{1}{5}$$

Therefore, answer B is correct.

Question 16. We again consider the Poverty dataset from Table 1 and the $N = 11$ observations we already encountered in Table 3. The first five features of the dataset is processed to produce five new, binary features such that $f_i = 1$ corresponds to a value x_i greater than the median², and we thereby arrive at the $N \times M = 11 \times 5$ binary matrix in Table 5. We wish to apply a Bayesian classifier to the dataset and as part of this task we have to estimate the probability

$$p(f_2 = 1, f_3 = 1 | y_b = 1).$$

For better numerical stability, we will use robust estimation to obtain the probability by introducing a regularization factor of α such that:

$$p(A|B) = \frac{\{\text{Occurrences matching } A \text{ and } B\} + \alpha}{\{\text{Occurrences matching } B\} + 2\alpha}.$$

Question 17. Consider again the binarized version of the Poverty dataset given in Table 5. We will no longer use robust estimation (i.e., we set $\alpha = 0$) and train a naïve-Bayes classifier in order to predict the class label y_b using only the features f_2 and f_3 . If for an observation we have

$$f_2 = 1, f_3 = 0$$

what is then the probability that the observation has high GNP (i.e., $y_b = 1$) according to a naïve-Bayes classifier trained using only the data in Table 5?

- A. $p_{\text{NB}}(y_b = 1|f_2 = 1, f_3 = 0) = \frac{2}{9}$
- B. $p_{\text{NB}}(y_b = 1|f_2 = 1, f_3 = 0) = \frac{1}{3}$
- C. $p_{\text{NB}}(y_b = 1|f_2 = 1, f_3 = 0) = \frac{2}{5}$
- D. $p_{\text{NB}}(y_b = 1|f_2 = 1, f_3 = 0) = \frac{16}{25}$
- E. Don't know.

Solution 17. To solve this problem, we simply use the general form of the naïve-Bayes approximation and plug in the relevant numbers. We get:

$$\begin{aligned} p_{\text{NB}}(y_b = 1|f_2 = 1, f_3 = 0) &= \\ &\frac{p(f_2 = 1|y = 1)p(f_3 = 0|y = 1)p(y_b = 1)}{\sum_{j=0}^1 p(f_2 = 1|y = j)p(f_3 = 0|y = j)p(y_b = j)} \\ &= \frac{\frac{1}{3} \frac{2}{3} \frac{3}{11}}{\frac{8}{8} \frac{8}{8} \frac{8}{11} + \frac{1}{3} \frac{2}{3} \frac{3}{11}} \\ &= \frac{2}{5}. \end{aligned}$$

Cross-Validation

28. oktober 2021 18:46

Spring 2020

Question 13. Suppose a neural network is trained to predict GNP. As part of training the network, we wish to select between three different model architectures respectively with 5, 10 and 20 hidden units and estimate the generalization error of the optimal choice. In the outer loop we opt for $K_1 = 4$ -fold cross-validation, and in the inner $K_2 = 7$ -fold cross-validation. The time taken to *train* a single model is 20 seconds, and this can be assumed constant for each fold. If the time taken to test a model is 1 second what is then the total time required to complete the 2-level cross-validation procedure?

D. 1848 seconds

E. Don't know.

Solution 13. Let $S = 3$ denote the three different models considered. Going over the 2-level cross-validation algorithm we see the total number of models to be *trained* is:

$$K_1(K_2S + 1) = 88$$

Multiplying by the time taken to train a single model we obtain a total training time of 1760 seconds.

As every model we use to train is also used for testing a dataset the number of times we test a model is:

$$K_1(K_2S + 1) = 88$$

As each of these take 1 second we obtain in total $1760+88=1848$ seconds.

Decision Boundary

28. oktober 2021 18:58

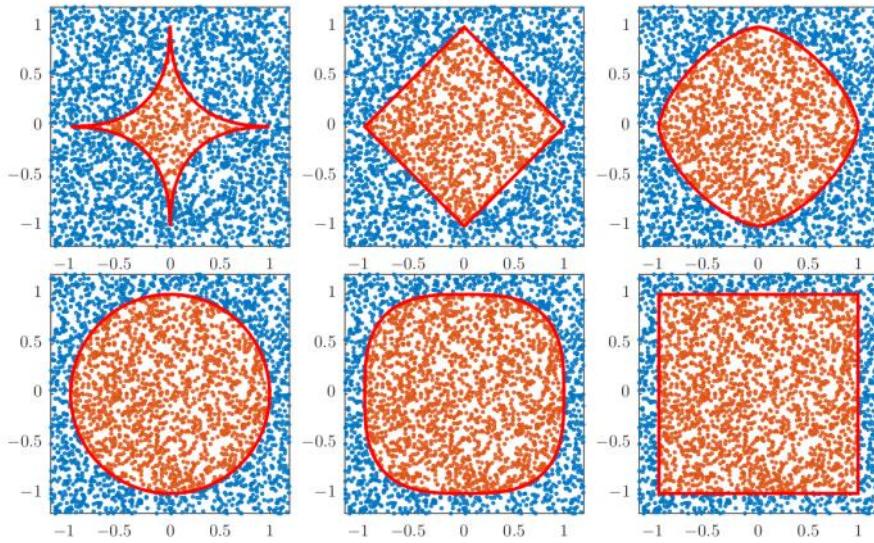


Fig. 4.2. Illustration of the p -distance for various values of p . A point \mathbf{x} is colored red if its p distance to the center $\mathbf{0}$ is less than 1, $d_p(\mathbf{x}, \mathbf{0}) \leq 1$. Top row: $p = \frac{1}{2}, 1, \frac{3}{2}$ and bottom row: $p = 2, 3, \infty$. The red line is the decision boundary $d_p(\mathbf{x}, \mathbf{0}) = 1$. Increasing p corresponds to “inflating” the red region.

Spring 2020

Question 14. We will fit a decision tree in order to determine based on the features x_1 and x_2 if a country has a relatively low or high GNP. In the top panel of Figure 7 is given the fitted decision tree and in the bottom panel is given four different decision boundaries in which one of the four decision boundaries corresponds to the boundaries generated by the decision tree given in the top panel.

Which one of the the four decision boundaries corresponds to the decision boundaries of the illustrated classification tree?

- A. Decision boundary of Classifier 1
- B. Decision boundary of Classifier 2
- C. **Decision boundary of Classifier 3**
- D. Decision boundary of Classifier 4
- E. Don't know.

Solution 14. The decision tree includes four decisions two based on x_1 and two based on x_2 . As such the decision boundaries must have two horizontal and vertical lines which only holds for Classifier 3.

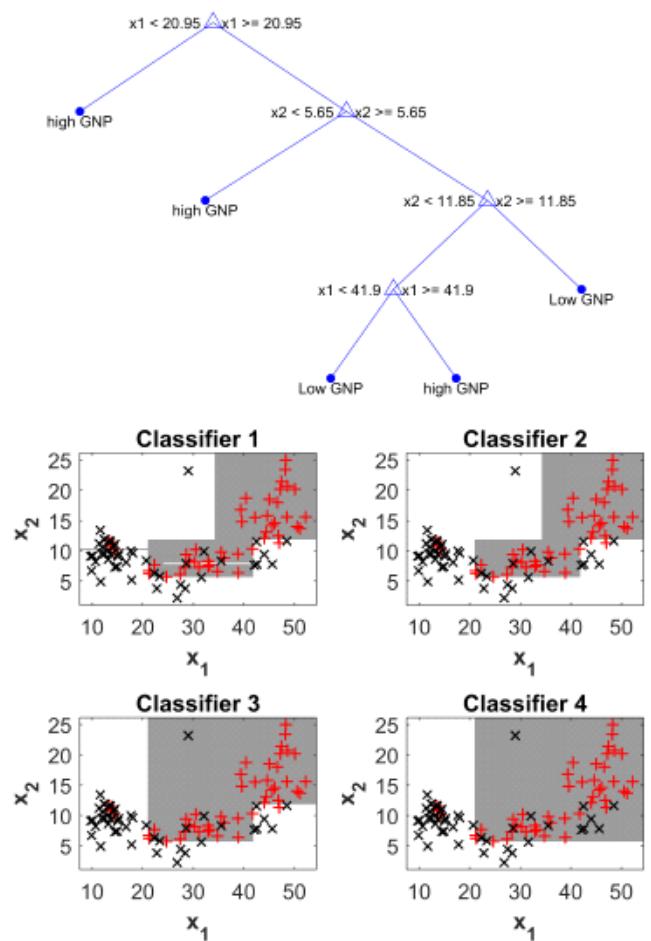


Figure 7: Top panel, a decision tree fitted to x_1 and x_2

Figure 7: Top panel, a decision tree fitted to x_1 and x_2 of the Poverty data in order to predict whether a country has relatively low or high GNP. Bottom panel, decision boundaries for four different decision trees in which gray regions correspond to regions predicted having low GNP ($y_b = 0$) and white regions to predictions having high GNP ($y_b = 1$). One of the four decision boundaries corresponds to the decision boundary of the classification tree given in the top panel.

Decision Tree Classifier

29. oktober 2021 10:01

Spring 2020

Question 18. We will develop a decision tree classifier in order to determine whether a country is relatively poor ($y_b = 0$) or rich ($y_b = 1$) considering only the data in Table 5. During the training of the classifier the purity gain using feature f_1 corresponding to thresholding x_1 by the median value is evaluated by Hunt's algorithm as the first decision in the tree (i.e., as decision for the root of the tree). As impurity measure we will use Gini which is given by $I(v) = 1 - \sum_c p(c|v)^2$.

What is the purity gain Δ of this considered split?

A. $\Delta = 0.000$

B. $\Delta = 0.059$

C. $\Delta = 0.125$

D. $\Delta = 0.148$

E. Don't know.

Solution 18. The purity gain is given by

$$\Delta = I(r) - \sum_{k=1}^K \frac{N(v_k)}{N} I(v_k),$$

where

$$I(v) = 1 - \sum_c p(c|v)^2.$$

Evaluating the purity gain for the split we have:

$$\begin{aligned} \Delta &= (1 - ((8/11)^2 + (3/11)^2)) \\ &\quad - [\frac{4}{11}(1 - ((2/4)^2 + (2/4)^2)) \\ &\quad + \frac{7}{11}(1 - ((6/7)^2 + (1/7)^2))] \\ &= 0.059 \end{aligned}$$

	f_1	f_2	f_3	f_4	f_5
o_1	1	1	1	0	0
o_2	1	1	1	0	0
o_3	1	1	1	0	0
o_4	1	1	1	0	0
o_5	1	1	1	0	0
o_6	0	1	1	0	0
o_7	0	1	0	1	1
o_8	1	1	1	0	0
o_9	1	0	1	0	0
o_{10}	0	0	0	1	1
o_{11}	0	1	0	1	1

Table 5: Binarized version of the Poverty dataset in which the features x_1, \dots, x_5 are binarized. Each of the binarized features f_i are obtained by taking the corresponding feature x_i and letting $f_i = 1$ correspond to a value x_i greater than the median (otherwise $f_i = 0$). As in Table 3 the colors indicate the two classes such that the red observations $\{o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8\}$ belong to class $y_b = 0$ (corresponding to a low GNP), and black observations $\{o_9, o_{10}, o_{11}\}$ belongs to class $y_b = 1$ (corresponding to a high GNP)

Expectation

29. oktober 2021 10:27

Spring 2020

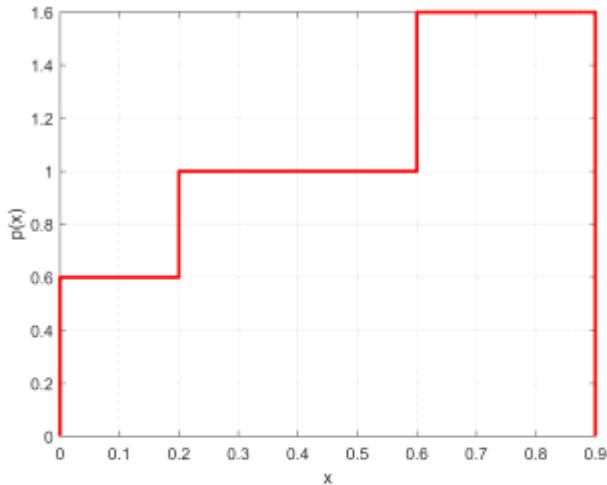


Figure 10: Probability density function for a random variable x . Outside the region from 0 to 0.9 the density function is zero.

Question 24. In Figure 10 is given the density function $p(x)$ of a random variable x . What is the expected value of x , i.e. $\mathbb{E}[x]$?

- A. 0.450
- B. 0.532**
- C. 0.600
- D. 1.000
- E. Don't know.

Solution 24.

$$\begin{aligned}\mathbb{E}[x] &= \int xp(x) = \int_0^{0.2} x \cdot 0.6 + \int_{0.2}^{0.6} x \cdot 1 + \int_{0.6}^{0.9} x \cdot 1.6 \\ &= 0.6 \cdot 0.5 \cdot (0.2^2 - 0^2) + 1 \cdot 0.5 \cdot (0.6^2 - 0.2^2) \\ &\quad + 1.6 \cdot 0.5 \cdot (0.9^2 - 0.6^2) = 0.532,\end{aligned}$$

where we have used that $\int_a^b x dx = 0.5 \cdot (b^2 - a^2)$

Forward-Backward

28. oktober 2021 18:40

Spring 2020

Question 12. Consider again the Poverty dataset in Table 1. We would like to predict GNP using a least squares linear regression model, and since we would like the model to be as interpretable as possible we will use variable selection to obtain a parsimonious model. We limit ourselves to the five features x_1, x_2, x_3, x_4 and x_5 . In Table 4 we have pre-computed the estimated training and test errors for all combinations of the five attributes. Which one of the following statements is correct?

- A. Forward selection will select attributes x_3 .
- B. Forward selection will select attributes x_1, x_3, x_4, x_5 .
- C. Forward selection will select attributes x_1, x_2, x_4 .
- D. Backward selection will select attributes x_1, x_4 .

Feature(s)	Training RMSE	Test RMSE
none	1.429	2.02
x_1	0.755	1.662
x_2	1.421	1.977
x_3	0.636	1.628
x_4	0.847	1.636
x_5	0.773	1.702
x_1, x_2	0.640	1.706
x_1, x_3	0.636	1.638
x_2, x_3	0.401	1.912
x_1, x_4	0.745	1.602
x_2, x_4	0.565	1.799
x_3, x_4	0.587	1.890
x_1, x_5	0.728	1.647
x_2, x_5	0.449	1.767
x_3, x_5	0.613	1.824
x_4, x_5	0.733	2.155
x_1, x_2, x_3	0.380	2.135
x_1, x_2, x_4	0.541	1.696
x_1, x_3, x_4	0.586	1.914
x_2, x_3, x_4	0.399	1.954
x_1, x_2, x_5	0.448	1.779
x_1, x_3, x_5	0.613	1.831
x_2, x_3, x_5	0.396	1.828
x_1, x_4, x_5	0.702	2.022
x_2, x_4, x_5	0.407	2.087
x_3, x_4, x_5	0.582	1.901
x_1, x_2, x_3, x_4	0.379	2.168
x_1, x_2, x_3, x_5	0.369	1.988
x_1, x_2, x_4, x_5	0.400	2.138
x_1, x_3, x_4, x_5	0.580	1.927
x_2, x_3, x_4, x_5	0.359	1.935
x_1, x_2, x_3, x_4, x_5	0.315	2.030

A. Forward selection will select attributes x_3 .

B. Forward selection will select attributes x_1, x_3, x_4, x_5 .

C. Forward selection will select attributes x_1, x_2, x_4 .

D. Backward selection will select attributes x_1, x_4 .

E. Don't know.

Solution 12. The correct answer is A. To solve this problem, it suffices to show which variables will be selected by forward or backward selection. First note that in variable selection, we only need to concern ourselves with the *test* error, as the training error should trivially drop when more variables are introduced and is furthermore not what we ultimately care about.

Forward selection: The method is initialized with the empty set $\{\}$ having an error of 2.02.

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}$. Since the lowest error of the available sets is 1.628, which is lower than 2.02, we update the current selected variables to $\{x_3\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_3\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}, \{x_1, x_4\}, \{x_2, x_4\}, \{x_3, x_4\}, \{x_1, x_5\}, \{x_2, x_5\}, \{x_3, x_5\}, \{x_4, x_5\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates at $\{x_3\}$.

Backward selection: The method is initialized with the set $\{x_1, x_2, x_3, x_4, x_5\}$ having an error of 2.03.

Logistic-Regression

28. oktober 2021 18:30

Spring2020

Question 9. A logistic regression model is trained to distinguish between the two classes $y_b \in \{0, 1\}$, i.e., relatively low GNP (negative class) vs. relative high GNP (positive class). The model is trained using all observations except the 11 observations given in Table 3 that are used for testing the model (i.e., using the hold-out method). The features x_1, \dots, x_5 are standardized (mean subtracted and each feature divided by its standard deviation). The feature x_6 is transformed using one-out-of-K coding and the last region removed to generate the new features c_1, c_2, c_3, c_4, c_5 that are included in the regression to produce the class-probability \hat{y} defined by the trained model:

$$\begin{aligned}\hat{y} = & \sigma(1.41 + 0.76x_1 + 1.76x_2 - 0.32x_3 - 0.96x_4 + 6.64x_5 \\ & - 5.13c_1 - 2.06c_2 + 96.73c_3 + 1.03c_4 - 2.74c_5).\end{aligned}$$

We will predict the estimated output of the sixth of the eleven test observations given by:

$$\mathbf{x}_6 = [-0.06 \ -0.28 \ 0.43 \ -0.30 \ -0.36 \ 0 \ 0 \ 0 \ 0 \ 1]$$

Which one of the following statements is correct?

- A. According to the estimated model an increase in a country's birth rate will increase the probability that the country is rich.
- B. The probability observation \mathbf{x}_6 belongs to class $y = 1$ is less than 1 %.
- C. The attribute *Region* has very little influence on whether a country is poor or rich.
- D. As the weight for x_1 and x_3 have opposing signs we can conclude the two features are negatively correlated.

- A. According to the estimated model an increase in a country's birth rate will increase the probability that the country is rich.**
- B. The probability observation \mathbf{x}_6 belongs to class $y = 1$ is less than 1 %.
- C. The attribute *Region* has very little influence on whether a country is poor or rich.
- D. As the weight for x_1 and x_3 have opposing signs we can conclude the two features are negatively correlated.
- E. Don't know.

Solution 9. As the coefficient in front of x_1 is positive this implies that increasing x_1 will according to the model increase the probability of being in the positive class (i.e., a rich country), thus this is a correct statement. The estimated output for \mathbf{x}_6 is

$$\begin{aligned}\hat{y} = & \sigma(1.41 + (0.76 \cdot -0.06) + (1.76 \cdot -0.28) \\ & - (0.32 \cdot 0.43) - (0.96 \cdot -0.30) + (6.64 \cdot -0.36) \\ & - (2.74 \cdot 1.00)) \\ = & \frac{1}{1+\exp(-(1.41+(0.76\cdot-0.06)+(1.76\cdot-0.28)-(0.32\cdot0.43)-(0.96\cdot-0.36)-\\ & - (2.74 \cdot 1.00)))} \\ = & 1.62\%.\end{aligned}$$

From the model it is further observed that *Region* has a very strong influence on the estimated output - in particular $c_3 = 1$ corresponding to the country being in the Western Europe/US/Canada/Australia/NewZealand/Japan region strongly influences that the country will be given a high probability of being in the positive class (i.e., rich), i.e. the coefficient in front of c_3 is positive and very large with magnitude of 96.73. Notably, we can not use the sign of the estimated weights to deduce anything about feature correlation.

KNN-Classifier

28. oktober 2021 17:58

Spring2020

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}	o_{11}
o_1	0.0	1.7	1.4	0.4	2.2	3.7	5.2	0.2	4.3	6.8	6.0
o_2	1.7	0.0	1.0	2.0	1.3	2.6	4.5	1.8	3.2	5.9	5.2
o_3	1.4	1.0	0.0	1.7	0.9	2.4	4.1	1.5	3.0	5.5	4.8
o_4	0.4	2.0	1.7	0.0	2.6	4.0	5.5	0.3	4.6	7.1	6.3
o_5	2.2	1.3	0.9	2.6	0.0	1.7	3.4	2.4	2.1	4.8	4.1
o_6	3.7	2.6	2.4	4.0	1.7	0.0	2.0	3.8	1.6	3.3	2.7
o_7	5.2	4.5	4.1	5.5	3.4	2.0	0.0	5.4	2.5	1.6	0.9
o_8	0.2	1.8	1.5	0.3	2.4	3.8	5.4	0.0	4.4	6.9	6.1
o_9	4.3	3.2	3.0	4.6	2.1	1.6	2.5	4.4	0.0	3.4	2.9
o_{10}	6.8	5.9	5.5	7.1	4.8	3.3	1.6	6.9	3.4	0.0	1.0
o_{11}	6.0	5.2	4.8	6.3	4.1	2.7	0.9	6.1	2.9	1.0	0.0

Table 3: The pairwise Euclidian distances, $d(o_i, o_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2 = \sqrt{\sum_{k=1}^M (x_{ik} - x_{jk})^2}$ between 11 observations from the Poverty dataset based on x_1, \dots, x_5 . Each observation o_i corresponds to a row of the data matrix \mathbf{X} of Table 1 (excluding x_6). The colors indicate classes such that the red observations $\{o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8\}$ belong to class $y_b = 0$ (corresponding to a low GNP level), and the black observations $\{o_9, o_{10}, o_{11}\}$ belongs to class $y_b = 1$ (corresponding to a relatively high GNP).

Question 6. To examine if observation o_3 may be an outlier we will calculate the average relative density using the Euclidean distance based on the observations given in Table 3 only. We recall that the KNN density and average relative density (ard) for the observation \mathbf{x}_i are given by:

$$\text{density}_{\mathbf{X}_{\setminus i}}(\mathbf{x}_i, K) = \frac{1}{K \sum_{\mathbf{x}' \in N_{\mathbf{X}_{\setminus i}}(\mathbf{x}_i, K)} d(\mathbf{x}_i, \mathbf{x}')},$$

$$\text{ard}_{\mathbf{X}}(\mathbf{x}_i, K) = \frac{\text{density}_{\mathbf{X}_{\setminus i}}(\mathbf{x}_i, K)}{\frac{1}{K} \sum_{\mathbf{x}_j \in N_{\mathbf{X}_{\setminus i}}(\mathbf{x}_i, K)} \text{density}_{\mathbf{X}_{\setminus j}}(\mathbf{x}_j, K)},$$

where $N_{\mathbf{X}_{\setminus i}}(\mathbf{x}_i, K)$ is the set of K nearest neighbors of observation \mathbf{x}_i excluding the i 'th observation, and $\text{ard}_{\mathbf{X}}(\mathbf{x}_i, K)$ is the average relative density of \mathbf{x}_i using K nearest neighbors. What is the average relative density for observation o_3 for $K = 2$ nearest neighbors?

- A. 0.59
- B. 1.00
- C. 1.05

D. 1.18

Question 6 Consider again the 11 observations in

D. 1.18

- E. Don't know.

Solution 6.

To solve the problem, first observe the $k = 2$ neighborhood of o_3 and density is:

$$N_{\mathbf{X}_{\setminus 3}}(\mathbf{x}_3) = \{o_5, o_2\}, \quad \text{density}_{\mathbf{X}_{\setminus 3}}(\mathbf{x}_3) = 1.053$$

For each element in the above neighborhood we can then compute their $K = 2$ -neighborhoods and densities to be:

$$N_{\mathbf{X}_{\setminus 5}}(\mathbf{x}_5) = \{o_3, o_2\}, \quad N_{\mathbf{X}_{\setminus 2}}(\mathbf{x}_2) = \{o_3, o_5\}$$

and

$$\text{density}_{\mathbf{X}_{\setminus 5}}(\mathbf{x}_5) = 0.909, \quad \text{density}_{\mathbf{X}_{\setminus 2}}(\mathbf{x}_2) = 0.870.$$

From these, the ARD can be computed by plugging in the values in the formula given in the problem and we obtain $1.053 / (0.5 \cdot (0.909 + 0.870))$.

Question 8. Consider again the 11 observations in Table 3. We will use a one-nearest neighbor classifier to classify the observations. What will be the error rate of the KNN classifier when considering a leave-one-out cross-validation strategy to quantify performance?

- A. 3/11
- B. 4/11**
- C. 5/11
- D. 6/11
- E. Don't know.

Solution 8. observation o_1 has o_8 as nearest neighbor and correctly classified.

observation o_2 has o_3 as nearest neighbor and correctly classified.

observation o_3 has o_5 as nearest neighbor and correctly classified.

observation o_4 has o_8 as nearest neighbor and correctly classified.

observation o_5 has o_3 as nearest neighbor and correctly classified.

observation o_6 has o_9 as nearest neighbor and incorrectly classified.

observation o_7 has o_{11} as nearest neighbor and incorrectly classified.

observation o_8 has o_1 as nearest neighbor and correctly classified.

observation o_9 has o_6 as nearest neighbor and incorrectly classified.

observation o_{10} has o_{11} as nearest neighbor and correctly classified.

observation o_{11} has o_7 as nearest neighbor and incorrectly classified.

4 out of 11 observations are thus incorrectly classified.

Spring 2020

Question 19. We again consider the dataset in Table 5. This time it is decided to group the observations according to f_2 corresponding to having a relatively low or high death rate (DeathRt). We will thereby cluster the observations such that $f_2 = 0$ corresponds to observations in the first cluster and $f_2 = 1$ corresponds to observations in the second cluster³. We wish to compare this clustering to that corresponding to the true class labels $y_b = 0$ and $y_b = 1$ according to the Jaccard index. Recall that the Jaccard index is given by $J = \frac{S}{N(N-1)/2-D}$ where S denotes the number of pairs of observations assigned to the same cluster that are in the same class, and D denotes the number of pairs of observations assigned to different clusters that are also in different classes. What is the value of J between the true class labels given by $y_b = 0$ and $y_b = 1$ and the two extracted clusters given by $f_2 = 0$ and $f_2 = 1$?

- A. $J = 0.0909$
- B. $J = 0.5273$
- C. $J = 0.7436$
- D. $J = 0.7838$
- E. Don't know.

Solution 19. for the eleven observations we have the true labels are $\mathbf{y}_b = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1]^\top$ and the clustering is given as $\mathbf{z} = [2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 2]^\top$. We now consider all object pairs in same cluster having same class, for the 9 observations in cluster 2 we have that 8 are in the same class giving $8(8-1)/2$ pairs and for the two observations in cluster 1 we have that both are in the same class giving $2(2-1)/2$ pairs. Thus, $S = 8(8-1)/2 + 2(2-1)/2 = 29$. For the 9 observations in cluster 2 we have that 8 have $y_b = 0$ whereas both the two observations in cluster 2 have $y_b = 1$. Thus, $D = 8 \cdot 2 = 16$. As a result, we have that $J = \frac{S}{N(N-1)/2-D} = 29/(11 \cdot (11 - 1)/2 - 16) = 0.7436$,

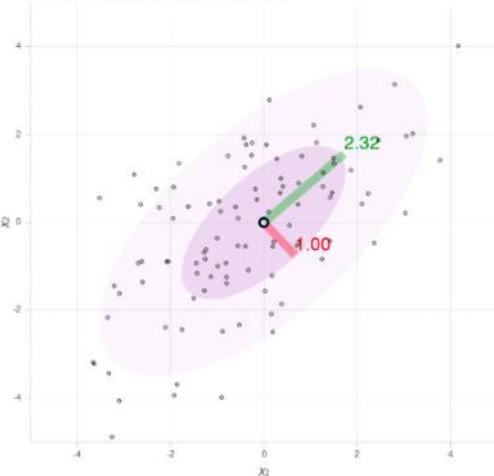
Mahalanobis distance

29. oktober 2021 10:26

To calculate the Mahalanobis distance, you need a covariance matrix. This matrix is given by the covariance of all the blue observations. On the plot, you can see that the correlation between x_1 and x_2 for those blue points is positive, i.e. when x_1 increases, x_2 increases too. The off diagonal elements of the covariance matrix are the covariance between x_1 and x_2 which will be positive (same sign as the correlation).

Therefore, what is meant by "the covariance matrix will have a shape in the direction of the green circles and blue plusses" is that if you draw an ellipse around the distribution, its major axis will be a line following the circles and plusses.

Multivariate normal distribution



Tap: select mean or direction
Double tap: de-select current selection
Move mouse: move current selection if any

$$\mu = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 3.42 & 2.19 \\ 2.19 & 2.99 \end{bmatrix}$$

Here is an image from a simulation done with <http://www2.imm.dtu.dk/courses/02450/DemoNormal.html> that approximates the exercise conditions. Each point on the edge of the ellipse has the same Mahalanobis distance from the centre.

EIGENVALUE DECOMPOSITION OF THE COVARIANCE MATRIX

The covariance matrix can be thought of it in terms of its eigendecomposition (or spectral decomposition):

$$\Sigma = Q\Lambda Q^{-1}$$

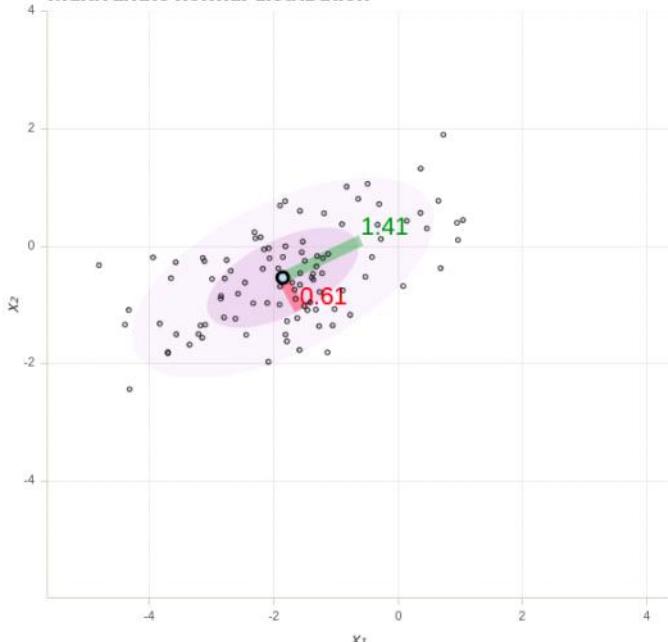
where Q is a square $d \times d$ matrix whose columns are the eigenvectors of the covariance matrix, and Λ is a diagonal matrix with diagonal elements corresponding to the eigenvalues of the covariance matrix. With this decomposition, the eigenvectors define the directions that the multivariate normal distribution "varies in", and the eigenvalues determines the degree of variation in that direction. Go the next slide to see what this means in terms of the plot.

direction changes the eigenvectors (Q). Notice that the Λ contains the lengths of the vectors squared; the length of the vectors define the standard deviation in the given direction, and so the squared value is the variance in the direction. The ellipses then define one standard deviation and two standard deviations away from the mean of the distribution.

PAF

For th
By ch
chang
blue c
plot.
The c
direct
eigenv
char
direct
conta
vecto
so the
ellips
devia

Multivariate normal distribution



Tap: select mean or direction
Double tap: de-select current selection
Move mouse: move current selection if any

$$\mu = \begin{bmatrix} -1.85 \\ 0.61 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1.66 & 0.65 \\ 0.65 & 0.70 \end{bmatrix}$$

ellipses then define one standard deviation and two standard deviations away from the mean of the distribution.

The diagonal elements in Σ signifies variation for the corresponding attribute. Similarly, the off-diagonals determine the amount of covariance between the two attributes.

Move mouse: move current selection if any

$$\mu = \begin{bmatrix} -1.85 \\ -0.54 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1.66 & 0.65 \\ 0.65 & 0.70 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} \textcolor{red}{0.37} & 0.00 \\ 0.00 & \textcolor{green}{1.98} \end{bmatrix} = \begin{bmatrix} \textcolor{red}{0.61^2} & 0.00 \\ 0.00 & \textcolor{green}{1.41^2} \end{bmatrix} \quad Q = \begin{bmatrix} \textcolor{red}{0.45} & \textcolor{green}{0.89} \\ -\textcolor{red}{0.89} & \textcolor{green}{0.45} \end{bmatrix}$$

Spring 2020

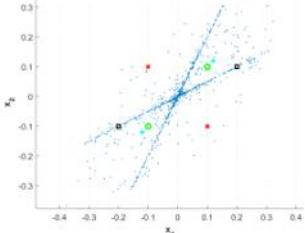


Figure 9: A dataset of 1000 observations given by the blue dots. In the plot is also given the location of two red crosses, two green circles, two cyan pluses and two black squares.

Question 23.

Consider the dataset given in Figure 9. We will consider the Mahalanobis distance using the empirical covariance matrix estimated based on the 1000 blue observations. Which one of the following statements is correct?

- A. The Mahalanobis distance between the two green circles is smaller than the Mahalanobis distance between the two black squares.
- B. The Mahalanobis distance between the two red crosses is the same as the Mahalanobis distance between the two green circles.
- C. The Mahalanobis distance between the two black squares is smaller than the Mahalanobis distance between the two cyan pluses.
- D. The empirical covariance matrix estimated based on the blue observations has at least one element that is negative.
- E. Don't know

Solution 23. As the correlation between x_1 and x_2 is positive the covariance matrix only has positive elements. The covariance matrix will have a shape in the direction of the green circles and blue pluses and therefore these pairs of observations will have relatively short Mahalanobis distance between each other, when compared to the other pairs of observations. Thus, the Mahalanobis distance between the two green circles is smaller than the Mahalanobis distance between the two black squares.

Multivariate gauss

28. oktober 2021 17:44

Spring2020

Question 5. In Figure 3 a Gaussian Mixture Model (GMM) is fitted to the standardized data projected onto the first two principal component directions using three mixture components (i.e., $K = 3$ clusters). Recall that the multivariate Gaussian distribution is given by:

$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Which one of the following GMM densities corresponds to the fitted density given in Figure 3?

A.

$$\begin{aligned} p(\mathbf{x}) &= 0.1425 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 3.3884 \\ 0.7424 \end{bmatrix}, \begin{bmatrix} 0.1695 & 0.0665 \\ 0.0665 & 0.1104 \end{bmatrix}) \\ &+ 0.3235 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} -1.9482 \\ 0.6132 \end{bmatrix}, \begin{bmatrix} 1.2137 & -0.0703 \\ -0.0703 & 0.3773 \end{bmatrix}) \\ &+ 0.5340 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 0.2756 \\ -0.5696 \end{bmatrix}, \begin{bmatrix} 2.0700 & 0.1876 \\ 0.1876 & 0.1037 \end{bmatrix}) \end{aligned}$$

B.

$$\begin{aligned} p(\mathbf{x}) &= 0.1425 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} -1.9482 \\ 0.6132 \end{bmatrix}, \begin{bmatrix} 0.1695 & 0.0665 \\ 0.0665 & 0.1104 \end{bmatrix}) \\ &+ 0.3235 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 3.3884 \\ 0.7424 \end{bmatrix}, \begin{bmatrix} 1.2137 & -0.0703 \\ -0.0703 & 0.3773 \end{bmatrix}) \\ &+ 0.5340 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 0.2756 \\ -0.5696 \end{bmatrix}, \begin{bmatrix} 2.0700 & 0.1876 \\ 0.1876 & 0.1037 \end{bmatrix}) \end{aligned}$$

C.

$$\begin{aligned} p(\mathbf{x}) &= 0.3235 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} -1.9482 \\ 0.6132 \end{bmatrix}, \begin{bmatrix} 0.1695 & 0.0665 \\ 0.0665 & 0.1104 \end{bmatrix}) \\ &+ 0.1425 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 3.3884 \\ 0.7424 \end{bmatrix}, \begin{bmatrix} 2.0700 & 0.1876 \\ 0.1876 & 0.1037 \end{bmatrix}) \\ &+ 0.5340 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 0.2756 \\ -0.5696 \end{bmatrix}, \begin{bmatrix} 1.2137 & -0.0703 \\ -0.0703 & 0.3773 \end{bmatrix}) \end{aligned}$$

D.

$$\begin{aligned} p(\mathbf{x}) &= 0.3235 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} -1.9482 \\ 0.6132 \end{bmatrix}, \begin{bmatrix} 0.1695 & 0.0665 \\ 0.0665 & 0.1104 \end{bmatrix}) \\ &+ 0.1425 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 3.3884 \\ 0.7424 \end{bmatrix}, \begin{bmatrix} 1.2137 & -0.0703 \\ -0.0703 & 0.3773 \end{bmatrix}) \\ &+ 0.5340 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 0.2756 \\ -0.5696 \end{bmatrix}, \begin{bmatrix} 2.0700 & 0.1876 \\ 0.1876 & 0.1037 \end{bmatrix}) \end{aligned}$$

E. Don't know.

Solution 5. Inspecting the GMM density we observe that the cluster located at $\begin{bmatrix} 3.3884 \\ 0.7424 \end{bmatrix}$ will have the lowest mixing proportion as only few observations belong to this cluster. Furthermore, the cluster located at $\begin{bmatrix} -1.9482 \\ 0.6132 \end{bmatrix}$ clearly has positive covariance between PCA1 and PCA2 and much smaller variance

(i.e., 0.1695) in the PCA1 direction when compared to the other cluster located at $\begin{bmatrix} 0.2756 \\ -0.5696 \end{bmatrix}$ having high variance (i.e., 2.0700) also having positive covariance. This only holds for answer option D.

PCA

28. oktober 2021 17:18

Spring2020

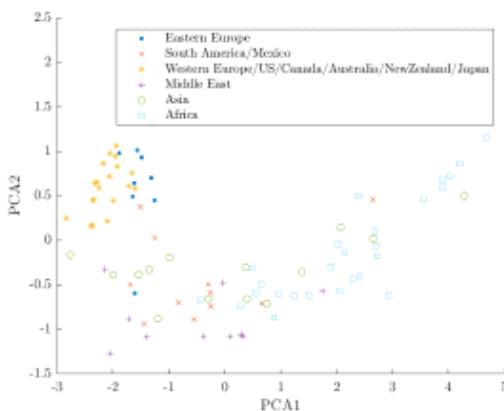


Figure 2: The Poverty data projected onto the first two principal component directions with each observation labelled according to the region it belongs to (given by x_6).

Question 3. A Principal Component Analysis (PCA) is carried out on the Poverty dataset in Table 1 based on the attributes x_1, x_2, x_3, x_4, x_5 .

The data is standardized by (i) subtracting the mean and (ii) dividing each column by its standard deviation to obtain the standardized data matrix $\tilde{\mathbf{X}}$. A singular value decomposition is then carried out on the standardized data matrix to obtain the decomposition $\mathbf{U}\mathbf{S}\mathbf{V}^T = \tilde{\mathbf{X}}$

$$\mathbf{V} = \begin{bmatrix} 0.43 & -0.5 & 0.7 & -0.25 & -0.07 \\ 0.38 & 0.85 & 0.3 & -0.2 & 0.03 \\ 0.46 & -0.13 & -0.61 & -0.61 & -0.15 \\ -0.48 & -0.0 & 0.13 & -0.63 & 0.6 \\ -0.48 & 0.1 & 0.16 & -0.36 & -0.78 \end{bmatrix} \quad (1)$$

$$\mathbf{S} = \begin{bmatrix} 19.64 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 6.87 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 3.26 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 2.30 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.12 \end{bmatrix}.$$

Which one of the following statements is true?

- A. The variance explained by the first four principal components is greater than 99 %.
- B. The variance explained by the last four principal components is greater than 15 %.
- C. The variance explained by the first two principal components is greater than 97 %.
- D. The variance explained by the first principal component is greater than 90 %.
- E. Don't know.

Solution 3. The correct answer is A. To see this, recall that the variance explained by a given component k of the PCA is given by

$$\frac{\sigma_k^2}{\sum_{j=1}^M \sigma_j^2}$$

where M is the number of attributes in the dataset being analyzed. The values of σ_k can be read off as entry $\sigma_k = S_{kk}$ where \mathbf{S} is the diagonal matrix of the SVD computed above. We therefore find the variance explained by the first four components is:

$$\text{Var.Expl.} = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2} = \frac{19.64^2 + 6.87^2 + 3.26^2 + 2.30^2}{19.64^2 + 6.87^2 + 3.26^2 + 2.30^2 + 1.12^2} = 0.9972.$$

Question 4. Consider again the PCA analysis of the Poverty dataset, in particular the SVD decomposition of $\tilde{\mathbf{X}}$ in Equation (1). In Figure 2 is given the data projected onto the first two principal components and each observation labelled according to the region it belongs to. Which one of the following statements is true?

- A. An observation from Africa will typically have a relatively high value of BirthRt, a high value of DeathRt, a high value of InfMort, a low value of LExpM and a low value of LExpF as observed from the projection onto principal component number 1.
- B. An observation from Western Europe/US/Canada/Australia/New Zealand/Japan will typically have a relatively high value of BirthRt, a low value of DeathRt, a high value of InfMort, and a low value of LExpF as observed from the projection onto principal component number 2.
- C. As observed from the projection onto principal component number 1 observations from Eastern Europe will typically have a relatively low value of BirthRt, a high value of DeathRt, a low value of InfMort, a high value of LExpM whereas LExpF will have almost no influence (the coefficient is only -0.07).
- D. As can be seen from the plot of the first and second principal components there is a negative correlation between the observations projected onto PC1 and PC2.
- E. Don't know.

Solution 4. The correct answer is A. Focusing on the correct answer, note the projection onto principal component \mathbf{v}_1 (i.e. column one of \mathbf{V}) is

$$b_1 = \mathbf{x}^\top \mathbf{v}_1 = [x_1 \ x_2 \ x_3 \ x_4 \ x_5] \begin{bmatrix} 0.43 \\ 0.38 \\ 0.46 \\ -0.48 \\ -0.48 \end{bmatrix}$$

for this projection to be (relatively large) and positive which is the case for observations coming from Africa, this occurs if x_1, x_2, x_3, x_4, x_5 has large magnitude and the sign convention given in option A. As the data projected onto PC1 and PC2 is given by $\tilde{\mathbf{X}}\mathbf{v}_1$ and $\tilde{\mathbf{X}}\mathbf{v}_2$

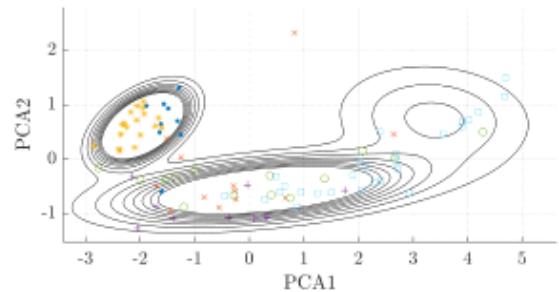


Figure 3: A GMM with $K=3$ clusters fitted to the poverty data projected onto the first two principal component directions. Each observation is again labelled according to the region it belongs to (given by x_6).

and the mean has been subtracted during standardization the mean values of the data projected onto \mathbf{v}_1 and \mathbf{v}_2 will be zero and thus the covariance between the observations projected onto PC1 and PC2 given by $\frac{1}{N-1}(\tilde{\mathbf{X}}\mathbf{v}_1)^\top(\tilde{\mathbf{X}}\mathbf{v}_2) = (\mathbf{U}\mathbf{S}\mathbf{V}^\top\mathbf{v}_1)^\top(\mathbf{U}\mathbf{S}\mathbf{V}^\top\mathbf{v}_2) = \mathbf{u}_1^\top s_{11} \mathbf{u}_2 s_{22} \mathbf{u}_1^\top \mathbf{u}_2 = 0$ and there can therefore be no correlation between the observations projected onto PC1 and PC2.

Summary Statistics

28. oktober 2021 17:06

Spring2020

	Mean	Std	$x_{p=25\%}$	$x_{p=50\%}$	$x_{p=75\%}$
BirthRt	29.46	13.62	14.6	29	42.575
DeathRt	10.73	4.66	7.7	9.5	12.4
InfMort	55.28	46.05	13.025	43	88.25
LExpM	61.38	9.67	55.2	63.4	68.55

Table 2: Summary statistics of the first four attributes of the Poverty dataset. The column $x_{p=25\%}$ refers to the 25'th percentile of the given attribute, $x_{p=50\%}$ to the median and $x_{p=75\%}$ to the 75'th percentile.

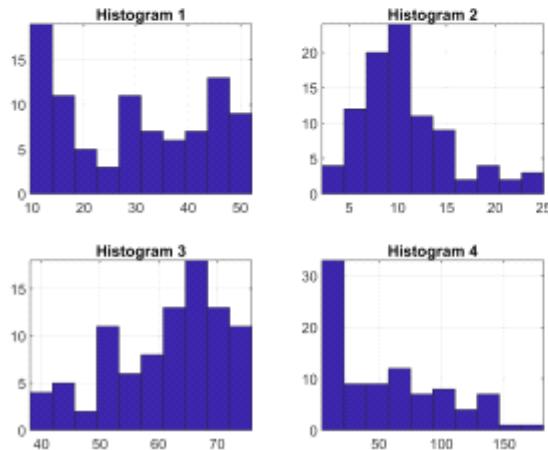


Figure 1: Four histograms corresponding to the variables with summary statistics given in Table 2 but not necessarily in that order.

Question 2.

Table 2 contains summary statistics of the first four attributes of the Poverty dataset. Which of the histograms in Figure 1 match which of the attributes according to their summary statistics?

- A. *BirthRt* matches histogram 4, *DeathRt* matches histogram 2, *InfMort* matches histogram 1 and *LExpM* matches histogram 3.
- B. *BirthRt* matches histogram 4, *DeathRt* matches histogram 1, *InfMort* matches histogram 3 and *LExpM* matches histogram 2.
- C. *BirthRt* matches histogram 2, *DeathRt* matches histogram 3, *InfMort* matches histogram 1 and *LExpM* matches histogram 4.
- D. *BirthRt* matches histogram 1, *DeathRt* matches histogram 2, *InfMort* matches histogram 4 and *LExpM* matches histogram 3.

Solution 2. To solve the problem, note that we can read off the median, 25'th, and 75'th percentiles from Table 2 as $q_{p=50\%}$, $q_{p=25\%}$, and $q_{p=75\%}$ respectively. These can be matched to the histograms in Figure 1 by observing histogram 2 does not have observations above 25 and thus must therefore be *DeathRt*. Histogram 4 is the only histogram having observations above 88.25 which only holds for *InfMort* (see 75th percentile). This only holds for answer option D.



02450: Introduction to Machine Learning and Data Mining

Measures of similarity, summary statistics and probabilities

Jes Frellsen

DTU Compute, Technical University of Denmark (DTU)

DTU Compute
Department of Applied Mathematics and Computer Science

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$
$$\int_a^b \Theta^{\sqrt{17}} + \Omega \delta e^{i\pi} = -1$$
$$\text{E}^{\infty} = \{2.7182818284\}^{\text{dtu}}$$
$$\chi^2 \sum!$$

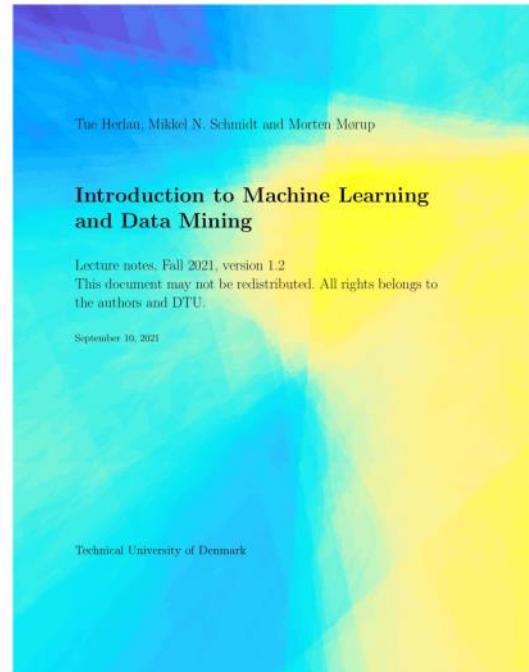
Today

Feedback Groups of the day:

Akshat Bhardwaj, Aldis Helga Bjorgvinsdottir, Alysha K Chamadia, Andreas Holmer Bigom, Andrew Boonto Blumensen, Anja Lykke Borre, Anna Brønd, Ashish Chawla, Aske Bruun-Ringgaard, August Hertz Bugge, August Valentin Nørgaard Birch, Belen Castellote Lopez, Benjamin Bogø, Bogdan Capsa, Calle Ryge Carlsen, Chinmay Bhalla, Christoffer Binzer Bjørner, Connor Ward Chewning, Dimitris Bokos-Zygouris, Frederik Bjørling Bornemann, Frederik Jakob Eskildsen Bruun, Fridtjof Cerup-Simonsen, Hanna Maria Börtin, Helga Pórey Björnsdóttir, Jakob Friis Christiansen, Joana Cardiff Aleu, Johan Matzen Brodt, Johan Stauner Bill, João Augusto César Moutinho, Juan Cervera Bustamante, Juliane Bjørn Budde, Julien François Per Chegaray, Jun Chen, Kian Bostani Nezhad, Kunt Alp Celebi, Lars Knuth Satoshi Boyens-Thiele, Lorenzo Capaldo, Ludovica Caccaro, Marion Josephine Isabelle Maëlys Cadet, Mikkel Wøidemann Klæbel Blomsterberg, Niels Karsten Bisgaard-Bohr, Oriol Cayon Domingo, Oscar Carpentier, Pau Carrascal Fabregat, Peter Sebastian Hein Bitsch, Peter Tønder Blendstrup, Philipp Bockshecker, Piet Johan Brochorst Christensen, Raquel Maria Casañ Crespo, Rares-Victor Botis, Ruixin Chen, Rune Yding Brogaard, Samy Chekkouri, Simon Buk-Mortensen, Sueanoi Chokswas, Tyme Chatupanyachotikul, Xenia Cohen Chime, Yuechen Chen, Zhe Chen

2 DTU Compute

Reading material: Chapter 4, Chapter 5



Lecture 3 14 September, 2021

Lecture Schedule

1 Introduction

31 August: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

7 September: C2, C3

3 Measures of similarity, summary statistics and probabilities

14 September: C4, C5

4 Probability densities and data visualization

21 September: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

28 September: C8, C9

6 Overfitting, cross-validation and Nearest Neighbor

5 October: C10, C12 (Project 1 due before 13:00)

7 Performance evaluation, Bayes, and Naive Bayes

12 October: C11, C13

8 Artificial Neural Networks and Bias/Variance

26 October: C14, C15

9 AUC and ensemble methods

2 November: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 November: C18

11 Mixture models and density estimation

16 November: C19, C20 (Project 2 due before 13:00)

12 Association mining

23 November: C21

Recap

13 Recap and discussion of the exam

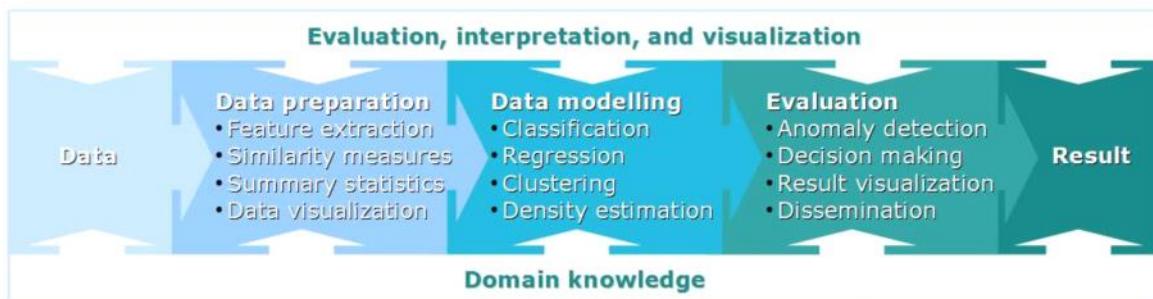
30 November: C1-C21

Online help: Forum on DTU Learn

Videos of lectures: <https://video.dtu.dk>

Streaming of lectures: Zoom (link on DTU Learn)

Lecture 3 14 September, 2021



Learning Objectives

- Compute measures of similarity/dissimilarity (L_p distance, cosine distance, etc.)
- Understand basics of probability theory and stochastic variables in the discrete setting
- Understand probabilistic concepts such as expectations, independence and the Bernoulli distribution
- Understand the maximum likelihood principle for repeated binary events

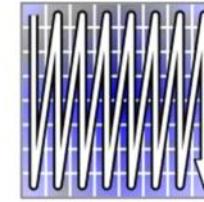
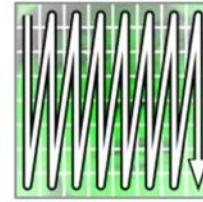
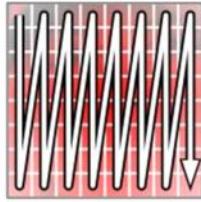
PCA recap: Principal component analysis on images



- 1000 images, 86 x 86 pixels, 3 RGB intensities

Tamara Berg "Faces in the wild"

Pre-processing



- Concatenate all pixel color values in one long vector
 - $86 \times 86 \times 3 = 22'188$
 - Image is now represented as a 22'188 dimensional vector
- Stack all 1000 images into a big matrix
 - $1000 \times 22'188$

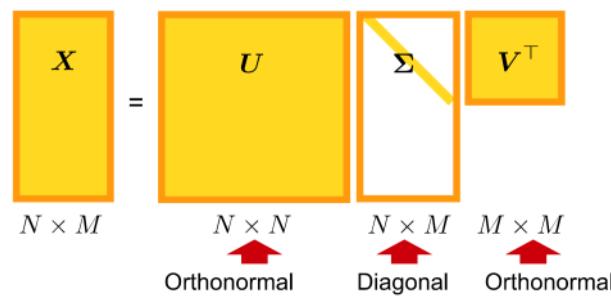
Principal component analysis (PCA)

1. Subtract the mean

- Consider dividing with variance; use 1-out-of-K coding for nominal attributes

2. Compute the singular value decomposition (SVD)

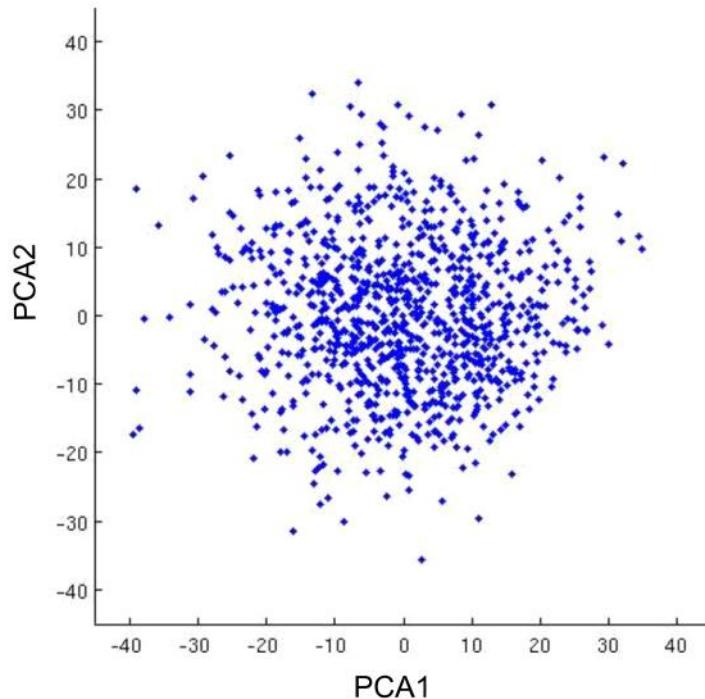
- Orthogonal linear transformation
- Transforms data to a new coordinate system
 - Greatest variance along the first axis (first column of V)
 - Second greatest variance along the second axis



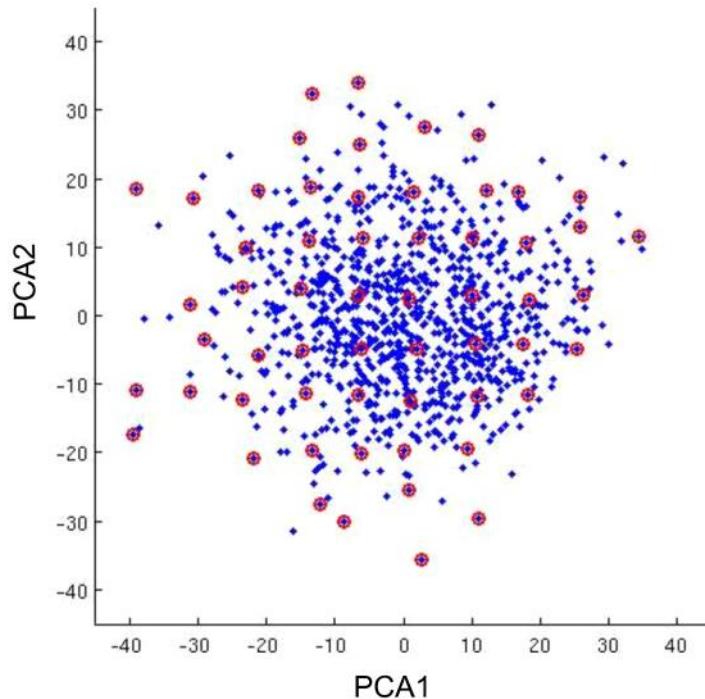
• Plot data in the transformed coordinate system

- Corresponds to looking at data from an angle where it is most spread out

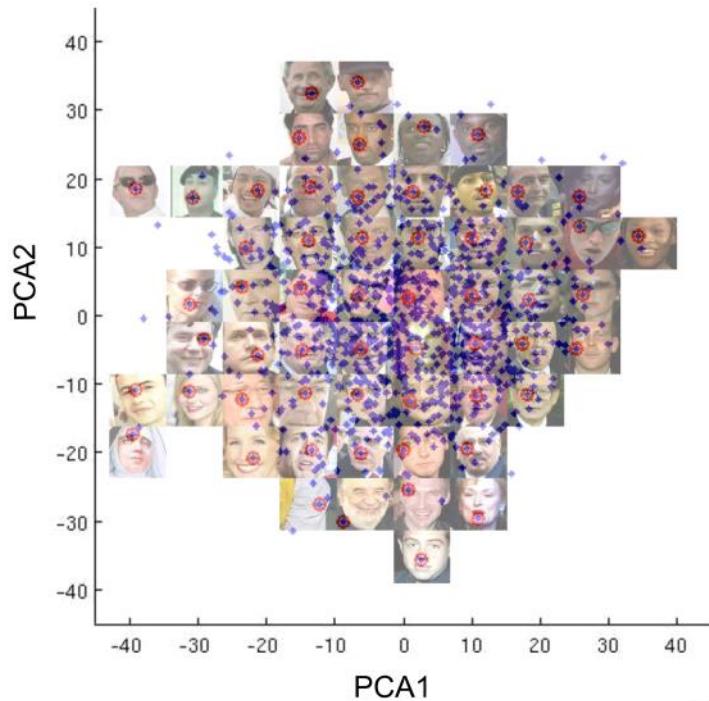
PCA on face images



PCA on face images

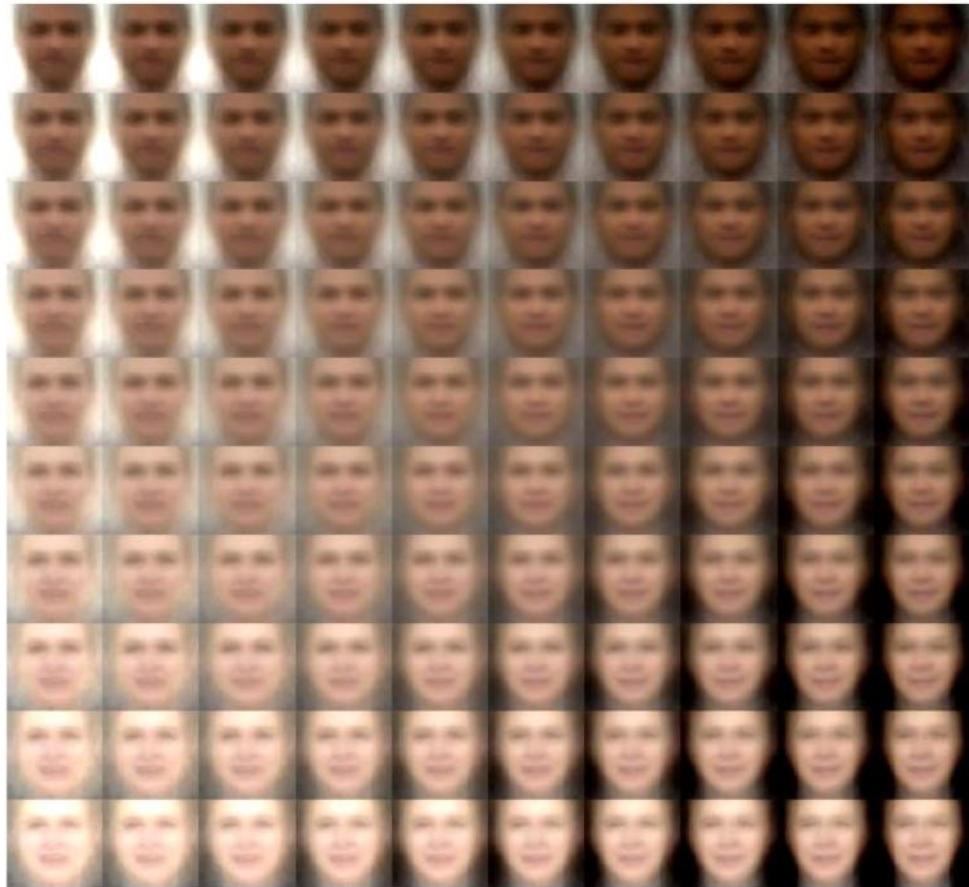


PCA on face images



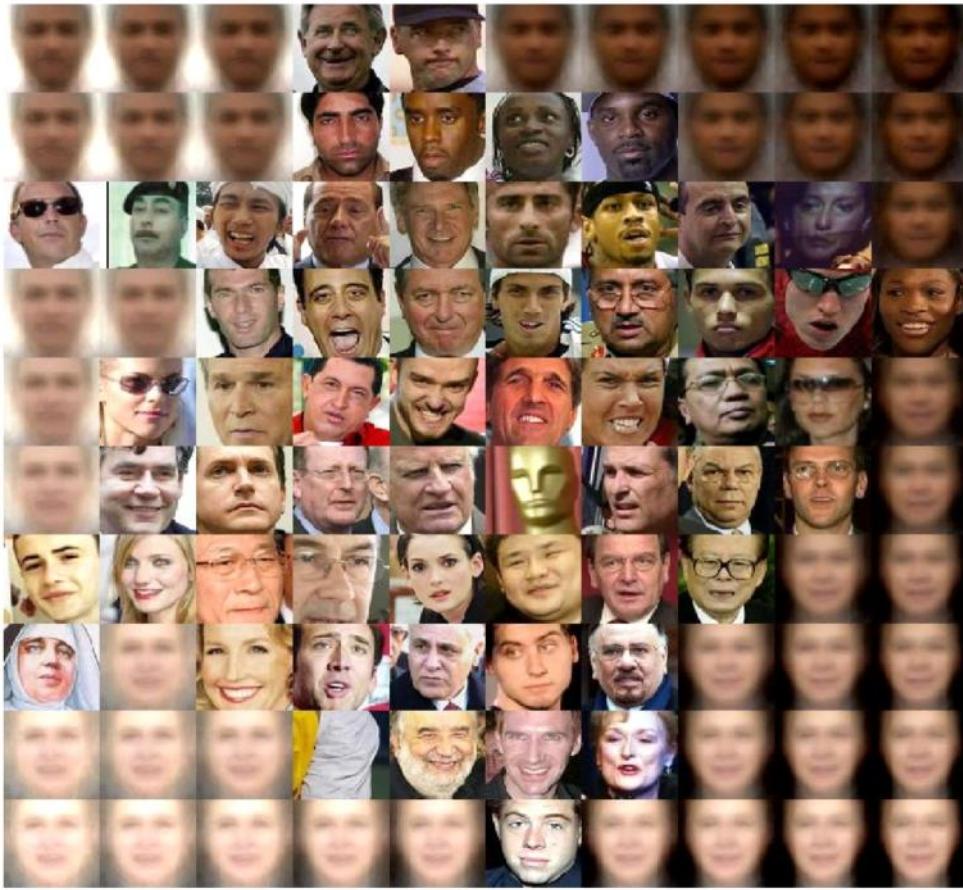


- What information do the two principal axes capture?





What information do the two principal axes capture?



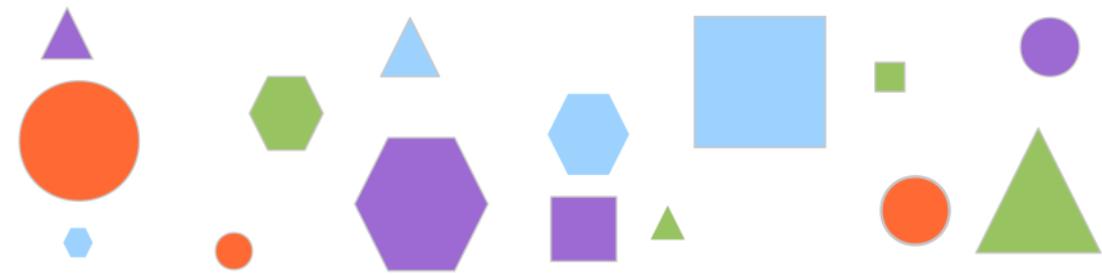
Similarity / Dissimilarity measures

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

Dissimilarity $d(\mathbf{x}, \mathbf{y})$ 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



Dissimilarity measures

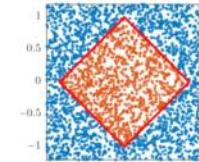
Minkowsky is the general rule: the other are derivations

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

- One-norm ($p = 1$)

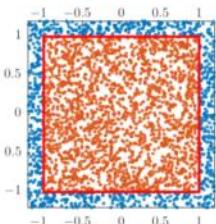
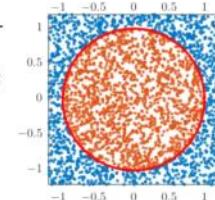
it'd be good if you don't want to be sensitive

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



- Euclidean ($p = 2$)

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



- Max-norm distance ($p = \infty$)
good choice for looking for outliers

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

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

Imagine X and Y binary vectors

Similarity measures

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

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)

it's symmetric in 0 and 1, if I change the 0s with the 1s in

the 2 vectors, the SMC value doesn't change:

useful because sometimes 0s or 1s could
be not significative

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

different vectors size means the number
of non-0s entries in the vector

* Symmetric

* Positive matches

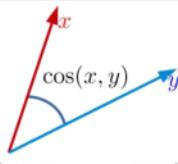
Jaccard Coefficient

only focuses in positive matches

but if the two vectors are different size we might get $J(x, y) = \frac{f_{11}}{K - f_{00}}$

high value -> solved with the COSINE

Cosine similarity



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

divide by the norm, and focuses on positive matches
works for both binary vectors but also for any other kind

Extended Jaccard coefficient

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

Also defined for continuous data $\|x^2\|$ counts the number of non-0s in the vector

* Positive matches
* Document length

Quiz 1, 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.

$F_{00} = 1$

$F_{11} = 2$

$k = 5$

$x \cdot T \cdot y = 2$

$\|x\| = 3^{(1/2)} = 0^2 + 1^2 + 1^2 \dots = \text{square root of non-0s entries}$

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

K : Total number of attributes

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

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

Which of the following statements are true?

- A. $SMC(o_1, o_2) = \frac{3}{5}, J(o_1, o_2) = \frac{1}{2}, \cos(o_1, o_2) = \frac{2}{3},$
- B. $SMC(o_1, o_2) = \frac{3}{5}, J(o_1, o_2) = \frac{3}{4}, \cos(o_1, o_2) = \sqrt{\frac{2}{3}},$
- C. $SMC(o_1, o_2) = \frac{2}{5}, J(o_1, o_2) = \frac{1}{3}, \cos(o_1, o_2) = \frac{2}{3},$
- D. $SMC(o_1, o_2) = \frac{2}{5}, J(o_1, o_2) = \frac{1}{3}, \cos(o_1, o_2) = \sqrt{\frac{2}{3}},$
- E. Don't know.

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

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

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

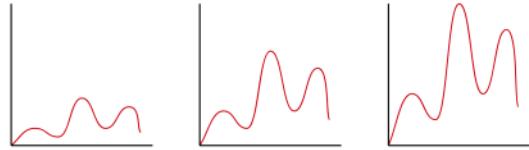
$$EJ(x, y) = \frac{x^\top y}{\|x\|_2^2 + \|y\|_2^2 - x^\top y}$$

Invariance

Scale invariance

the COSINE similarity is scale invariant

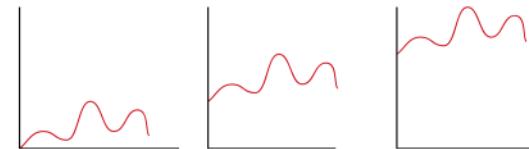
$$d(\mathbf{x}, \mathbf{y}) = d(\alpha \mathbf{x}, \mathbf{y})$$



Translation invariance

required in image recognition for instance,
in I scale the image, or add some light, or rotate, I
want my method to be robust

$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{x} + \mathbf{z}, \mathbf{y})$$



General invariances

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

Transformations

Standardization: Ensure a single attribute will not dominate:

IMPORTANT = the scale of the 3 attributes

is really different, if I don't standardize them
all the differences will focus mainly on $\tilde{x}_{ik} = \frac{x_{ik} - \hat{\mu}_k}{\hat{\sigma}_k}$

If I standardize when not needed I may lose some information

Example:

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

Combining heterogeneous attributes Transform measures and combine

- we first get a similarity measure between the education

- $S = 1 / (1+D)$: if I want to transform a Dissimilarity measure into a Similarity one.

$$s_{Edu.} = SMC(x_{Edu.}, y_{Edu.})$$

the choose of the function

$$s_{Age.} = a (a + d_1(x_{Age.}, y_{Age.}))^{-1}, \quad a = 1$$

'a' depends on the application

to combine the two

$$s(x, y) = \frac{1}{2} (s_{Edu.} + s_{Age.})$$

- use the mean

- a particular weight

Example:

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

Weighting Attributes have different importance

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

Empirical statistics

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

- Empirical mean

$$\hat{\mu} = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- Empirical variance

$$\hat{s} = \text{var}[x] = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

- Empirical covariance

$$\text{cov}[x, y] = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$$

- Empirical standard deviation

$$\hat{\sigma} = \text{std}[x] = \sqrt{\hat{s}}$$

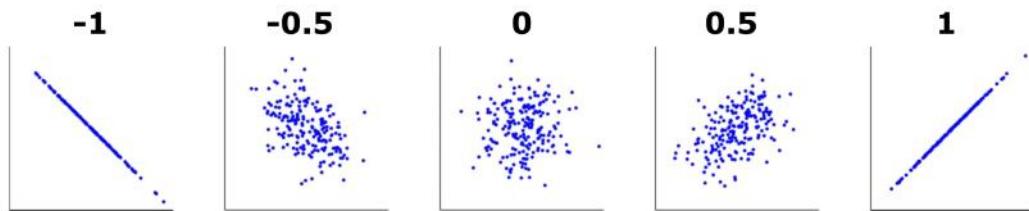
Correlation

- 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 there is a perfect linear relation

$$x_k = ay_k + b$$



Quantiles

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

Quantiles describe the *points* that divide the underlying distribution into intervals of equal probability:

- 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 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(N)}$

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

Probabilities

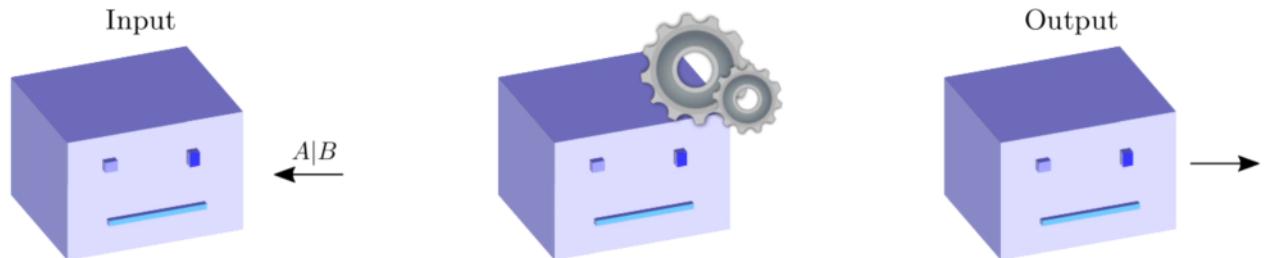
Continuous optimization properties: if I have an image classification, I want the output says that the one shown is a cat with a certain amount of probability, instead of saying only is a cat or dog (discrete optimization)

Pragmatically: A big part of AI is dealing with **uncertainty** and **incomplete information**. Probabilities is the formal framework for doing so.

Algorithmically: If an image belongs to a particular category is a discrete event. The **probability** it belongs to a category is continuous. Algorithmically, easier to optimize continuous quantities.

Convenience: There are boiler-plate ideas for transforming a **probabilistic** assumption into an algorithm (maximum likelihood).

Probabilities



Assuming B is true, how plausible is it A is true?

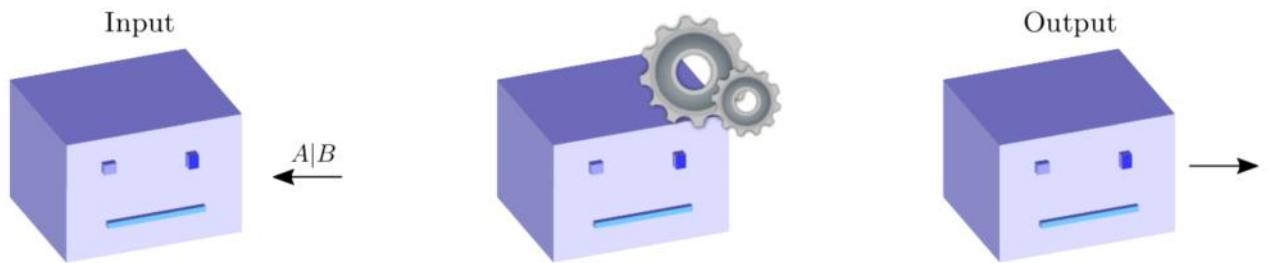
Assuming B , the plausibility of A is
(low / medium / high / certain)

We reason about a proposition A in light of evidence B :
the probability of A given B is $x \rightarrow$ the degree of probability
that A is true knowing that B is true, is equal to $P(A|B) = x$

The degree-of-belief that A is true given B is accepted as true is at a level x

- A number between 0 and 1
- A and B are always binary (true/false) propositions
- Represents a *state of knowledge*

Probabilities: Trial example



Assuming B is true, how plausible is it A is true?

Assuming B , the plausibility of A is
(low / medium / high / certain)

G : *The accused is guilty*

E_1 : *A car similar to his was seen at the crime scene.*

E_2 : *A large sum of money was found in his posession*

E_3 : *His fingerprints was found at the door of the bank.*

Probabilities express states-of-knowledge

the probability of A given B is $x \rightarrow$ the degree of probability

that A is true knowing that B is true, is equal to X

$E \equiv E_1$ and E_2 and E_3

$P(G|E) > P(G|E_2)$ if we get more knowledge, we become more certain about something

Binary propositions

A binary proposition is a statement which is either true or false (we might not know, but someone with complete knowledge would)

A : In 49 BCE, Caesar crossed the Rubicon

B : Acceleration sensor 39 measures more than 0.85

C : Patient 901 has high cholesterol

Propositions can be combined with **and**, **or** and **not**:

$AB \equiv$ True if A and B are both true

$A + B \equiv$ True if either A or B are true

$\bar{A} \equiv$ True if A is false

We define two special propositions which is always **true/false**:

1 : A proposition which is always true

0 : A proposition which is always false

...and the following identities: $A1 = A$, $A + \bar{A} = 1$, $\bar{\bar{A}} = A$ and

$$A(B_1 + B_2 + \cdots + B_n) = AB_1 + AB_2 + \cdots + AB_n$$

Quiz 2, Probabilities

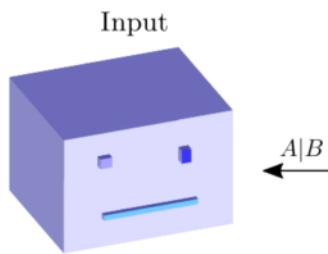
Assume we define the following 4 boolean variables.

- R_1 : Handed in report 1
- R_2 : Handed in report 2
- R_3 : Handed in report 3
- F : Student failed 02450

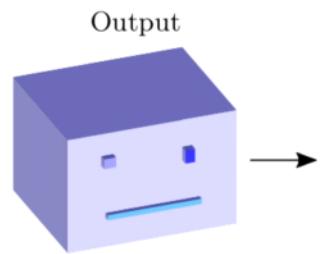
How would you express the probability of the statement:

If a student hand in report 1, 2 and 3, the chance of passing 02450 is greater than 90%?

- A. $P(R_1 R_2 R_3 | F) > 0.9$
- B. $P(\overline{F} | R_1 + R_2 + R_3) > 0.9$
- C. $P(\overline{F} | R_1 R_2 R_3) > 0.9$
- D. $P(R_1 + R_2 + R_3 | F) > 0.9$
- E. Don't know.



Assuming B is true, how plausible is it A is true?

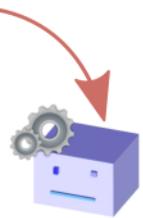


Assuming B , the plausibility of A is
(low / medium / high / certain)

Rules of probability

The sum rule: $P(A|C) + P(\bar{A}|C) = 1$

The product rule: $P(AB|C) = P(B|AC)P(A|C)$



Interpretation:

$P(A|B) = 0$ (*interpretation: given B is true, A is certainly false*)

$P(A|B) = 1$ (*interpretation: given B is true, A is certainly true*)

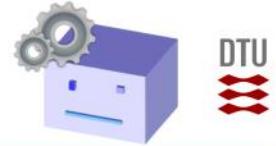
We also use the shorthand:

$$P(A|1) = P(A)$$

$$p(A) + p(\bar{A}) = 1$$

$$p(AB) = P(A|B)P(B)$$

Remarkably, this is the mathematical basis for this course



Marginalization and Bayes' theorem

$$\text{Sum rule} \quad P(A|C) + P(\bar{A}|C) = 1$$

$$\text{Product rule} \quad P(AB|C) = P(B|AC)P(A|C)$$

Marginalization

$$\begin{aligned} P(B|C) &= P(B|C) [P(A|BC) + P(\bar{A}|BC)] = P(AB|C) + P(\bar{A}B|C) \\ &= P(B|AC)P(A|C) + P(B|\bar{A}C)P(\bar{A}|C). \end{aligned}$$

Bayes theorem

$$\begin{aligned} P(A|BC) &= \frac{P(B|AC)P(A|C)}{P(B|C)} \\ &= \frac{P(B|AC)P(A|C)}{P(B|AC)P(A|C) + P(B|\bar{A}C)P(\bar{A}|C)}. \end{aligned}$$

$$p(AB) = p(B|A) * p(A)$$

$$p(AB) = p(A|B) * p(B)$$

$$p(A|B) * p(B) = p(B|A) * p(A) \rightarrow p(A|B) = p(B|A) * p(A)$$

DNA



Bayes theorem

$$P(A|BC) = \frac{P(B|AC)P(A|C)}{P(B|AC)P(A|C) + P(B|\bar{AC})P(\bar{A}|C)}$$

Crimes may be solved by matching crime-scene DNA to DNA in a database

- If the two samples are from the same person, a DNA test will always give a positive match
- If the DNA are from different persons, DNA will incorrectly give a positive match one time out of a million

A crime is committed in Racoon City by an unidentified male. Assume all 8000 possible perpetrators undergo a DNA test, and suppose the DNA test gives a positive result for George. What is the chance George is guilty?

G : George is guilty, D : There was a positive DNA match



$$p(G|D) = \frac{p(D|G) * P(G)}{p(D)}$$

$$p(D) = p(GD) + p(\bar{G}D) = p(D|G) * p(G) + p(D|\bar{G}) * p(1 - G) = \\ = 1 * 1/8000 + 10^{-6} * (1 - 1/8000) = 99\%$$

$$p(D|G) = 1$$

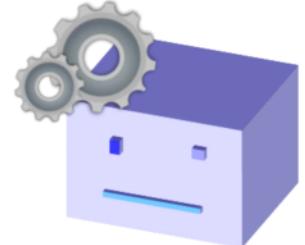
$$p(G) = 1/8000$$

Exclusive and exhaustive events

- | | |
|-----------------------------|-----------------------------|
| A_1 : The side ☐ face up. | A_2 : The side ☑ face up. |
| A_3 : The side ☒ face up. | A_4 : The side ☓ face up. |
| A_5 : The side ☔ face up. | A_6 : The side ☕ face up. |

- When no two propositions can be true at the same time, they are said to be **mutually exclusive**: $A_i A_j = 0$ for $i \neq j$
- Consider any two events A and B

$$P(A + B) = P(A) + P(B) - P(AB)$$



- In general, for n mutually exclusive events

$$P(A_1 + A_2 + \dots + A_n) = \sum_{i=1}^n P(A_i)$$

- A set of events is **exhaustive** if one has to be true: $A_1 + \dots + A_n = 1$. Then:

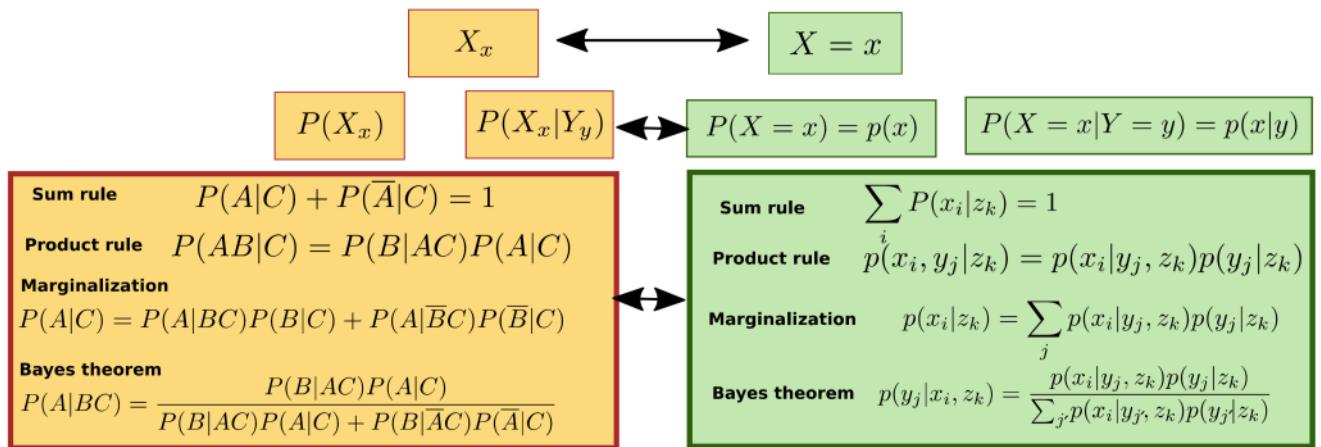
$$\sum_{i=1}^n P(A_i) = P(A_1 + A_2 + \dots + A_n) = 1$$

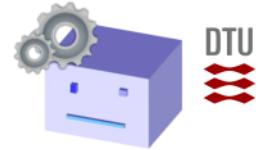
Stochastic variables

- Often, we will measure numerical quantities (number of children, age of a patient, etc.)
- Suppose a quantity X (number of children) takes a value $x = 3$. We can write this as the binary event X_3 and in general:

$X_x : \{\text{The binary event that } X \text{ is equal to the number } x\}$

- Stochastic variable simplify this notation by the definition:





Quiz 3, Avila bible (Fall 2018)

$p(\tilde{x}_2, \tilde{x}_{10} y)$	$y = 1$	$y = 2$	$y = 3$
$\tilde{x}_2 = 0, \tilde{x}_{10} = 0$	0.19	0.3	0.19
$\tilde{x}_2 = 0, \tilde{x}_{10} = 1$	0.22	0.3	0.26
$\tilde{x}_2 = 1, \tilde{x}_{10} = 0$	0.25	0.2	0.35
$\tilde{x}_2 = 1, \tilde{x}_{10} = 1$	0.34	0.2	0.2

Table 1: Probability of observing particular values of \tilde{x}_2 and \tilde{x}_{10} conditional on y .

We will consider a dataset based on the Avila bible. We wish to predict the copyist ($y = 0, 1, 2$) of a bible based on the two typographic attributes *upperm* and *mr/is*. We suppose the attributes have been binarized such that *upperm* corresponds to $\tilde{x}_2 = 0, 1$ and *mr/is* to $\tilde{x}_{10} = 0, 1$. Suppose the probability for each of the configurations of \tilde{x}_2 and \tilde{x}_{10} conditional on the copyist y are as given in Table 1. and the prior probability of

the copyists is

$$p(y = 1) = 0.316, p(y = 2) = 0.356, p(y = 3) = 0.328.$$

Using this, what is then the probability an observation was authored by copyist 1 given that $\tilde{x}_2 = 1$ and $\tilde{x}_{10} = 0$?

- A. $p(y = 1|\tilde{x}_2 = 1, \tilde{x}_{10} = 0) = 0.25$
- B. $p(y = 1|\tilde{x}_2 = 1, \tilde{x}_{10} = 0) = 0.313$
- C. $p(y = 1|\tilde{x}_2 = 1, \tilde{x}_{10} = 0) = 0.262$
- D. $p(y = 1|\tilde{x}_2 = 1, \tilde{x}_{10} = 0) = 0.298$
- E. Don't know.

Sum rule $\sum_i p(x_i|z_k) = 1$

Product rule $p(x_i, y_j|z_k) = p(x_i|y_j, z_k)p(y_j|z_k)$

Marginalization $p(x_i|z_k) = \sum_j p(x_i|y_j, z_k)p(y_j|z_k)$

Bayes theorem $p(y_j|x_i, z_k) = \frac{p(x_i|y_j, z_k)p(y_j|z_k)}{\sum_j p(x_i|y_j, z_k)p(y_j|z_k)}$

Lecture 3 14 September, 2021

Independence



Independent: $p(x_i, y_j) = p(x_i)p(y_j)$

Conditionally independent given z_k : $p(x_i, y_j | z_k) = p(x_i | z_k)p(y_j | z_k)$

Expectations

$$\text{Expectation: } \mathbb{E}[f] = \sum_{i=1}^N f(x_i)p(x_i). \quad (2)$$

$$\text{mean: } \mathbb{E}[x] = \sum_{i=1}^N x_i p(x_i), \quad \text{Variance: } \text{Var}[x] = \sum_{i=1}^N (x_i - \mathbb{E}[x])^2 p(x_i). \quad (3)$$

Example: Uniform probability

$$\begin{aligned} p(x_i) &= \frac{1}{N} \\ \mathbb{E}[f] &= \frac{1}{N} \sum_{i=1}^N f(x_i) \\ \mathbb{E}[x] &= \frac{\sum_{i=1}^N x_i}{N} \\ \text{Var}[x] &= \frac{1}{N} \sum_{i=1}^N (x_i - \mathbb{E}[x])^2 \end{aligned}$$

Densities and models

- In machine learning, we want to learn a parameter from data
- Models of the data which use parameters are how we do that
- We build models out of simpler building blocks (distributions (discrete variables see chapter 5) and densities (continuous variables see chapter 6)). In this course we will use four:

Bernoulli distribution

The Categorical distribution

The Beta density

The Multivariate normal density

The Bernoulli distribution

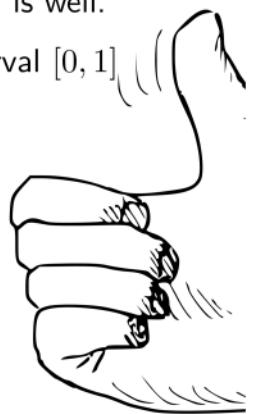
- Let $b = 0, 1$ denote a binary event.

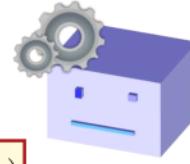
- For instance,

- $b = 0$ corresponds to heads, and $b = 1$ to tails, or
- $b = 0$ corresponds to a person being ill, and $b = 1$ that a person is well.

- The probability of b is expressed using a parameter θ in the unit interval $[0, 1]$

$$\text{Bernoulli distribution: } p(b|\theta) = \theta^b(1-\theta)^{1-b}.$$





The Bernoulli distribution, repeated events

Conditional independence $p(x_i, x_j | z_k) = p(x_i | z_k)p(x_j | z_k)$

- Suppose we observe a sequence b_1, \dots, b_N of Bernoulli (binary) events.
- For instance, for N patients we record whether person 1 is ill or well ($b_1 = 0$ or $b_1 = 1$) and up to whether patient N is ill or well ($b_N = 0$ or $b_N = 1$)
- When we **know** θ (the chance a person is well or ill), the events are **independent**

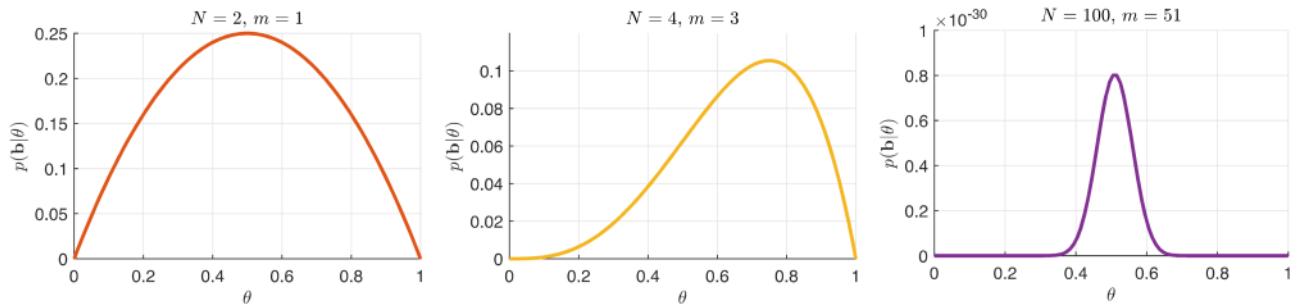
Bernoulli distribution: $p(b|\theta) = \theta^b(1-\theta)^{1-b}$.

$$p(b_1, \dots, b_N | \theta) = \prod_{i=1}^N p(b_i | \theta) = \prod_{i=1}^N \theta^{b_i}(1-\theta)^{1-b_i} = \theta^{\sum_{i=1}^N b_i}(1-\theta)^{N-\sum_{i=1}^N b_i}$$

product of N bernoulli
distribution

$$= \theta^m(1-\theta)^{N-m}, \quad m = b_1 + b_2 + \dots + b_N$$

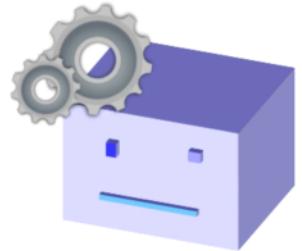
The Bernoulli distribution, maximum likelihood



$$p(b_1, \dots, b_N | \theta) = \theta^m (1 - \theta)^{N-m}$$

An idea for selecting θ^* is **Maximum likelihood**

$$\theta^* = \arg \max_{\theta} p(b_1, \dots, b_N | \theta)$$



The value of θ according to which the data is most plausible

Resources

<https://bayes.wustl.edu> Classical textbook which treats probabilities as states-of-knowledge and discuss many practical and philosophical issues (this book converted me to ML!)
(<https://bayes.wustl.edu/etj/prob/book.pdf>)

<https://02402.compute.dtu.d> A more in-depth description of summary statistics (see chapter 1) (<https://02402.compute.dtu.dk>)

<https://www.khanacademy.org> An excellent introduction to probability theory which we recommend as a go-to resource

(<https://www.khanacademy.org/math/statistics-probability/probability-library>)

<http://citeseervx.ist.psu.edu> A more in-depth discussion of Bayes in the court room (<http://citeseervx.ist.psu.edu/viewdoc/download;jsessionid=EF0328140036A4C668BB5B9FC76C9BE?doi=10.1.1.599.8675&rep=rep1&type=pdf>)



lec4



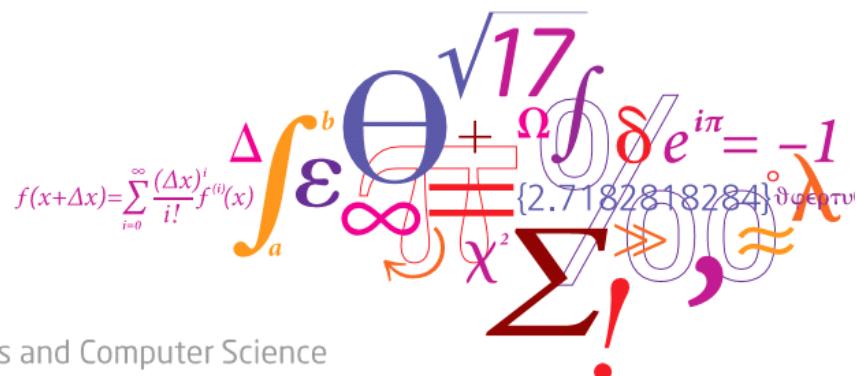
02450: Introduction to Machine Learning and Data Mining

Probability densities and data visualization

Jes Frellsen

DTU Compute, Technical University of Denmark (DTU)

DTU Compute
Department of Applied Mathematics and Computer Science

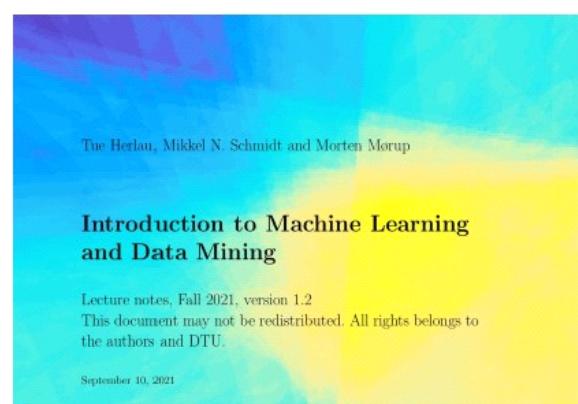


Today

Feedback Groups of the day:

Adèle des Moutis, Alexander Ulbæch Dupont, Alonso Espasandín Hernán, Andrea Maria Pierri Enevoldsen, Anna Bøgevang Ekner, Arian Dapouyeh, Artur Coelho Fabrício Rodrigues, Azra Cigura, Benjamin Becker Danielsen, Bjarke Lars Verner Bruhn Erichsen, Bufan Deng, Cassandra Desvig-Forss, Chiara de Geeter, Christina Davidsen Damm, Cornelius Erichs, Dídac Clària Hernández, Eliot Claret, Elisabeth Barfod Damgaard, Emeline Debalme, Emilie Sofie Engdal, Evangelos Diouslous, Fernando Cruz Ceravalls, Filip Furbo Enevoldsen, Gül Sude Demircan, Ilektra Amrali, Jiyuan Cui, Joao Maria De Sacadura Botte Corte-Real, Julian Teglgaard Drachmann, Julie Favre, Kristin Engel, Laura Degand, Laurine Dargaud, Lautaro Cabrera De Pietri, Leonard Diederich, Lluís Colomer Coll, Louis Kamp Eskildsen, Magnus Døvle Andrews, Marc Dvhiring, Marcus Deichmann

Reading material:
Chapter 6, Chapter 7



Kristin Engel, Laura Degaud, Laurine Dargaud,
Lautaro Cabrera De Pietri, Leonard Diederich, Lluís
Colomer Coll, Louis Kamp Eskildsen, Magnus Døvle
Andrews, Marc Dyhring, Marcus Deichmann
Christiansen, Mathias Hadi Dyhr, Michael Damholt,
Nicolás Ellarby Sánchez, Oliver Rosbæk Elmgreen,
Patrick Hørlykke Clemmensen, Pauline Lystbæk
Christiansen, Peter Munch Ejlev, Robert Edwin Peter
Wilsch, Ruben Eland, Sebastian Ekpete, Signe Maite
Conde Frieboes, Simone Elisabeth Engelbrecht,
Smilla Due, Tanish Dhagat, Thor Nørgaard Eriksen,
Vlad Dragusan, Yanghao Cui, Zakaria Abdikarim Ali

This document may not be redistributed. All rights belongs to
the authors and DTU.

September 10, 2021

Technical University of Denmark

2 DTU Compute

Lecture 4 21 September, 2021



Lecture Schedule

1 Introduction

31 August: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

7 September: C2, C3

3 Measures of similarity, summary statistics and probabilities

14 September: C4, C5

4 Probability densities and data visualization

21 September: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

28 September: C8, C9

6 Overfitting, cross-validation and Nearest Neighbor

5 October: C10, C12 (Project 1 due before 13:00)

7 Performance evaluation, Bayes, and Naive Bayes

12 October: C11, C13

8 Artificial Neural Networks and Bias/Variance

26 October: C14, C15

9 AUC and ensemble methods

2 November: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 November: C18

11 Mixture models and density estimation

16 November: C19, C20 (Project 2 due before 13:00)

12 Association mining

23 November: C21

Recap

13 Recap and discussion of the exam

30 November: C1-C21

Online help: Forum on DTU Learn

Videos of lectures: <https://video.dtu.dk>

Streaming of lectures: Zoom (link on DTU Learn)

3 DTU Compute

Lecture 4 21 September, 2021

Probabilities

- In more common notation we have

- Sum rule

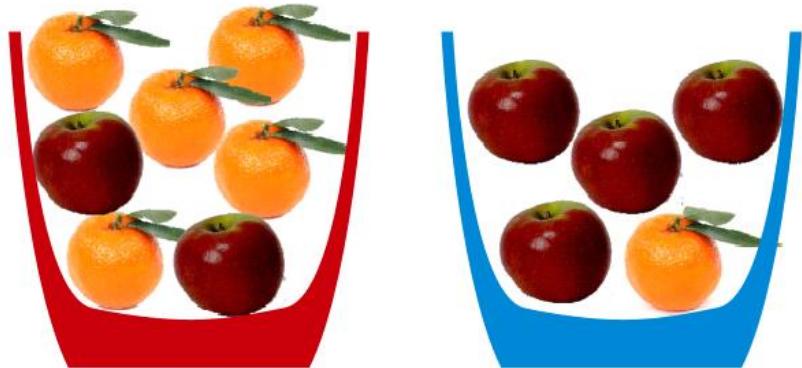
$$p(x) = p(x, y=0) + p(x, y=1)$$

- Product rule

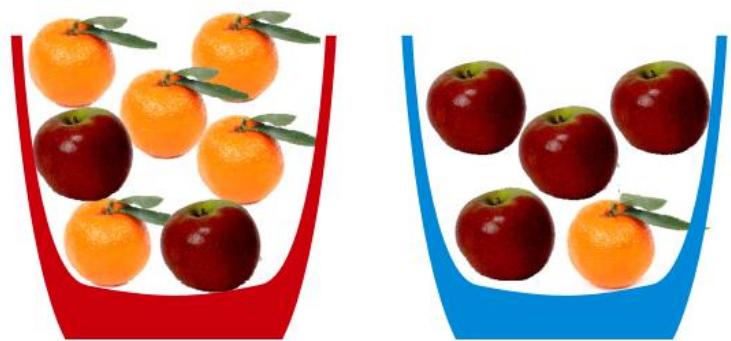
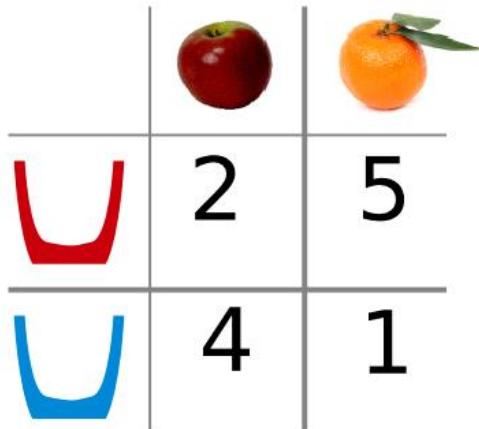
$$p(x, y) = p(x|y)p(y)$$

- Bayes' rule

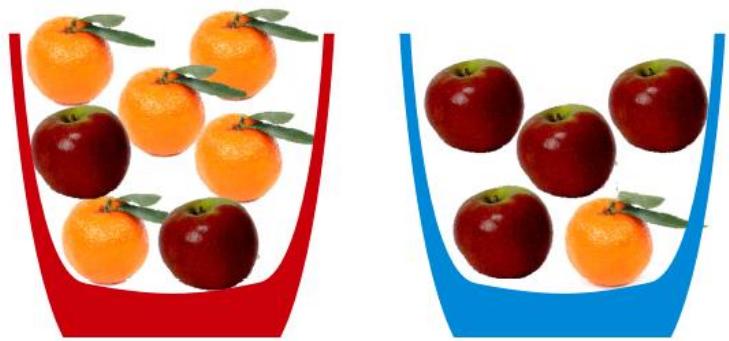
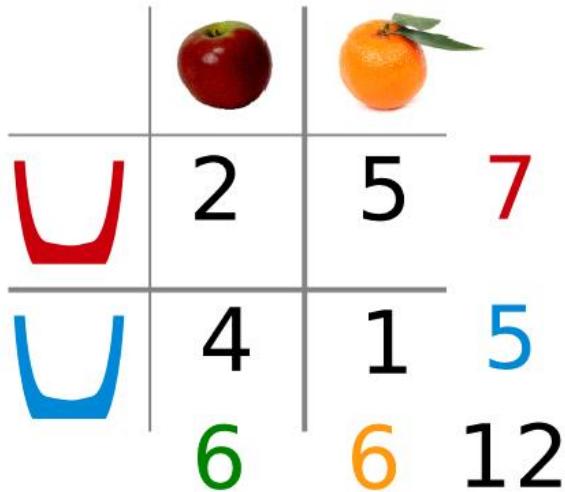
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$



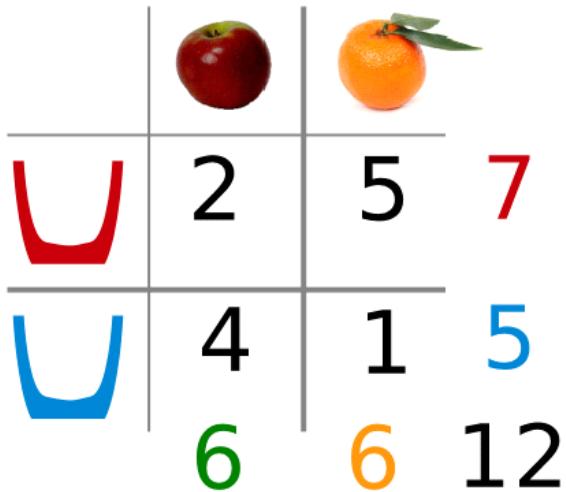
- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

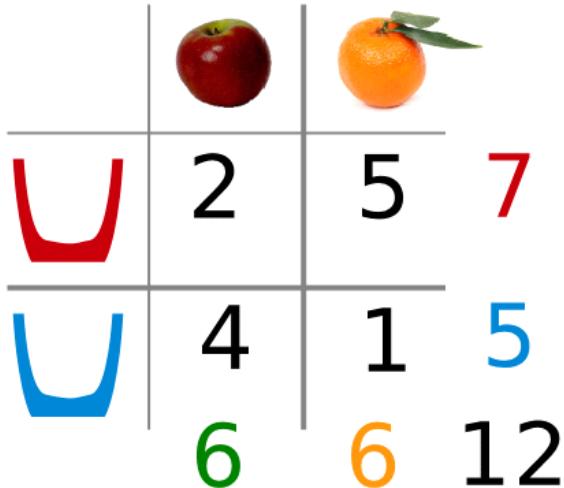


- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

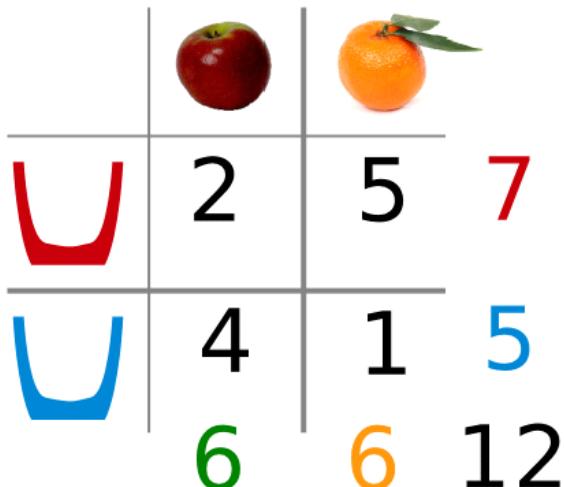
- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

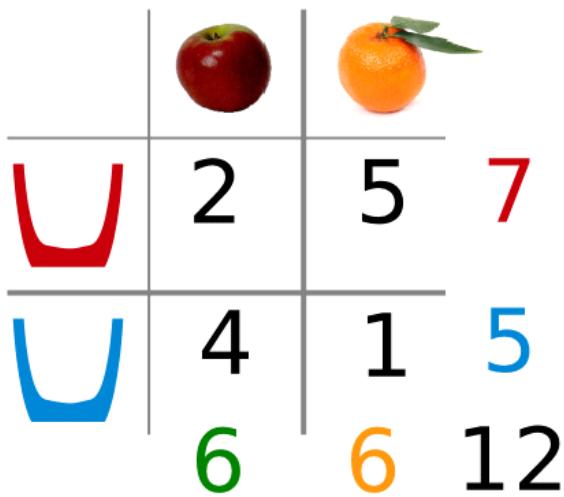
- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$\begin{aligned} p(r|o) &= \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6 \\ &= \frac{p(o|r)p(r)}{p(o)} \end{aligned}$$

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$\begin{aligned} p(r|o) &= \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6 \\ &= \frac{p(o|r)p(r)}{p(o)} \\ &= \frac{5/7 \cdot 7/12}{6/12} = 5/6 \end{aligned}$$



Medical test

A medical test for a given disease

- Correctly identifies the disease 99% of the time (true positives), and
- Incorrectly turns out positive 2% of the time (false positives).

You know that

- 1% of the population suffers from the disease.

You go to the doctor to get tested, and the test turns out to be positive.

What is the probability you have the disease?

Hints:

- Identify from the text: ($x=Positive$,
 $y=0: no\ disease$, $y=1: Disease$)

$p(Positive|Disease)$

$p(Positive|No\ Disease)$

$p(Disease)$

$p(No\ Disease)$

- Use the basic rules of probability given to the right to find:

$p(Disease|Positive)$

$$p(+|D) = 99\%$$

$$p(+|!D) = 2\%$$

$$p(D) = 1\%$$

$$\rightarrow p(D|+) = \frac{p(+|D)*p(D)}{p(+|D)*p(D)+p(+|!D)*p(D)} = \frac{p(+|D)*p(D)}{p(+|D)*p(D)+p(+|!D)*p(D)}$$

$$p(y) = \sum_x p(y, x)$$

$$= p(y|x)p(x) + p(y|\bar{x})p(\bar{x})$$

$$p(x, y) = p(x|y)p(y)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Quiz 1, Probabilities (Spring 2014)

Consider a dataset which describe the consumption of delicatessen products in different cities. Each observation in the dataset is a customer, and we record the city the customer is from as well as their consumption of delicatessen. Suppose you are told:

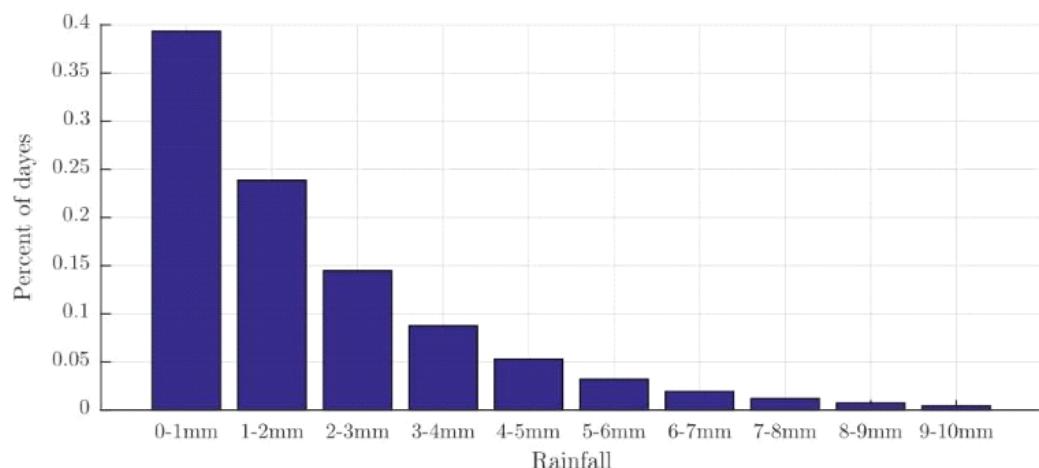
- 17.5 % were from Lisbon, 10.7 % were from Oporto and 71.8 % from the Other region.
- 44.1 % of the costumers from Lisbon spent above the median consumption on delicatessen (DELI).
- 48.9 % of the costumers from Oporto spent above the median consumption on delicatessen (DELI).
- 51.6 % of the costumers from the Other region spent above the median consumption on delicatessen (DELI).

What is the probability based on the wholesale data that a costumer that spent above the median consumption on delicatessen (DELI) come from Lisbon?

- A. 7.7 %
- B. 15.4 %
- C. 44.1 %
- D. 59.6 %
- E. Don't know.

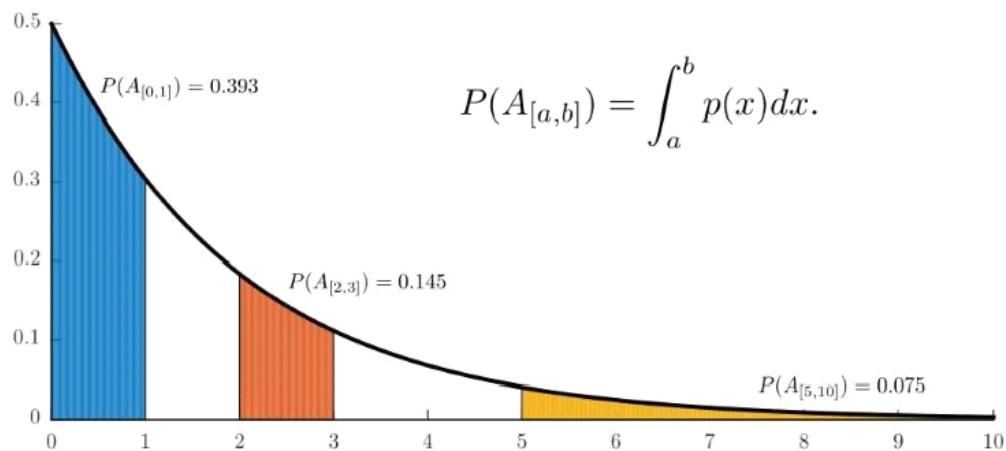
Probability vs. Density

- Suppose we consider the rainfall on an average day r
- **Can't** talk about the probability there will be **exactly** $r=2.3$ mm of rain, $P(r=2.3\text{mm})$
- **Can** talk about the probability there will be **between** 1 and 2 mm of rain



Probability vs. Density

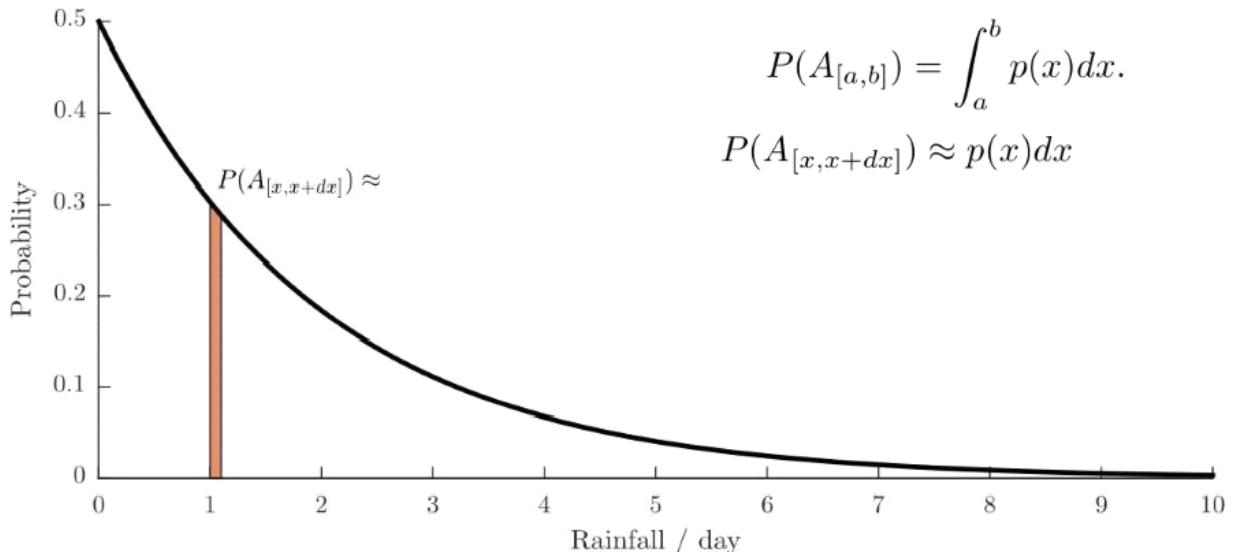
- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the **curve** is the **density**



$A_{[a,b]}$: There will be between a and b mm of rain

Probability vs. Density

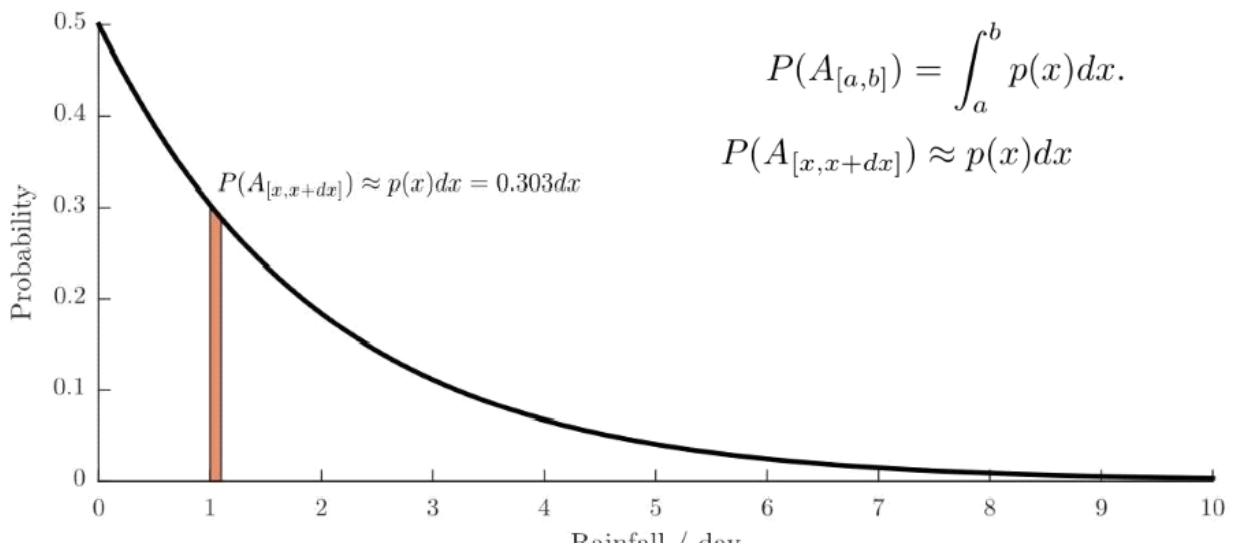
- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the **curve** is the **density**
- What is the probability there will be between 1 and 1.1 mm of rain?

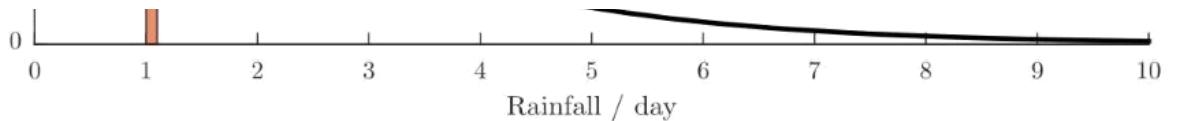


$A_{[a,b]}$: There will be between a and b mm of rain

Probability vs. Density

- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the **curve** is the **density**
- What is the probability there will be between 1 and 1.1 mm of rain?





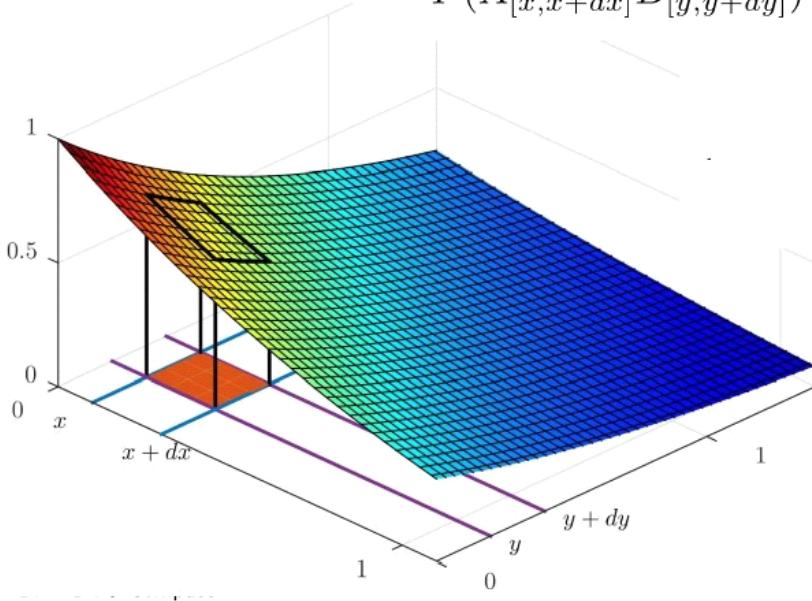
$A_{[a,b]} : \text{There will be between } a \text{ and } b \text{ mm of rain}$

Probability vs. Density

- For two variables x and y , the **probability** is an integral over an **area**

$$P((x, y) \in D) = \int_{(x,y) \in D} p(x, y) dx dy$$

$$\widehat{P}(A_{[x,x+dx]} B_{[y,y+dy]}) =$$



This implies:

$$p(x, y) = p(y|x)p(x)$$

$$\int_{(x,y) \in D}$$

$$P(A_{[x,x+dx]} B_{[y,y+dy]}) = P(A_{[x,x+dx]} | B_{[y,y+dy]}) P(B_{[y,y+dy]})$$

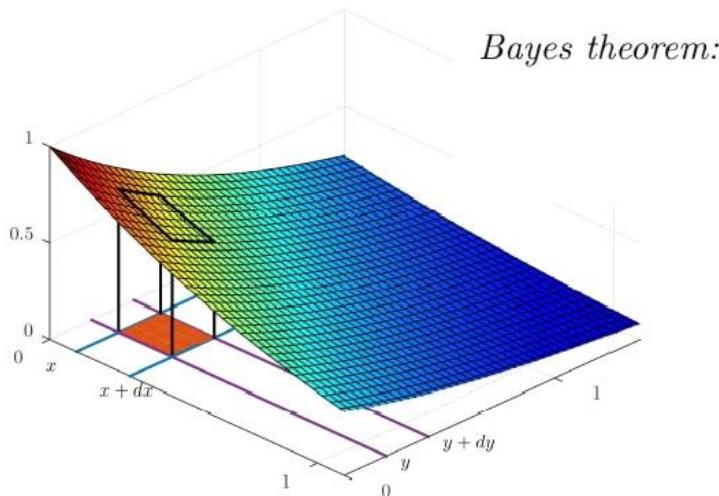
$$p(x, y) dx dy = p(x|y) dx p(y) dy$$

Probability vs. Density

- Thus, we have shown the rules of probability theory also holds for densities

$$\text{The sum rule: } \int dx p(x|z) = 1$$

$$\text{The product rule: } p(x, y|z) = p(y|x, z)p(x|z)$$



$$\begin{aligned} p(x|y, z) &= \frac{p(y|x, z)p(x|z)}{p(y|z)} \\ &= \frac{p(y|z)p(x|y, z)}{\int p(y|x', z)p(x'|z)dx'}. \end{aligned}$$

Collecting all of this we obtain:

- Rules of probability for densities

$$\text{Marginalization: } \int p(x, y|z)dx = p(y|z)$$

$$\text{The product rule: } p(x, y|z) = p(y|x, z)p(x|z)$$

$$\text{Bayes theorem: } p(x|y, z) = \frac{p(y|x, z)p(x|z)}{\int p(y|x', z)p(x'|z)dx'}.$$

- Rules of probability for discrete variables

$$\text{Marginalization: } \sum_c p(x = c, y|z) = p(y|z)$$

$$\text{The product rule: } p(x, y|z) = p(y|x, z)p(x|z)$$

$$\text{Bayes theorem: } p(x|y, z) = \frac{p(y|x, z)p(x|z)}{\sum_c p(y|x = c, z)p(x = c|z)}.$$

Expected values

- Discrete random variable

$$\mathbb{E}[g] = \sum_i g(x_i)P(x_i)$$

- Continuous random variable

$$\mathbb{E}[g] = \int_{-\infty}^{\infty} g(x)p(x)dx$$

Statistics

- **Mean**

$$\bar{x} = \mathbb{E}[x]$$

- **Covariance**

$$\text{cov}(x, y) = \mathbb{E}[(x - \bar{x})(y - \bar{y})]$$

- **Variance**

$$\text{var}(x) = \text{cov}(x, x) = \mathbb{E}[(x - \bar{x})^2]$$

- **Standard deviation**

$$\text{std}(x) = \sqrt{\text{var}(x)}$$

Densities and models

- In machine learning, we want to learn a parameter from data
- Models of the data which use parameters are how we do that
- We build models out of simpler building blocks (distributions (discrete variables see chapter 5) and densities (continuous variables see chapter 6)). In this course we will use four:

Bernoulli distribution

The Categorical distribution

The Beta density

The Multivariate normal density

The multivariate normal distribution

A distribution for M -dimensional vectors \mathbf{x} :

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{M}{2}} \sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$
$$M = 1 : \quad \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

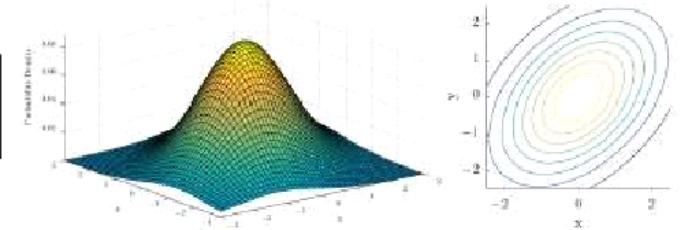
$\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix:

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}], \quad \Sigma_{ij} = \text{cov}[x_i, x_j]$$

- Example: 2-dimensional Normal distribution

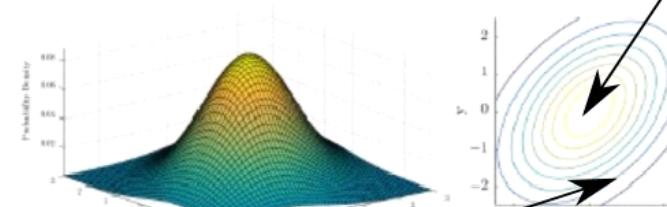
$$\boldsymbol{\mu} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{bmatrix}$$



Quiz 2, Covariance

- Match the covariances to the contour plots



$$\Sigma = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{bmatrix}$$

$$\mu = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

A. Covariance of A is $\Sigma_A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\Sigma_C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

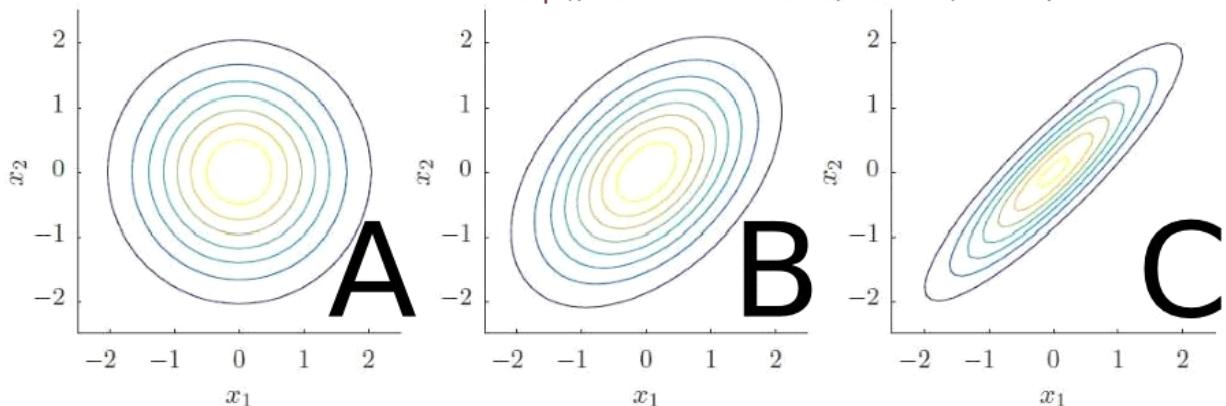
B. $\Sigma_B = \begin{bmatrix} 1 & -0.45 \\ 0.45 & 1 \end{bmatrix}$, $\Sigma_C = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$

C. $\Sigma_B = \begin{bmatrix} 10 & 4.5 \\ 4.5 & 10 \end{bmatrix}$, $\Sigma_C = \begin{bmatrix} 10 & 9 \\ 9 & 10 \end{bmatrix}$

D. $\Sigma_A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\Sigma_C = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$

E. Don't know.

Check out the online demo <http://www2.imm.dtu.dk/courses/02450/DemoNormal.html>



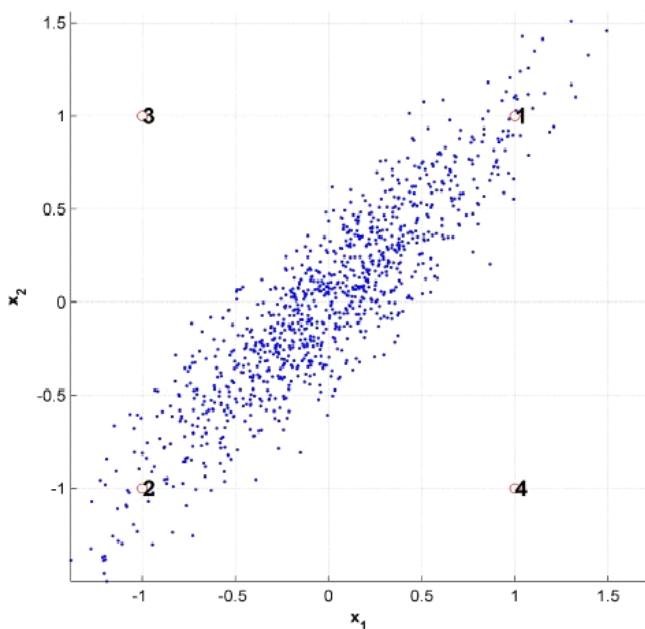
The Mahalanobis distance

How far are x_1 and x_2 apart?

- $\text{mahalanobis}(x_1, x_2) = 4.2$
- $d_{\text{Euclidean}}(x_1, x_2)^2 = 8.0$

How far are x_3 and x_4 apart?

- $\text{mahalanobis}(x_3, x_4) = 80$
- $d_{\text{Euclidean}}(x_3, x_4)^2 = 8.0$



$$\text{mahalanobis}(x, y)^2 = (x - y)^\top \Sigma^{-1} (x - y)$$

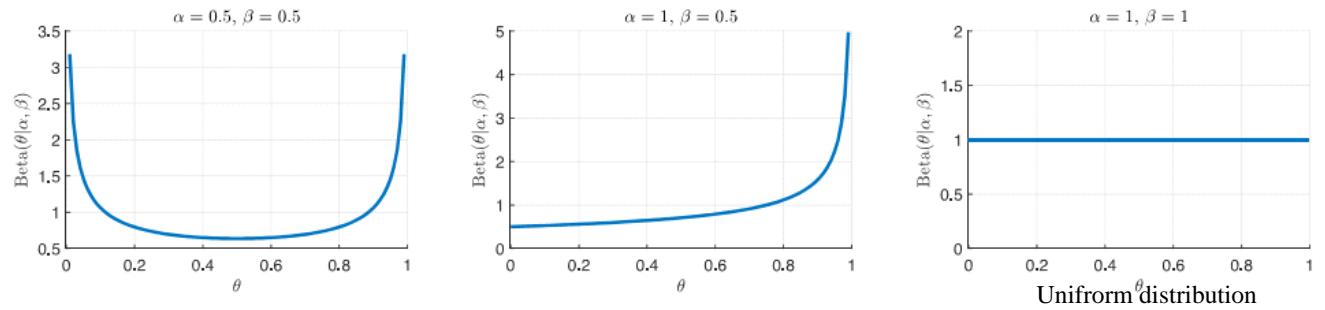
$$\text{mahalanobis}(x, y)^2 = (x - y)^\top \Sigma^{-1} (x - y)$$

$$d_{\text{euclidian}}(x, y)^2 = (x - y)^\top I^{-1} (x - y)$$

Beta distribution

Suppose θ is defined on the unit interval $[0, 1]$

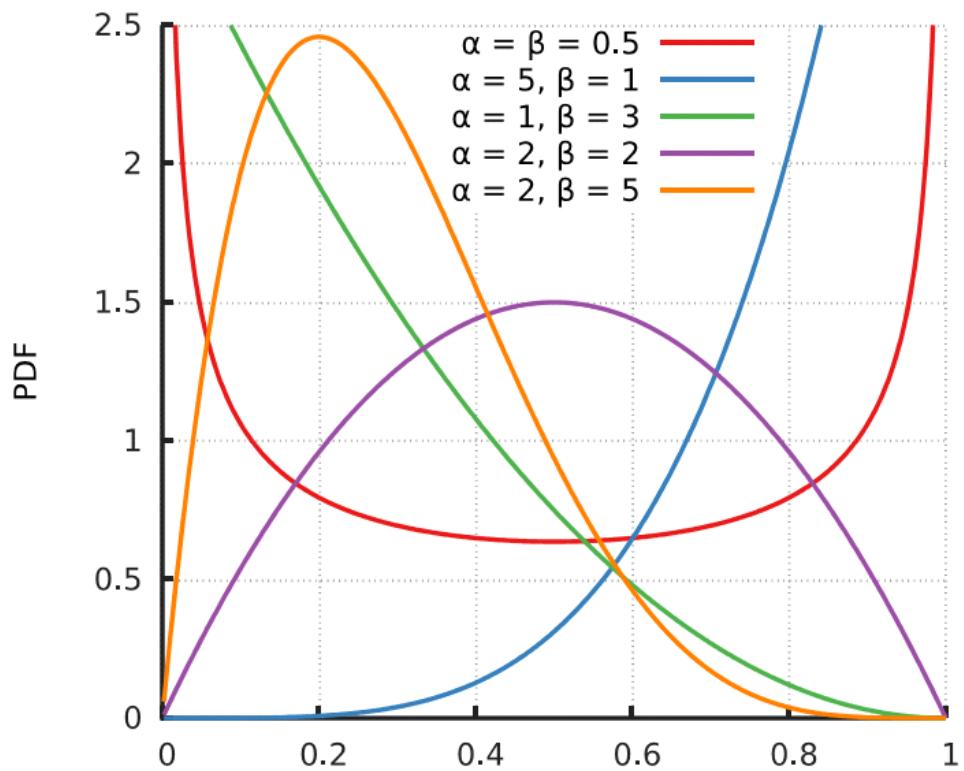
Beta density: $p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$.



$\alpha, \beta > 0$ are related to the variance and mean

$$\mathbb{E}_{p(\theta|\alpha, \beta)}[\theta] = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}_{p(\theta|\alpha, \beta)}[\theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Beta distribution



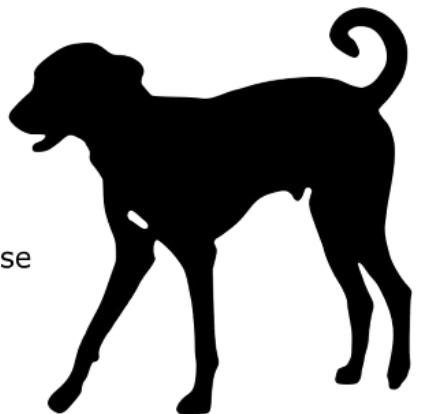
Picture from Wikipedia

24 DTU Compute

Lecture 4 21 September, 2021

Probabilities and learning

- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?



- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?



Intuition tells us the answers are different, but the situation seems similar...

Recall from lecture 3: The Bernoulli distribution

- Suppose a coin come up heads with probability θ
 - Suppose $b = 1$ is the event the coin land heads
 - $b = 0$ is the event the coin land tails
- The density is given by the **Bernoulli distribution**

$$p(b|\theta) = \theta^b(1-\theta)^{1-b}$$

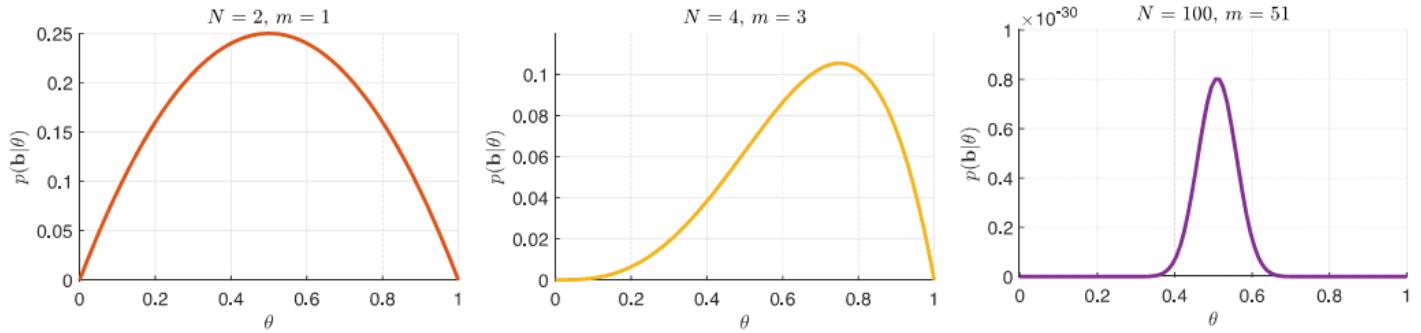
- For a sequence of N flips b_1, b_2, \dots, b_N
Model of the process of flipping a coin

Independence

$$\begin{aligned} p(b_1, \dots, b_N | \theta) &= \prod_{i=1}^N p(b_i | \theta) \\ &= \theta^{\sum_{i=1}^N b_i} (1-\theta)^{N - \sum_{i=1}^N b_i} \\ &= \theta^m (1-\theta)^{N-m}, \quad m = b_1 + b_2 + \dots + b_N \end{aligned}$$

- What is θ ?

Recall from lec. 3: Bernoulli distribution, maximum likelihood



$$p(b_1, \dots, b_N | \theta) = \theta^m (1 - \theta)^{N-m}$$

An idea for selecting θ^* is **Maximum likelihood**

Choose theta that makes the data the most likely

$$\theta^* = \arg \max_{\theta} p(b_1, \dots, b_N | \theta)$$

The value of θ according to which the data is most plausible

The Bernoulli distribution

- A magic coin is a coin that comes up heads with probability θ
 - Suppose $b = 1$ is the event the coin land heads
 - $b = 0$ is the event the coin land tails
- For a sequence of N flips b_1, b_2, \dots, b_N

$$p(b_1, \dots, b_N) = \theta^m(1 - \theta)^{N-m}, \quad m = b_1 + b_2 + \dots + b_N$$

- **What is θ ?** Answer: **Use Bayes' Theorem!**

$$p(\theta|\mathbf{b}) =$$

- Assume $p(\theta) =$

$$p(\theta|\mathbf{b}, \alpha, \beta) =$$

The Bernoulli distribution

- A magic coin is a coin that comes up heads with probability θ
 - Suppose $b = 1$ is the event the coin land heads
 - $b = 0$ is the event the coin land tails
- For a sequence of N flips b_1, b_2, \dots, b_N

b is a sequence of event

$$p(b_1, \dots, b_N) = \theta^m(1 - \theta)^{N-m}, \quad m = b_1 + b_2 + \dots + b_N$$

- **What is θ ?** Answer: **Use Bayes' Theorem!**

Theta is the parameter I want
To know

$$p(\theta|b) = \frac{p(b|\theta)p(\theta)}{p(b)} = \frac{p(b|\theta)p(\theta)}{\int_0^1 p(b|\theta')p(\theta')d\theta'}$$

Theta' because we are modelizing theta inside the integral

- Assume $p(\theta) =$

$$p(\theta|b, \alpha, \beta) =$$

The Bernoulli distribution

- A magic coin is a coin that comes up heads with probability θ
 - Suppose $b = 1$ is the event the coin land heads
 - $b = 0$ is the event the coin land tails
- For a sequence of N flips b_1, b_2, \dots, b_N

$$p(b_1, \dots, b_N) = \theta^m(1 - \theta)^{N-m}, \quad m = b_1 + b_2 + \dots + b_N$$

- **What is θ ?** Answer: **Use Bayes' Theorem!**

$$p(\theta|\mathbf{b}) = \frac{p(\mathbf{b}|\theta)p(\theta)}{p(\mathbf{b})} = \frac{p(\mathbf{b}|\theta)p(\theta)}{\int_0^1 p(\mathbf{b}|\theta')p(\theta')d\theta'}$$

- Assume $p(\theta) = p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$p(\theta|\mathbf{b}, \alpha, \beta) =$$

The Bernoulli distribution

- A magic coin is a coin that comes up heads with probability θ
 - Suppose $b = 1$ is the event the coin land heads
 - $b = 0$ is the event the coin land tails
- For a sequence of N flips b_1, b_2, \dots, b_N

$$p(b_1, \dots, b_N) = \theta^m(1 - \theta)^{N-m}, \quad m = b_1 + b_2 + \dots + b_N$$

- **What is θ ?** Answer: **Use Bayes' Theorem!**

$$p(\theta|\mathbf{b}) = \frac{p(\mathbf{b}|\theta)p(\theta)}{p(\mathbf{b})} = \frac{p(\mathbf{b}|\theta)p(\theta)}{\int_0^1 p(\mathbf{b}|\theta')p(\theta')d\theta'}$$

- Assume $p(\theta) = p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$p(\theta|\mathbf{b}, \alpha, \beta) = \frac{\theta^m(1-\theta)^{N-m} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \theta'^m(1-\theta')^{N-m} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta'^{\alpha-1}(1-\theta')^{\beta-1}d\theta'}$$

Alpha and beta are into the equation in the same way N and m are

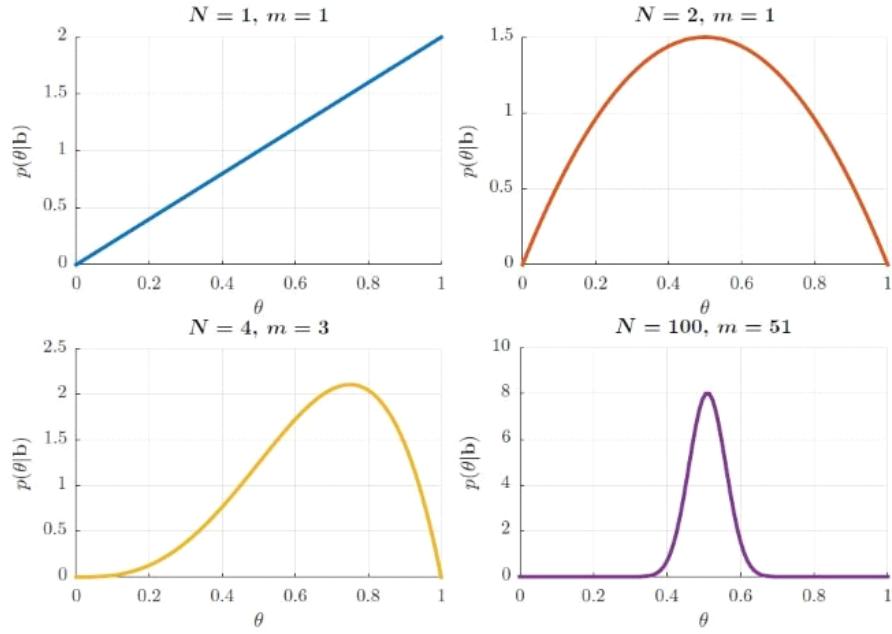
$$= \frac{\Gamma(\alpha + \beta + N)}{\Gamma(\alpha + m)\Gamma(\beta + N - m)}\theta^{\alpha+m-1}(1-\theta)^{\beta+N-m-1}$$

This is a BETA distribution

--> I assume I didn't see anything: it's a scaled version of the one shown before

Example: $\alpha = \beta = 1$

$$\begin{aligned} p(\theta|\mathbf{b}, \alpha, \beta) &= \frac{\Gamma(\alpha + \beta + N)}{\Gamma(\alpha + m)\Gamma(\beta + N - m)} \theta^{\alpha+m-1} (1-\theta)^{\beta+N-m-1} \\ &= \frac{(N+1)!}{m!(N-m)!} \theta^m (1-\theta)^{N-m} \end{aligned}$$



Dogs and coins



- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?



- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?

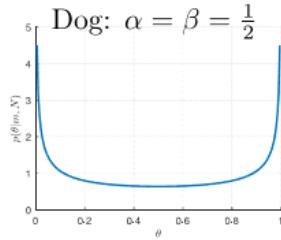
Dogs and coins



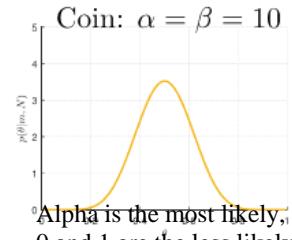
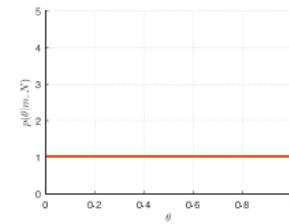
- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

Prior

$$p(\theta|\alpha, \beta) =$$



- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?



Alpha is the most likely, and 0 and 1 are the less likely

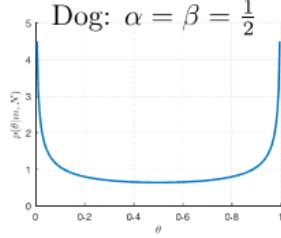
Dogs and coins



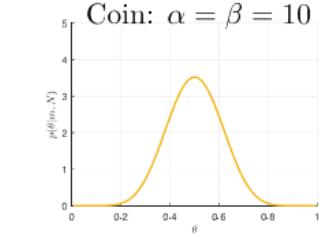
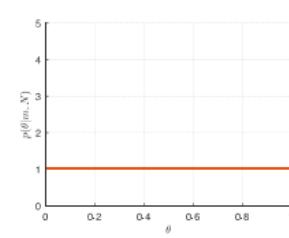
- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

Prior

$$p(\theta|\alpha, \beta) =$$



- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?



Likelihood

$$p(m = 4, N = 4|\theta) = \theta^m(1 - \theta)^{N-m} = \theta^4$$

Alpha and beta specify my prior knowledge: if $\alpha = 0$ and $\beta = 0$, before flipping a coin, I assume that coin are fair (central image) that all the values are equally likely. LEFT and RIGHT are the prior updated based on the data.

A priori i thing coin are fair, but then based on the likelihood, a posteriori, the probability of flipping a coin change.

A priori i thing coin are fair, but then based on the likelihood, a posteriori, the propability of flipping a coin change.



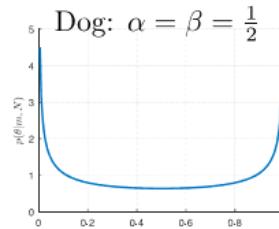
Dogs and coins



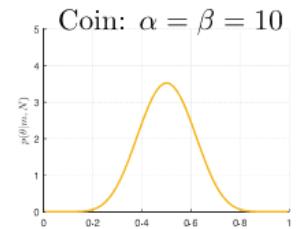
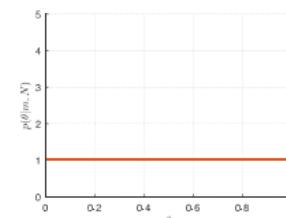
- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

Prior

$$p(\theta|\alpha, \beta) =$$



- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?



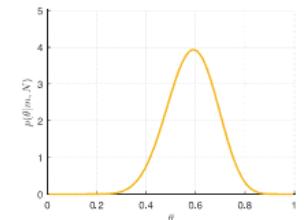
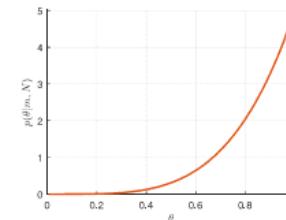
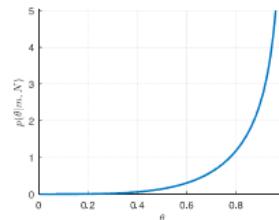
Likelihood

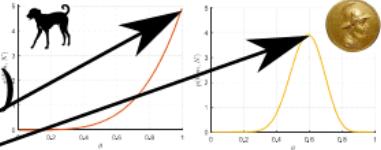
$$p(m=4, N=4|\theta) = \theta^m(1-\theta)^{N-m} = \theta^4$$

The difference between the two cases is that we have prior knowledge which tell us most coins are fair, and this affects our conclusions.
In most practical situations, we should assume as little as possible and choose $\alpha = \beta = \frac{1}{2}$

Posterior

$$p(\theta|m, N) = \frac{p(m, N|\theta)p(\theta|\alpha, \beta)}{p(m, N)} =$$





Learning principle: Maximum a posteriori (MAP)

- Another idea: Select θ which is "most probably"

$$\theta^* = \arg \max_{\theta} p(\theta|M, N) = \arg \max_{\theta} \left[\frac{p(m, N|\theta)p(\theta|\alpha, \beta)}{p(M, N)} \right]$$

- Use that $\arg \max_x f(x) = \arg \min_x [-\log f(x)]$ if $f(x) > 0$:

$$\theta^* = \arg \min_{\theta} \left[-\log \frac{p(m, N|\theta)p(\theta|\alpha, \beta)}{p(m, N)} \right]$$

(likelihood)
 $p(m, N|\theta) = \theta^m(1-\theta)^{N-m}$

(prior)
 $p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

- All in all:

$$\theta^* = \arg \min_{\theta} E(\theta), \quad E(\theta) = -\log p(m, N|\theta) - \log p(\theta|\alpha, \beta)$$

E(theta) is what so called COST function

Maximum a posteriori (MAP) learning

- Consider some data $\mathbf{X}_{\text{data}} = \mathbf{x}_1, \dots, \mathbf{x}_N$ and $\mathbf{y}_{\text{observation}} = y_1, \dots, y_N$
- Suppose we think x_i relates to y_i by some parameters θ
- Assume**

Observations are not informative about each other when we know parameters

- Then

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \mathbf{w}), \quad p(\mathbf{w}|\mathbf{X}) = p(\mathbf{w})$$

Without \mathbf{y} , we cannot learn the parameters

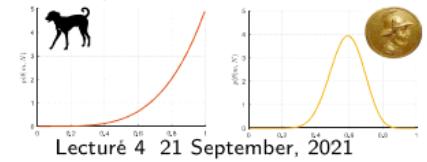
- The following are equivalent:

$$\text{Maximize: } \mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{y})$$

$$\text{Minimize: } \mathbf{w}^* = \arg \min_{\mathbf{w}} E(\mathbf{w}), \quad E(\mathbf{w}) = \left[\frac{1}{N} \sum_{i=1}^N -\log p(y_i|\mathbf{x}_i, \mathbf{w}) \right] - \frac{1}{N} \log p(\mathbf{w})$$

- All we need is a likelihood (**usually pretty simple**) and a prior (**can be omitted**) and we have a machine-learning method.

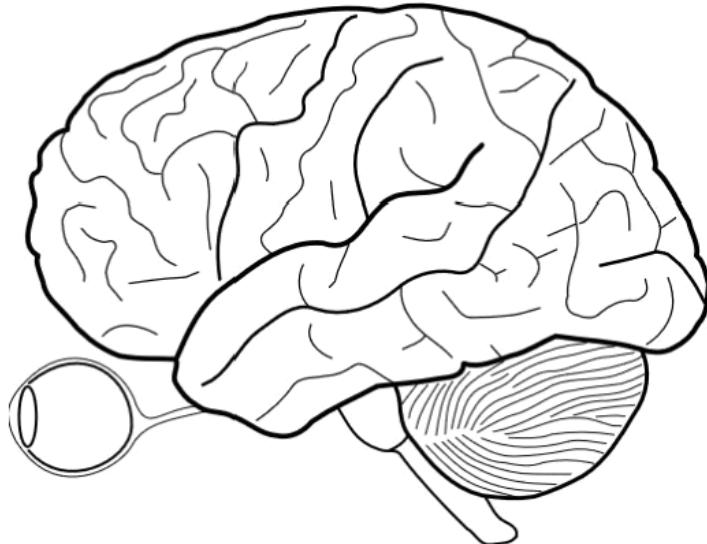
- Pro:** Easy, conceptually simple, efficient
- Con:** Can sometimes give spurious results (overfit)



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

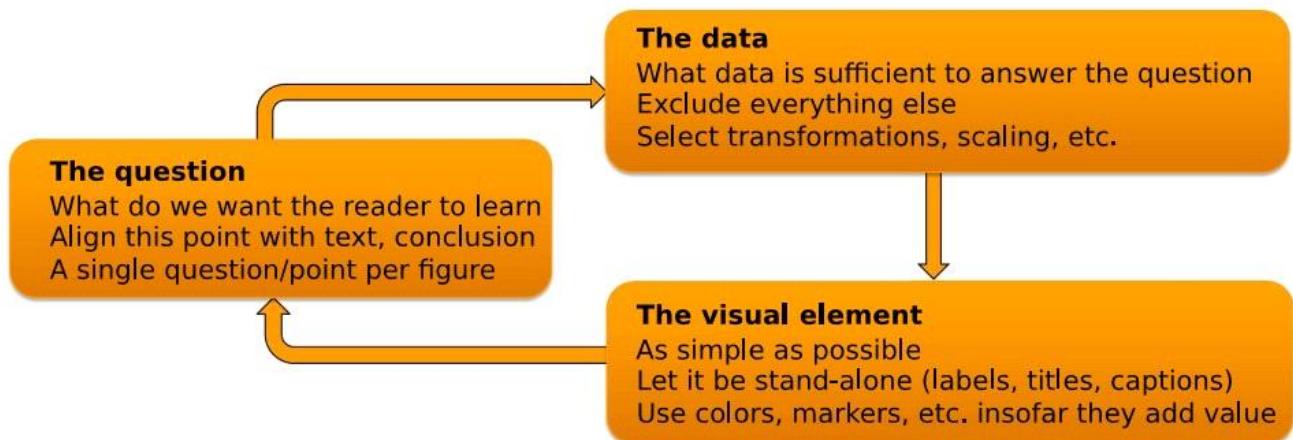
Visualization

- 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 exaggerate. (*vs. distort data in a plot*)

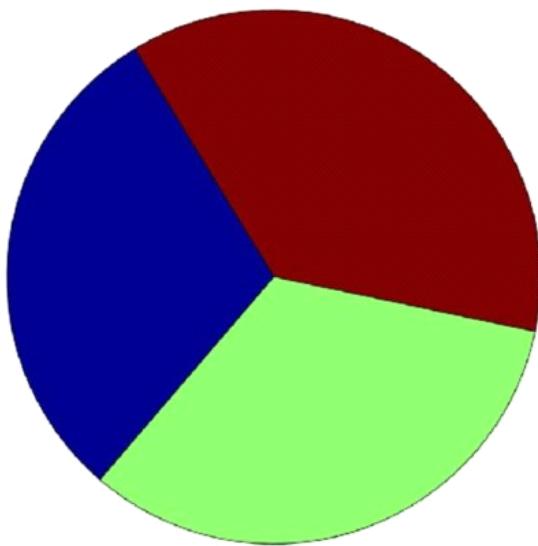


Important choices for visualizations

- **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

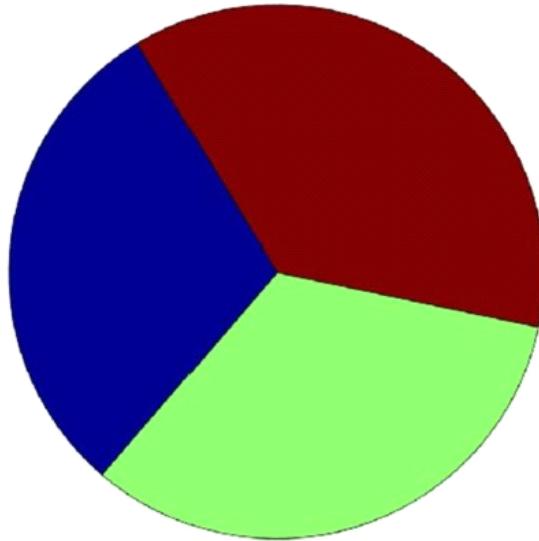
- **Area represents proportion**
 - Which is smallest, middle, and largest?
 - What are the proportions approximately?



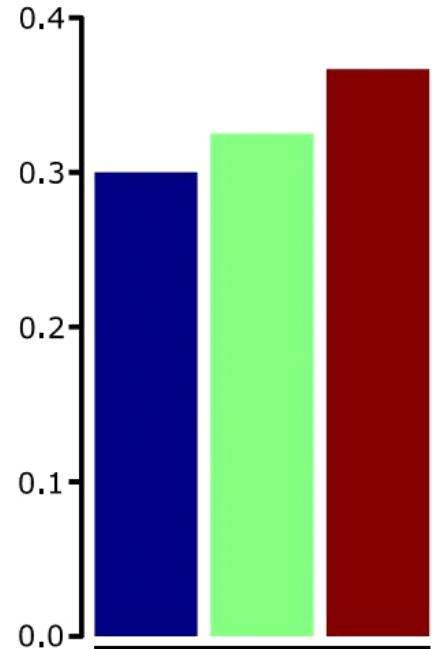
Representation

- **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
 - **How?**
 - 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 between attributes**

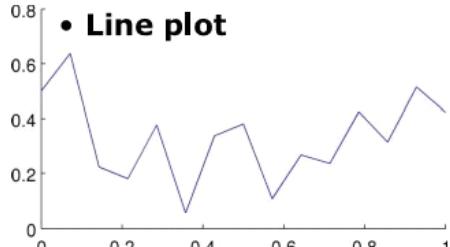
- 2D histogram
- Heat maps and contour plots
- Scatter plots

- **Visualization of high-dimensional objects**

- Matrix plots
- Parallel coordinates
- Star plots

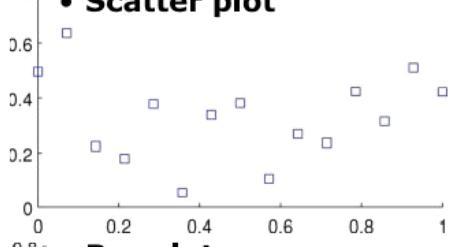
Basic plots

• Line plot



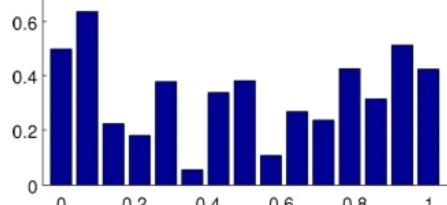
```
plot(x, y);
```

• Scatter plot



```
plot(x, y, 's');  
scatter(x, y, 's')
```

• Bar plot



```
bar(x, y);
```

The iris data set

- **Three 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

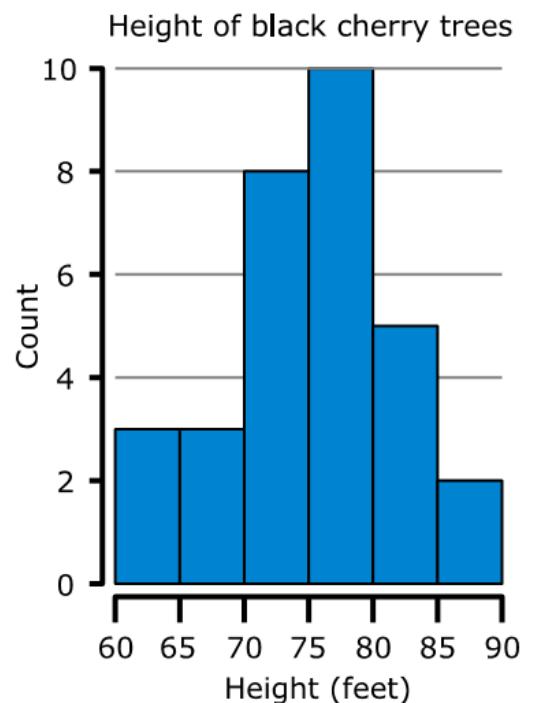
Flower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
...
...
150	5.9	3.0	5.1	1.8

X ^{Observation x Attribute}

Distribution of a single attribute

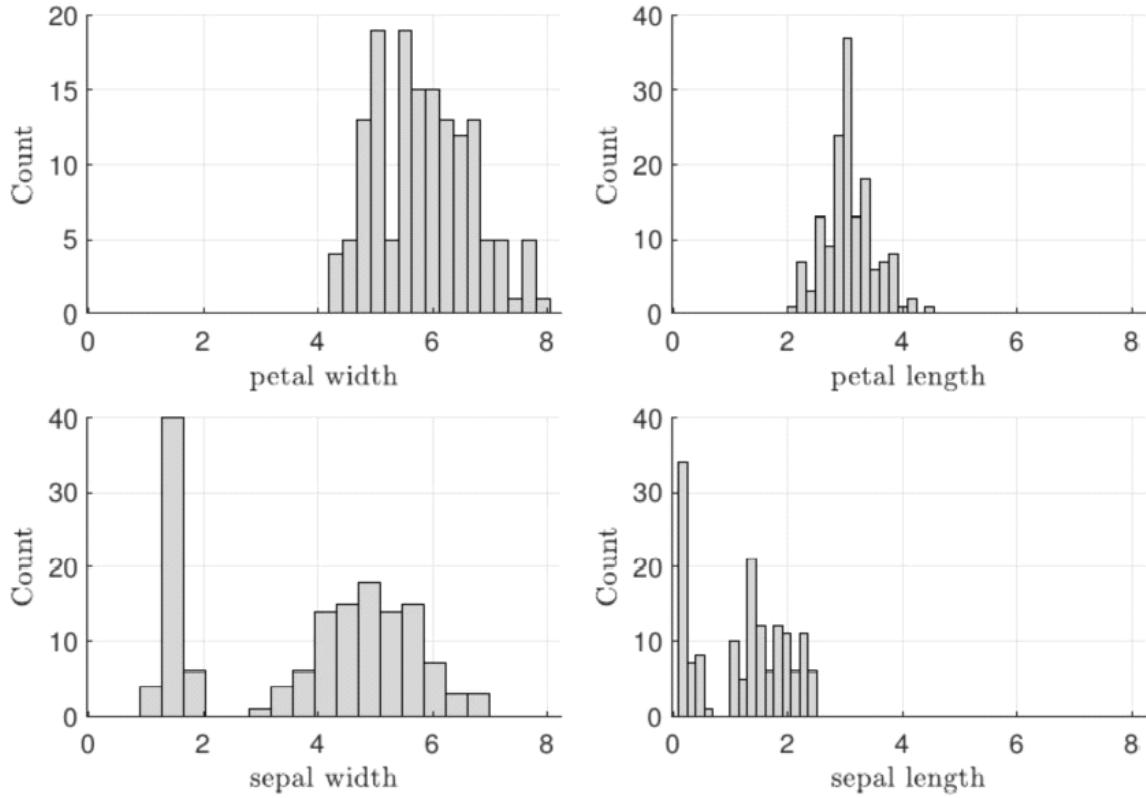
Histograms

- **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

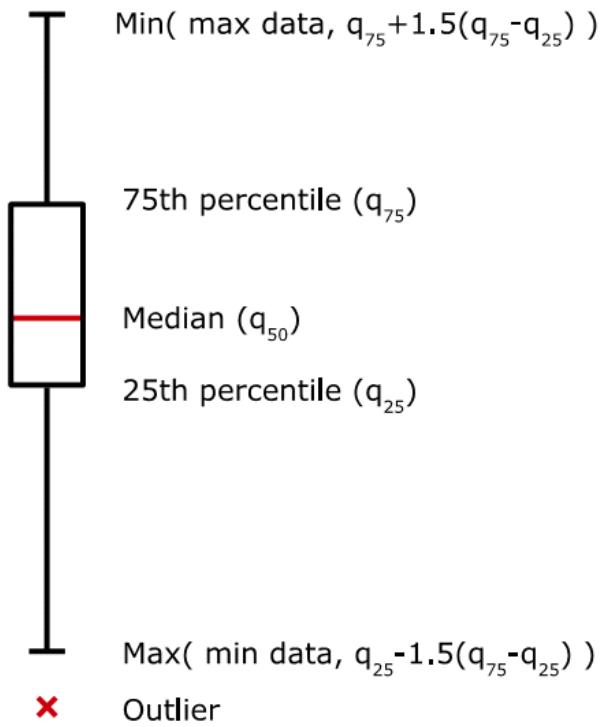


$H = \{60, 64, 64, 66, 67, 69, 71, 72, 72, 72, 72, 73, 74, 74, 75, 75, 76, 76, 76, 77, 77, 78, 78, 79, 80, 80, 81, 82, 84, 85, 85, 89\}$

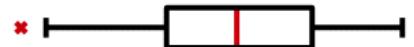
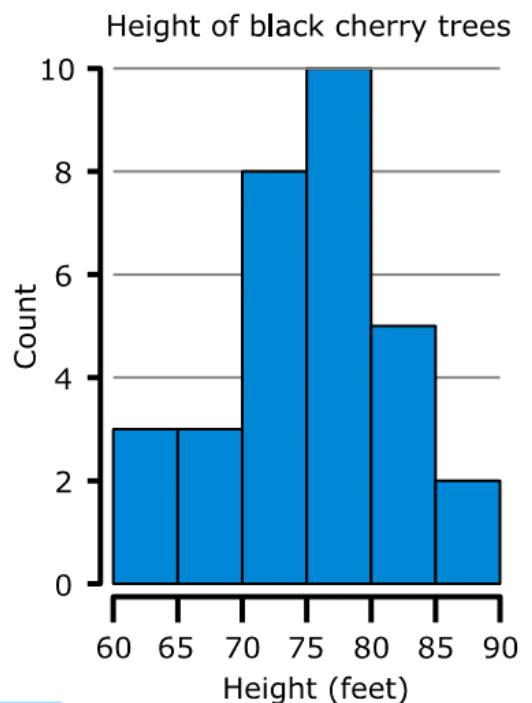
Histograms of the Iris data attributes



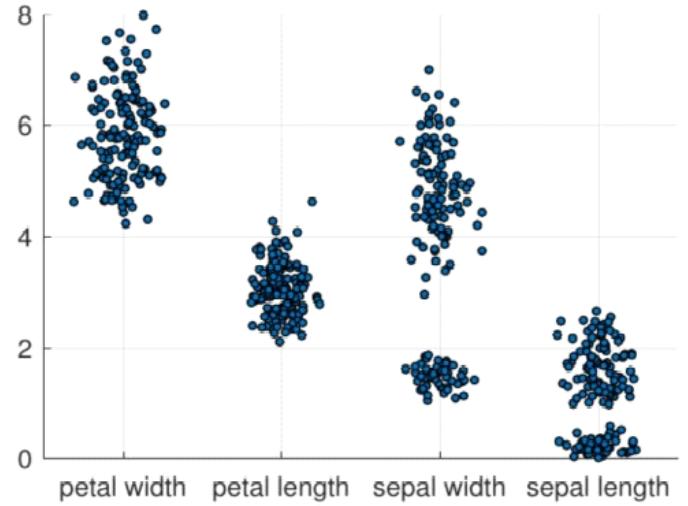
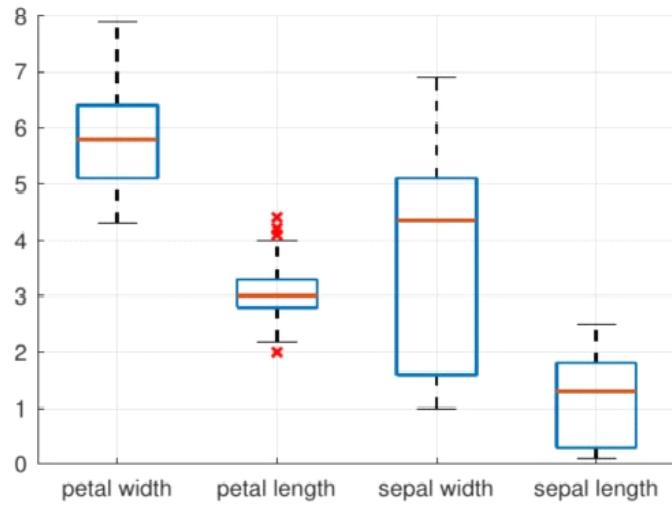
Box plots



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



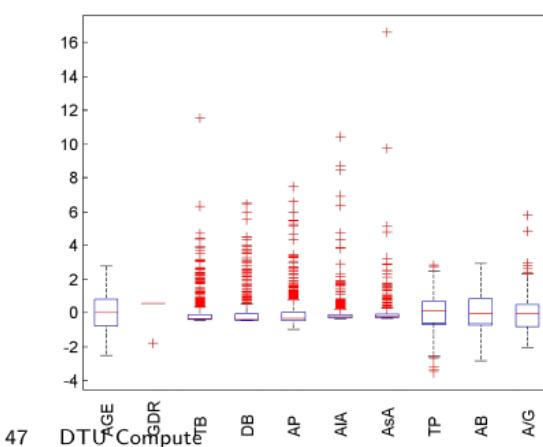
Box plots



Quiz 3, Boxplots (Fall 2012)

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.



47

DTU Compute

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*?

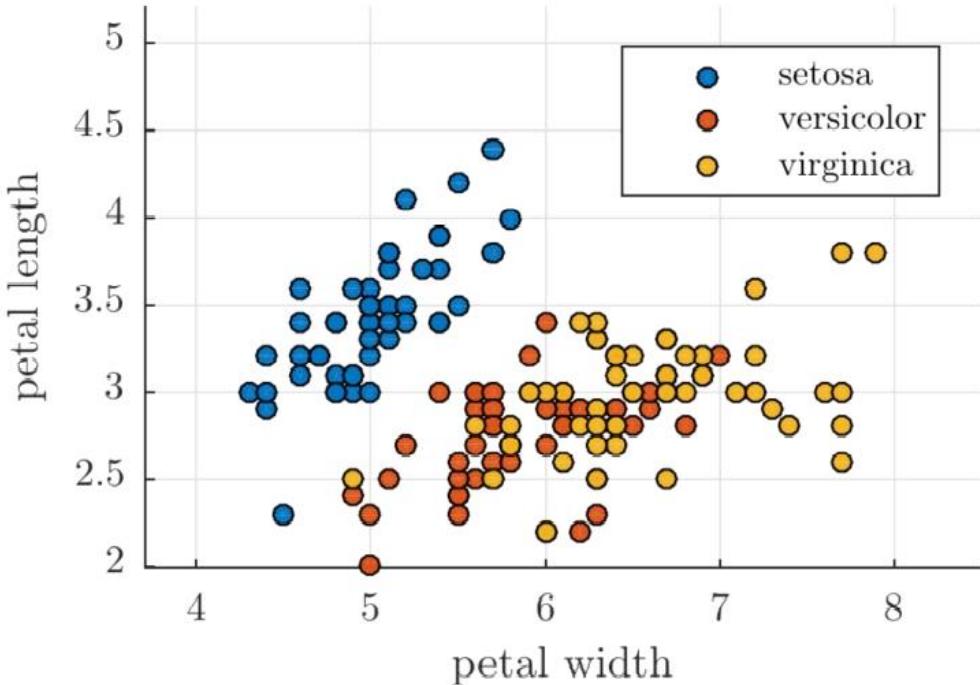
- A. The value of the 50th and 75th percentiles of the attribute DB coincides.
- B. Even though the distribution of AlA and AsA may have a similar shape this does not imply that the two attributes are correlated.
- C. The attribute TB is likely to be normal distributed.
- D. The attribute GDR has a clear outlier that should be removed.
- E. Don't know.

Lecture 4 21 September, 2021

Relation between attributes

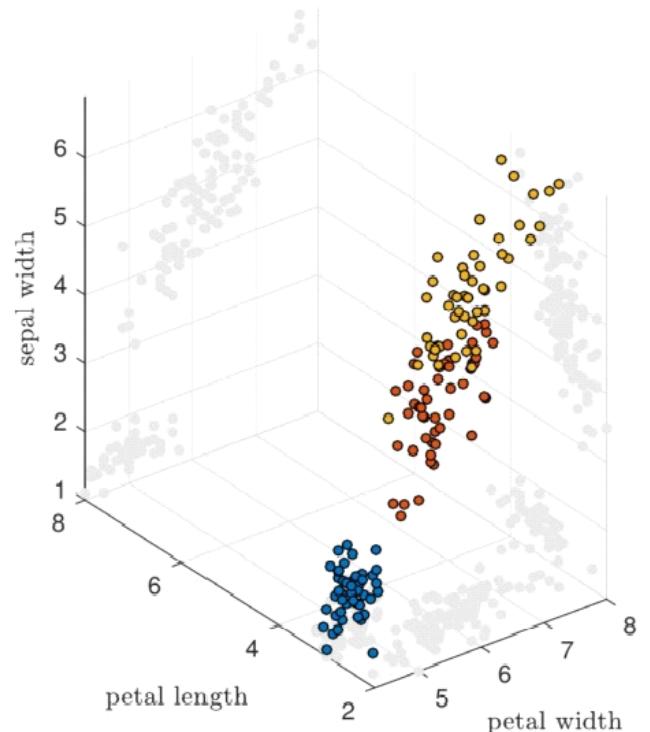
Scatter plots

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



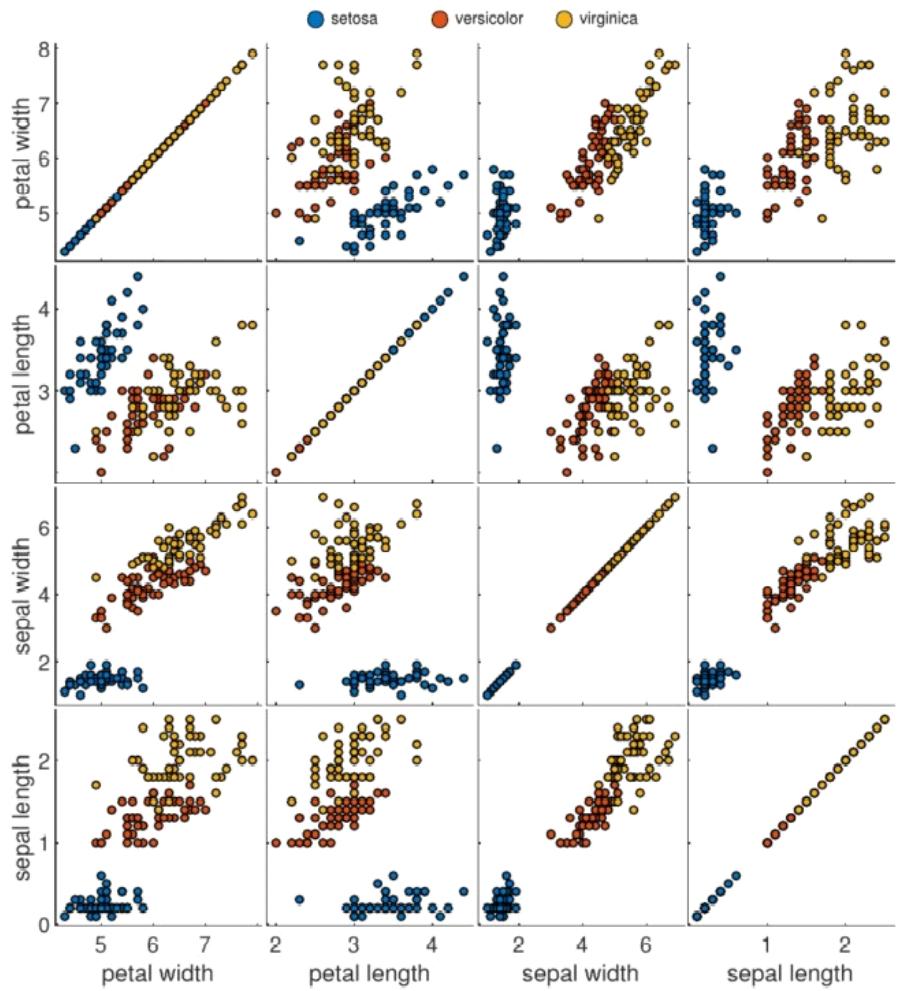
Scatter plots

- Shows **relation** 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**

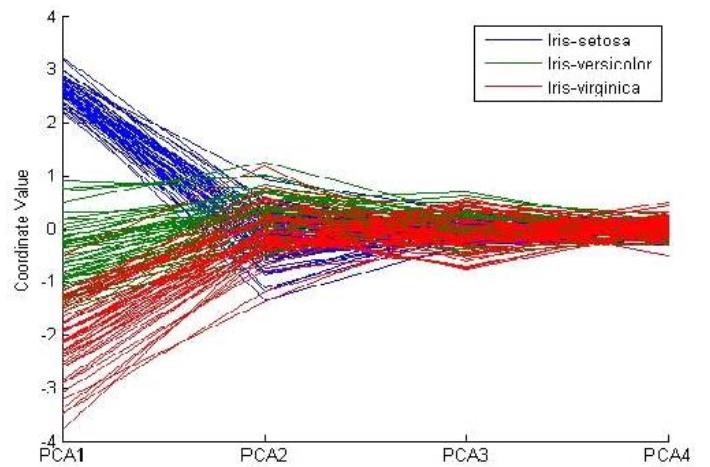
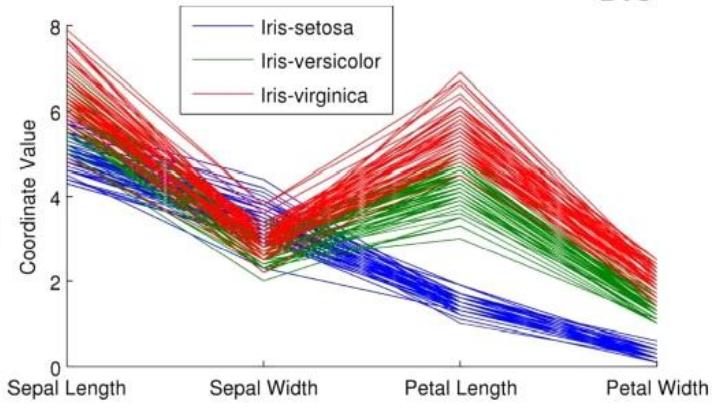
- Useful for visualizing the relation between objects



Lecture 4 21 September, 2021

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:

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



Edward Tufte

https://commons.wikimedia.org/wiki/File:Edward_Tufte_-_cropped.jpg
Made available by Keegan Peterzell

Lecture 4 21 September, 2021

Making good data visualizations is an art

For some interesting data visualizations see also

http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html

<http://www.informationisbeautiful.net/>

<http://www.junkcharts.typepad.com/>

Resources

<https://www.khanacademy.org> An excellent introduction to probability theory which we recommend as a go-to resource

(<https://www.khanacademy.org/math/statistics-probability/probability-library>)

<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>)



02450: Introduction to Machine Learning and Data Mining

Decision trees and linear regression

Tommy Sonne Alstrøm

DTU Compute, Technical University of Denmark (DTU)

DTU Compute

Department of Applied Mathematics and Computer Science

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

$\Delta \int_a^b \mathcal{E} \Theta + \Omega \delta e^{i\pi} = -1$

$\infty = \{2.7182818284\}_{\text{euler}}$

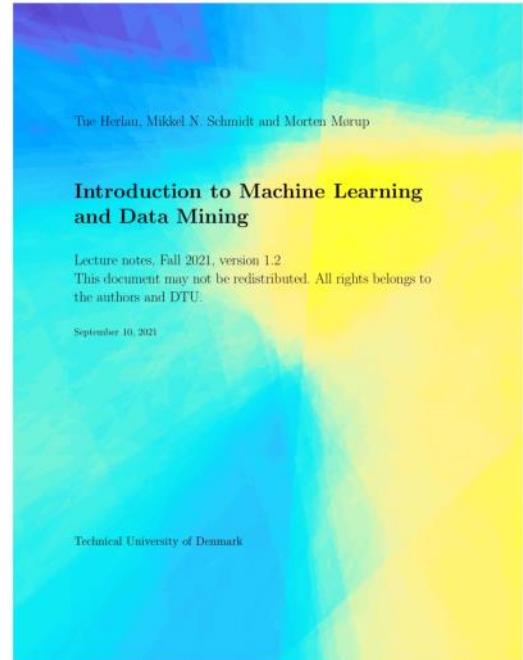
$\chi^2 \sum! \gg,$

Today

Feedback Groups of the day:

Alex Abades Grimes, Alexander Samuel Bendix Gosden, Andreas Halkjær-Knudsen, Andreas Handberg, Andreas Lyhne Fiehn, Andreu Guitart Gendre, Anna Aisha Goedhart, Anna Josefina Grillenberger, Anne Kargaard Gjelstrup, Antonio Ferrara, Benjamin Kock Fazal, Carmen Gomez Delicado, Clara Giarrusso, Daniel Grathwol, Elvis Antônio Ferreira Camilo, Emil Grovn, Gabriel Benjamin Bendix Gosden, Herdís Rún Halldórsdóttir, Hildur Margret Gunnarsdóttir, Jacopo Gaetani, Jingrui Ge, Johan Böcher Hanehøj, Johanne Christine Arboe Franck, Jonas Roslev Gjerris, Konstantinos Grammenidis, Lilian Mette Geese, Lucie Fontaine, Mackenzie Guynn, Magnus Frederiksen, Matteo Fineschi, Mattia Guarneri, Merkourios Giannikos, Mikkel Feng Frederiksen, Mário Gazo, Nikolai Gulbrandsen, Orhun Görkem, Paolo Federico, Peter Patrick Kjølstad Frederiksen, Pálína Kröyer Guðmundsdóttir, Ramon Fortuny Cuartielles, Rasmus Lyhne Fiehn, Rasmus Scott Grunnet-Jepsen, Rebeca González Revilla, Sebastian Nicolai Fabricius Grut, Sándor Földi, Theodor Peter Guttesen, Tine Preston Frederiksen, Tobias Tandrup Vinding Granell, Varun Ghatrau, Victor André Marcel Guerin, Victor Girardin Flindt, Victor Hamel, Viktor Helgi Gizurarson, Xingguang Geng, Yeray Garcia Concejero, Zeyneb Hady, Zhijian Feng

Reading material: Chapter 8, Chapter 9



Lecture 5 28 September, 2021

Lecture Schedule

1 Introduction

31 August: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

7 September: C2, C3

3 Measures of similarity, summary statistics and probabilities

14 September: C4, C5

4 Probability densities and data visualization

21 September: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

28 September: C8, C9

6 Overfitting, cross-validation and Nearest Neighbor

5 October: C10, C12 (Project 1 due before 13:00)

7 Performance evaluation, Bayes, and Naive Bayes

12 October: C11, C13

8 Artificial Neural Networks and Bias/Variance

26 October: C14, C15

9 AUC and ensemble methods

2 November: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 November: C18

11 Mixture models and density estimation

16 November: C19, C20 (Project 2 due before 13:00)

12 Association mining

23 November: C21

Recap

13 Recap and discussion of the exam

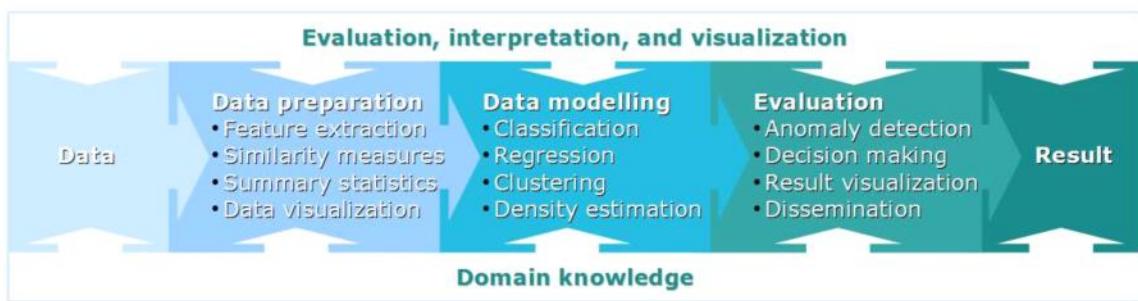
30 November: C1-C21

Online help: Forum on DTU Learn

Videos of lectures: <https://video.dtu.dk>

Streaming of lectures: Zoom (link on DTU Learn)

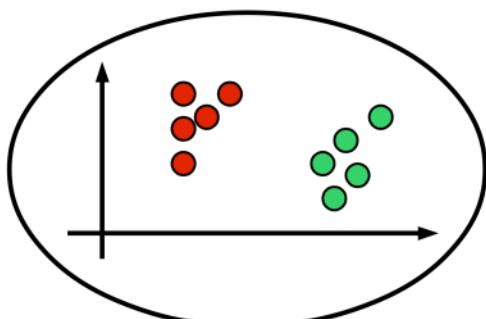
Lecture 5 28 September, 2021



Learning Objectives

- Explain what supervised learning is
- Explain the difference between classification and regression
- Be able to evaluate classifiers in terms of the confusion matrix, error rate and accuracy
- Understand the principle behind decision trees and Hunt's algorithm
- Apply and interpret decision trees, linear regression and logistic regression

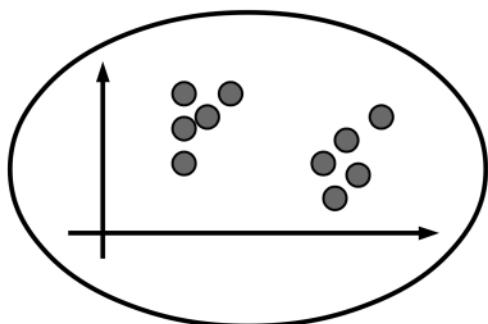
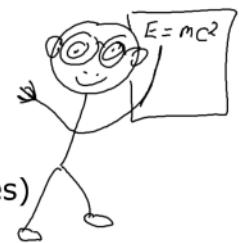
Supervised and Unsupervised learning



Supervised Learning

Input data x_n and output y_n

(Generalize from known examples)



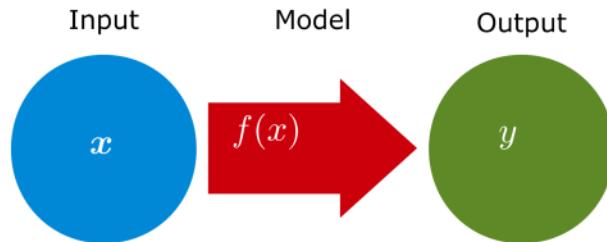
Unsupervised Learning

Input data x_n alone

(Exploratory analysis)



Supervised learning



- **Data**

- Inputs and outputs (*this is what we are given*)

$$\{\mathbf{x}_n, y_n\}_{n=1}^N$$

- **Model**

- Function that maps inputs to outputs (*what we are trying to determine*)

$$f(\mathbf{x})$$

- **Cost function**

- Dissimilarity measure between observation and prediction (*how we tell if a model is good or bad*)

$$d(y, f(\mathbf{x}))$$

How good I perform with the model:
y: what I predict,
f(x) : what I should predict

- **Types of supervised learning**

- Regression: Continuous output **y**
 - Classification: Discrete output **y**

Classification

- **Definition:** Learning a function that maps a data object to a discrete class
- **Why classify?**
 - Descriptive modeling
 - Explain / understand the relation between attributes and class
 - Predictive modeling
 - Predict the class of a new data object

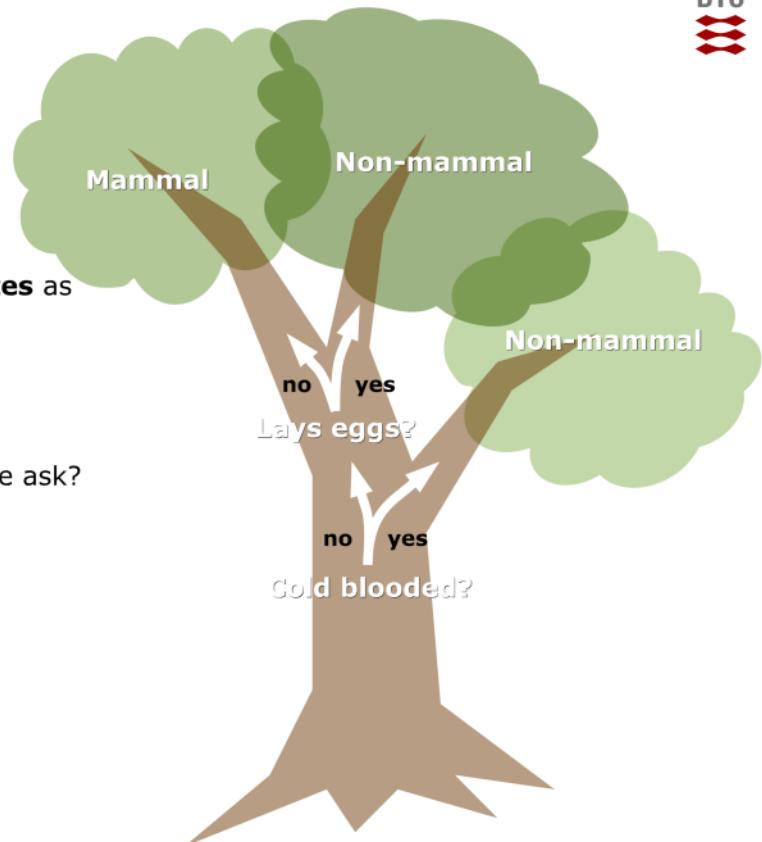
Decision trees

- Remember the game "20 questions to the professor"? (see also www.20q.net)

Q1. Is it an Animal? Yes.
Q2. Can you hold it? No.
Q3. Does it live in groups (gregarious)? Yes.
Q4. Are there many different sorts of it? No.
Q5. Can it jump? Yes.
Q6. Does it eat seeds? No.
Q7. Is it white? Sometimes.
Q8. Is it black and white? No.
Q9. Does it have paws? Yes.
Q10. Can you see it in a zoo? Yes.
Q11. Does it roar? Yes.
Q12. Is it worth a lot of money? Yes.
Q13. Does it have spots? Yes.
Q14. Is it multicoloured? Yes.
Q15. Can you make money by selling it? Yes.
Q16. Does it live in the jungle? Yes.
Q17. I guessed that it was a leopard? Wrong.
Q18. Does it like to play? Yes.
Q19. I guessed that it was a cheetah? Wrong.
Q20. I am guessing that it is a siberian tiger? Correct.

Decision trees

- Ask a series of questions until a conclusion is reached
- **Example:** Classify **vertebrates** as
 - **Mammal** or
 - **Non-mammal**
- **Learning task**
 - Which questions should we ask?



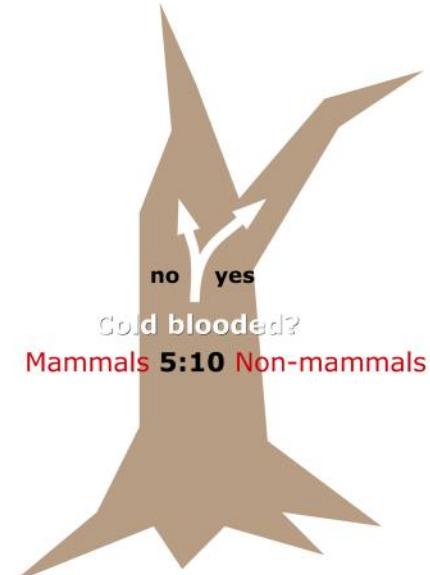
Hunts algorithm

- Assign all data objects to the root



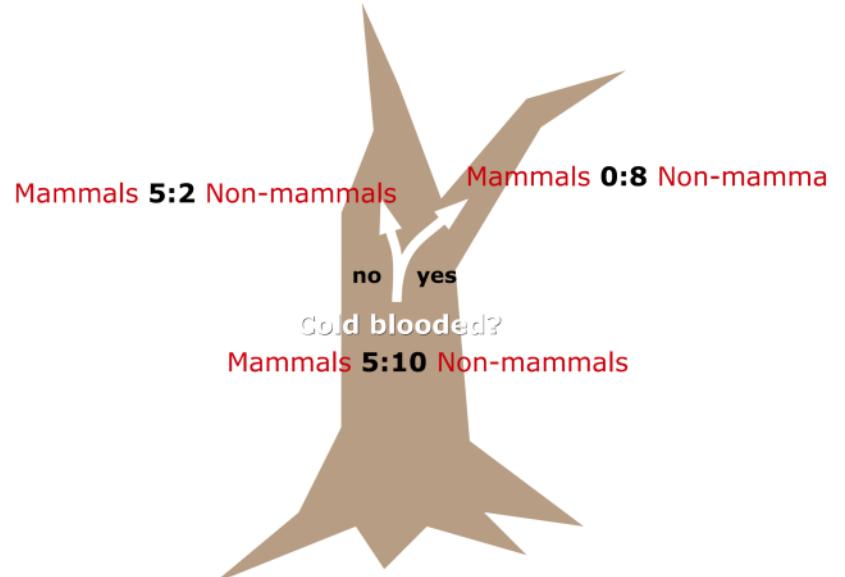
Hunts algorithm

- Select an attribute test condition
 - Find a good question to ask



Hunt's Algorithm

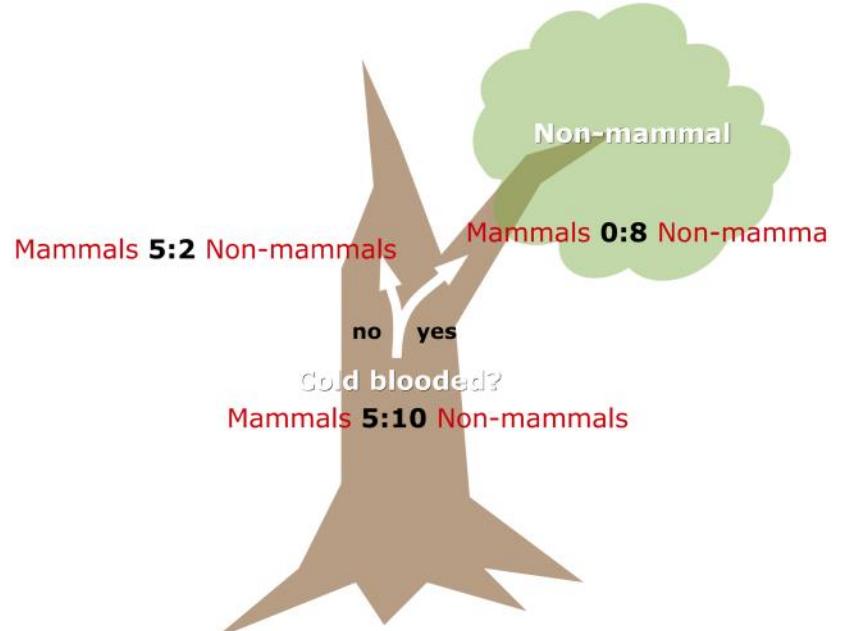
- Partition the data objects into subsets according to the test condition



Hunts algorithm

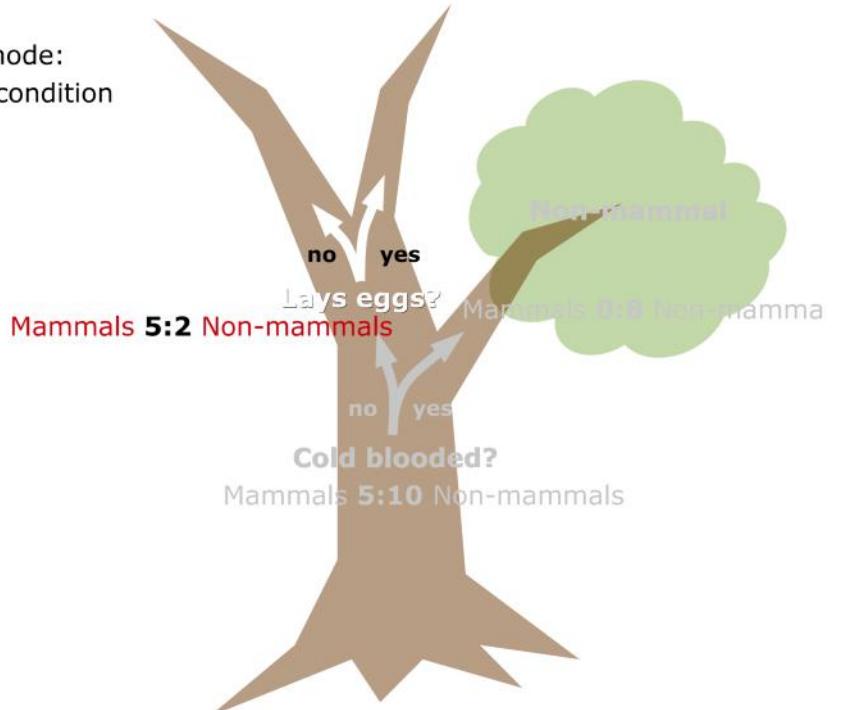
- If all data objects belong to the same class
 - Create a leaf node

This is a leaf node.
This leafnode is classified as PURE, since all the animal
In this group are of the same type (non-mammal).



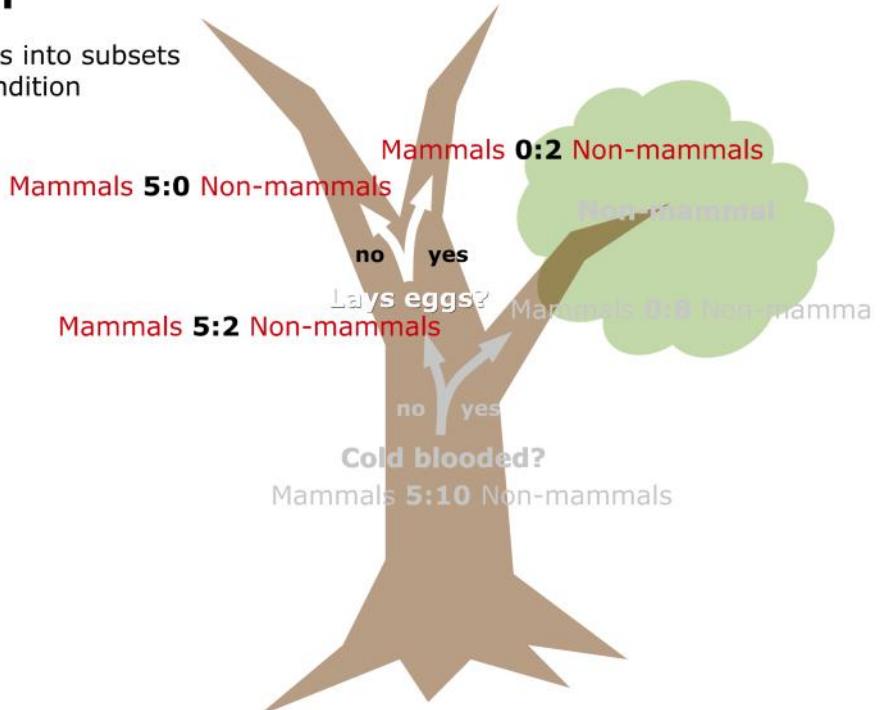
Hunts algorithm

- Repeat for each non-leave node:
 - Select an attribute test condition



Hunts algorithm

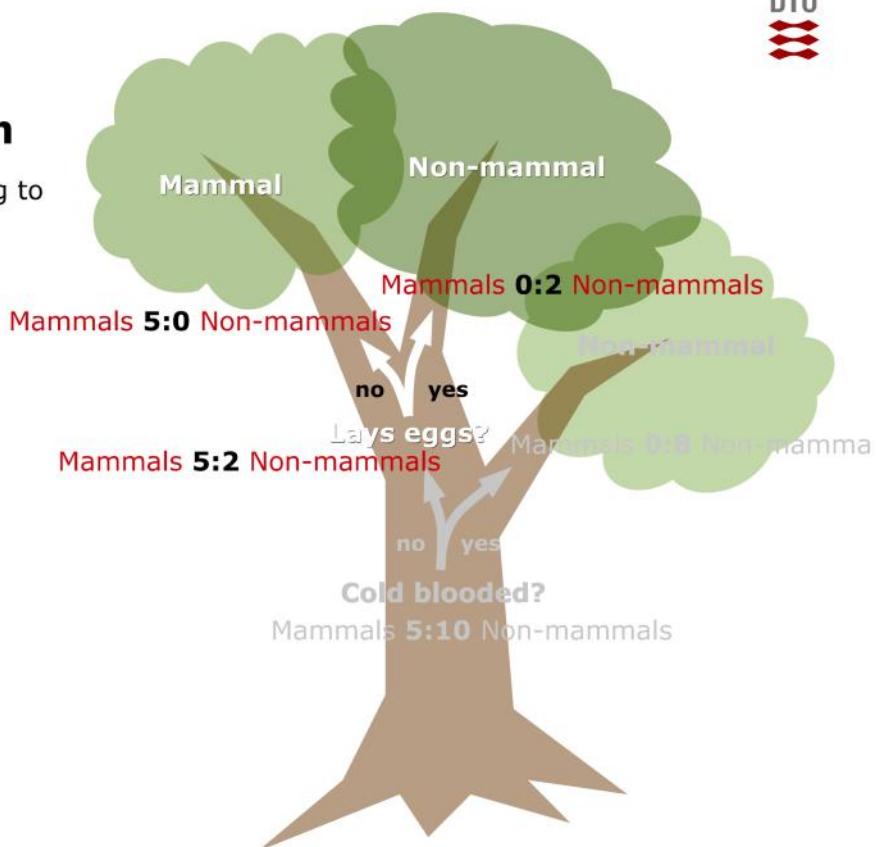
- Partition the data objects into subsets according to the test condition



Hunts algorithm

- If all data objects belong to the same class
 - Create a leaf node

We have 3 leaf nodes which are PURE: they completely make a distinction



Hunt's algorithm

- But how do we find the **best question** at each step?

Algorithm 2: Hunt's algorithm for decision trees

Require: Initial tree T only containing the root node

Require: D_r : Dataset associated with the current branch.
Initially just the full dataset

if The stop criterion is met **then**

 Add a leaf node to the tree which assigns every
 observation to the most prevalent class in D_r

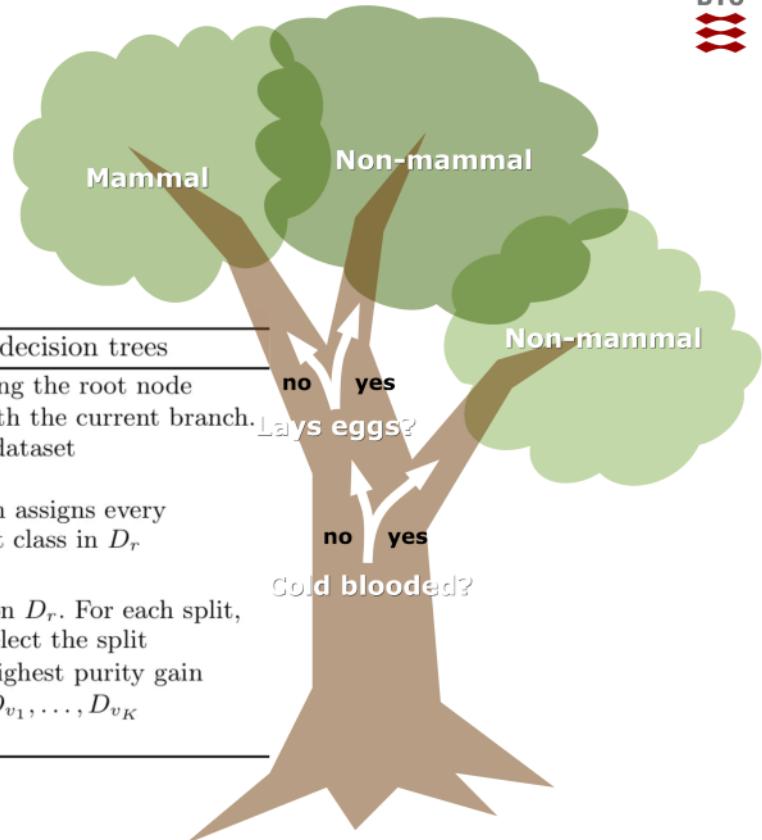
else

 Try a number of different splits on D_r . For each split,
 compute the **purity gain** and select the split

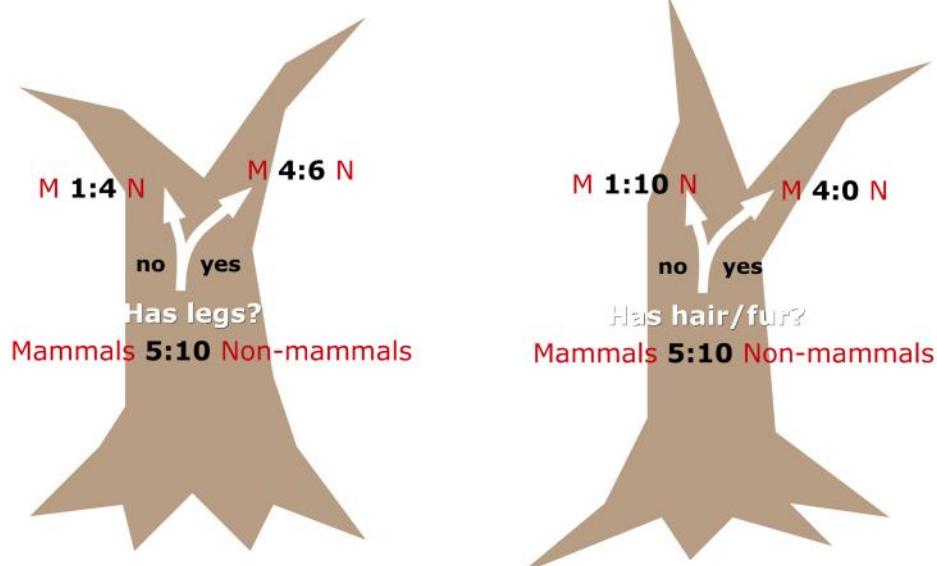
$D_r = \{D_{v_1}, \dots, D_{v_K}\}$ with the highest purity gain

 Recursively call the method on D_{v_1}, \dots, D_{v_K}

end if



Which split is best?

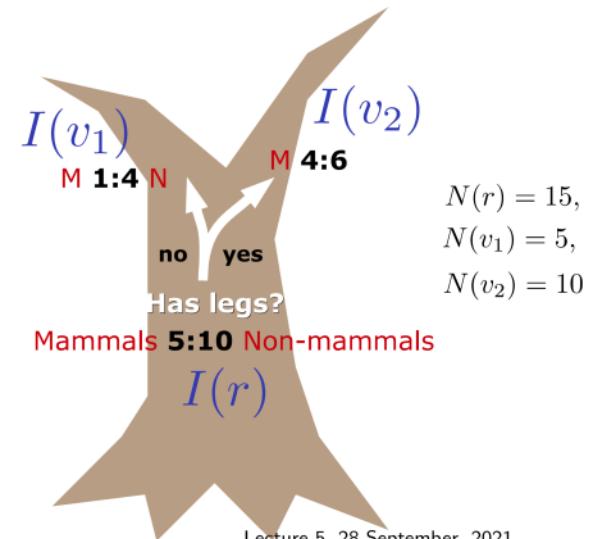


Which split is best?

- Create a measure Δ (**the purity gain**) of how good a split is
- A binary split creates 3 partitions: the root r and the right/left branches v_1, v_2 .
- For each partition, we compute $I(r), I(v_1), I(v_2)$ (the **impurity**)
- Purity gain is the **weighted reduction in impurity**:

$$\Delta = I(r) - \sum_{k=1}^{K=2} \frac{N(v_k)}{N(r)} I(v_k)$$

Delta calculate how much purity we gain passing from
The root to the leaves



Which split is best?

- Create a measure Δ (**the purity gain**) of how good a split is
- A binary split creates 3 partitions: the root r and the right/left branches v_1, v_2 .
- For each partition, we compute $I(r), I(v_1), I(v_2)$ (**the impurity**) of each partition
- Purity gain is the **weighted reduction in impurity**:

$$\Delta = I(r) - \sum_{k=1}^{K=2} \frac{N(v_k)}{N(r)} I(v_k)$$

The impurity measure $I()$ can be one of the following

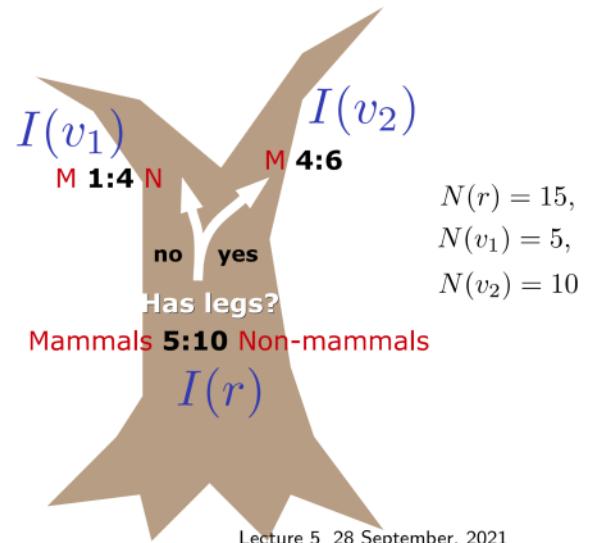
$$\text{Entropy}(v) = - \sum_{c=1}^C p(c|v) \log_2 p(c|v),$$

$$\text{Gini}(v) = 1 - \sum_{c=1}^C p(c|v)^2,$$

$$\text{ClassError}(v) = 1 - \max_c p(c|v).$$

$$p(c|v) = \frac{\{\text{Nr. in class } c \text{ in branch } v\}}{N(v)}$$

20 DTU Compute



Lecture 5 28 September, 2021

Quiz 1, Impurity gain

If we use the Gini index as impurity measure I , what is the purity gain Δ for the split indicated by the tree?

$$\Delta = I(r) - \sum_{k=1}^{K=2} \frac{N(v_k)}{N(r)} I(v_k)$$

$$\text{Gini}(v1) = 1 - (1/5)^2 - (4/5)^2$$

$$\text{Gini}(v2) = 1 - (4/10)^2 - (6/10)^2$$

$$\text{Gini}(r) = 1 - (5/15)^2 - (10/15)^2$$

The impurity measure $I()$ can be one of the following

$$\text{Entropy}(v) = - \sum_{c=1}^C p(c|v) \log_2 p(c|v),$$

$$\text{Gini}(v) = 1 - \sum_{c=1}^C p(c|v)^2,$$

$$\text{ClassError}(v) = 1 - \max_c p(c|v).$$

$$p(c|v) = \frac{\{\text{Nr. in class } c \text{ in branch } v\}}{N(v)}$$

21 DTU Compute

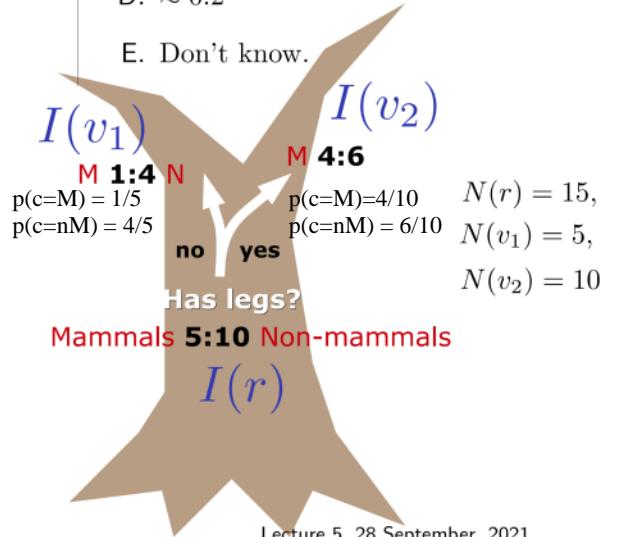
A. ≈ 0.0177

B. ≈ 0.104

C. ≈ 0.129

D. ≈ 0.2

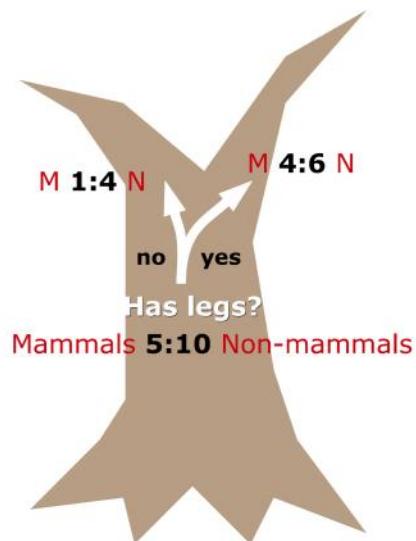
E. Don't know.



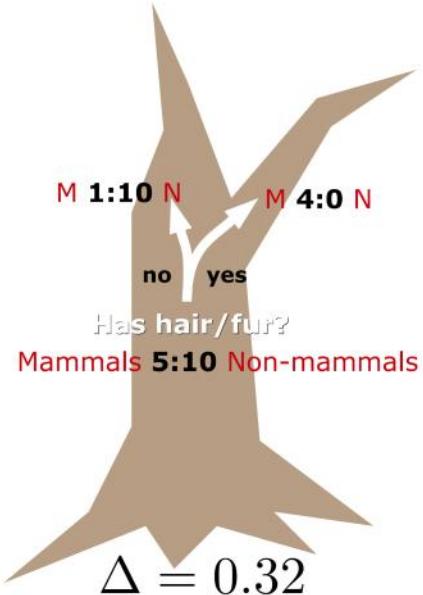
Lecture 5 28 September, 2021

Selecting the best split

- Consider a large number of possible splits
- Compute a measure of impurity after the proposed split
 - For each new branch of the tree
 - Compute weighted average impurity
- Choose split that reduces impurity most

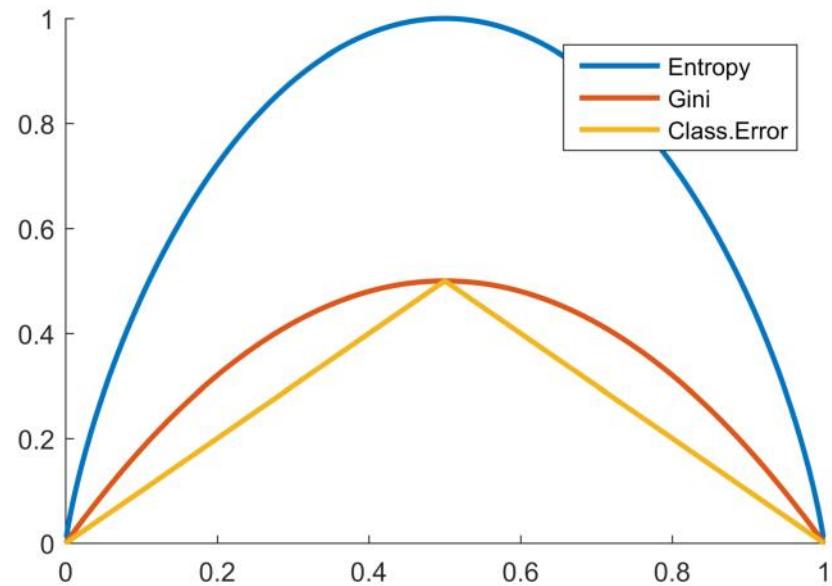


22 DTU Compute



Lecture 5 28 September, 2021

For a two class problem

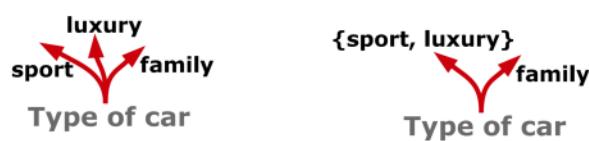


Which splits to consider

- Binary



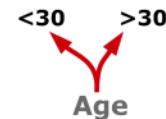
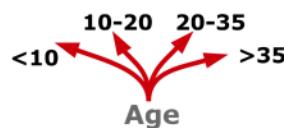
- Nominal



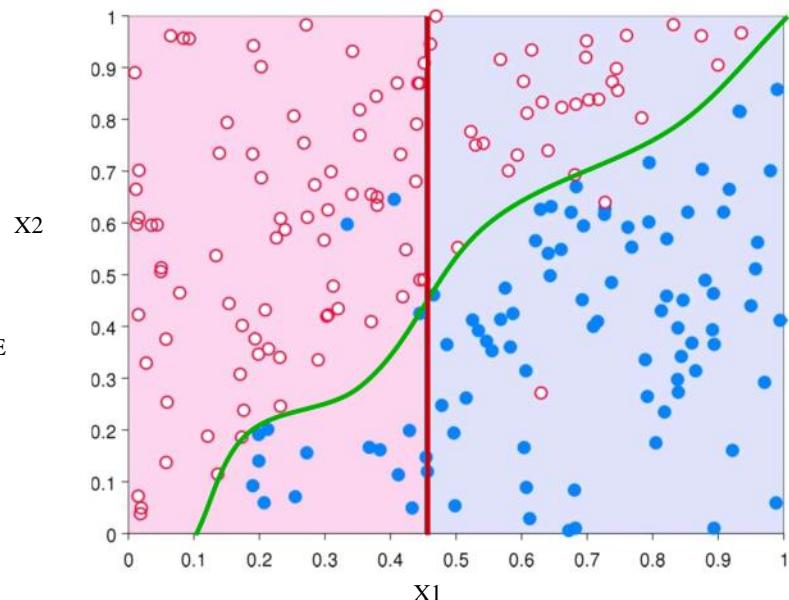
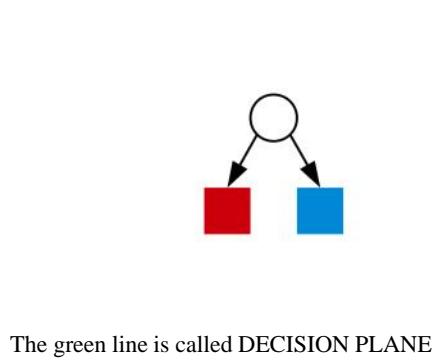
- Ordinal



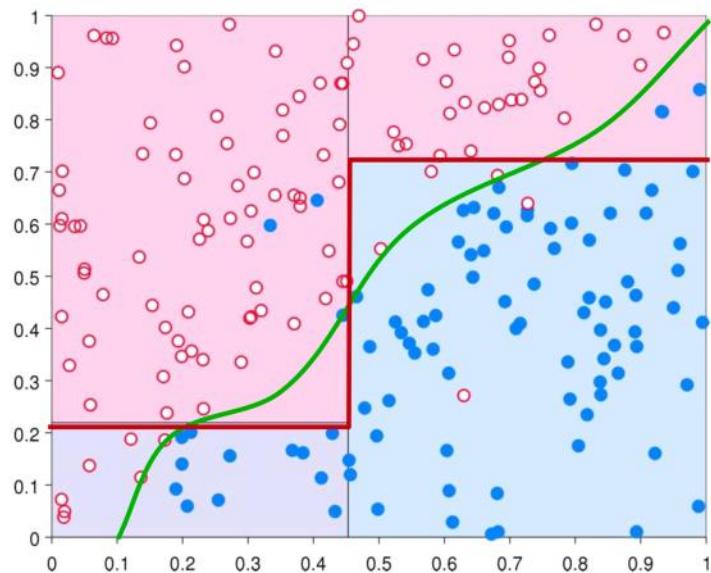
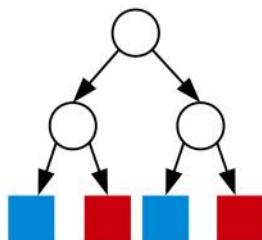
- Continuous



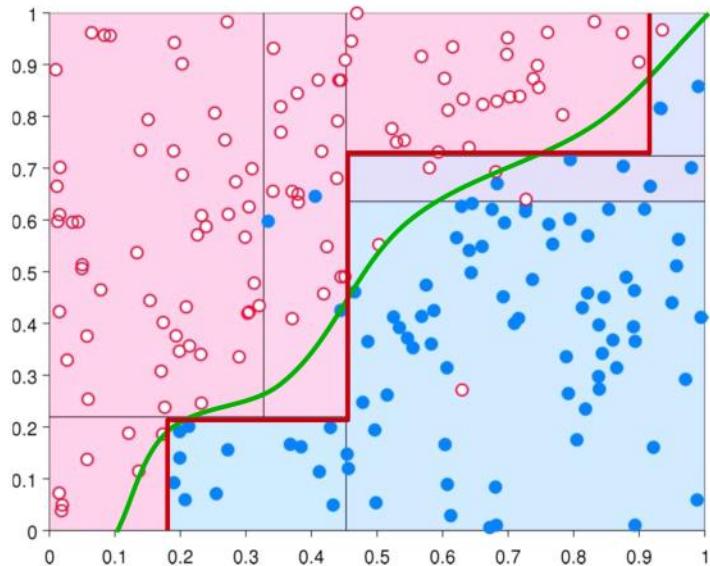
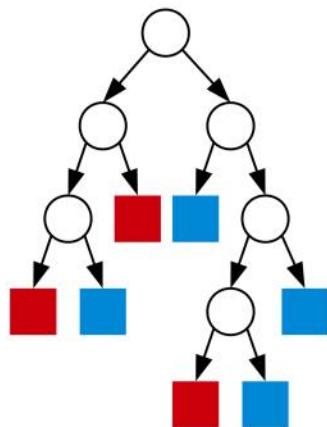
Classification Trees



Classification trees



Classification trees



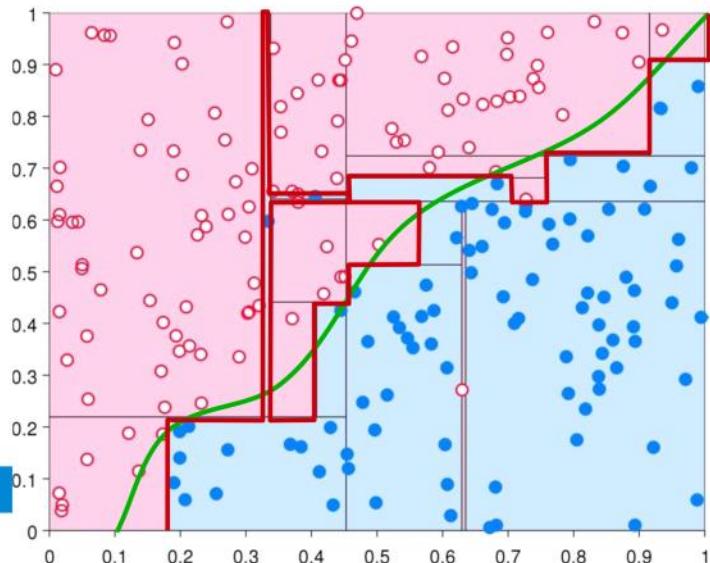
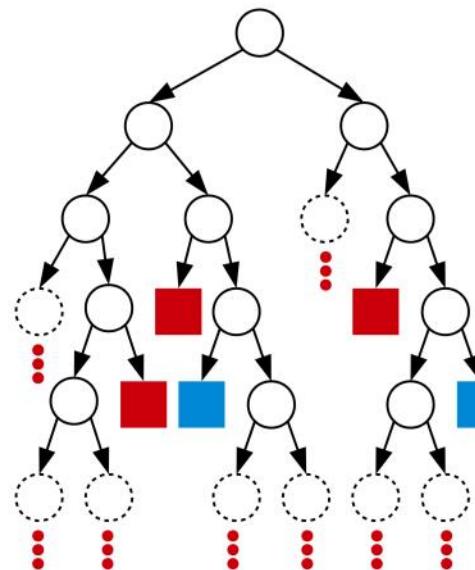
Classification trees



Common stopping criteria:

All records have the same class label

The number of observations have fallen below some minimum threshold



Regression trees

Algorithm 4: Hunt's algorithm for regression trees

```

Require: Initial tree  $T$  only containing the root node
Require:  $D_r$  : Dataset associated with the current branch. Initially just the full dataset
if The stop criterion is met then
    Add a leaf node to the tree which assigns every observation the mean value of the nodes in  $D_r$ :
     $y(r) = \frac{1}{N(r)} \sum_{i \in r} y_i$ 
else
    Try a number of different splits on  $D_r$ . For each split, compute the purity gain using the
    sum-of-squares impurity measure and select the split  $D_r = \{D_{v_1}, \dots, D_{v_K}\}$  with the highest purity
    gain
    Recursively call the method on  $D_{v_1}, \dots, D_{v_K}$ 
end if
```

Use mean square error as purity gain

$$I(v) = \frac{1}{N(v)} \sum_{i \in v} (y_i - \hat{y}_v)^2, \quad \hat{y}_v = \frac{1}{N(v)} \sum_{i \in v} y_i$$

It takes the output of the decision tree

Evaluating a classifier

Confusion matrix

- Visualization of actual versus predicted class labels

- Accuracy**

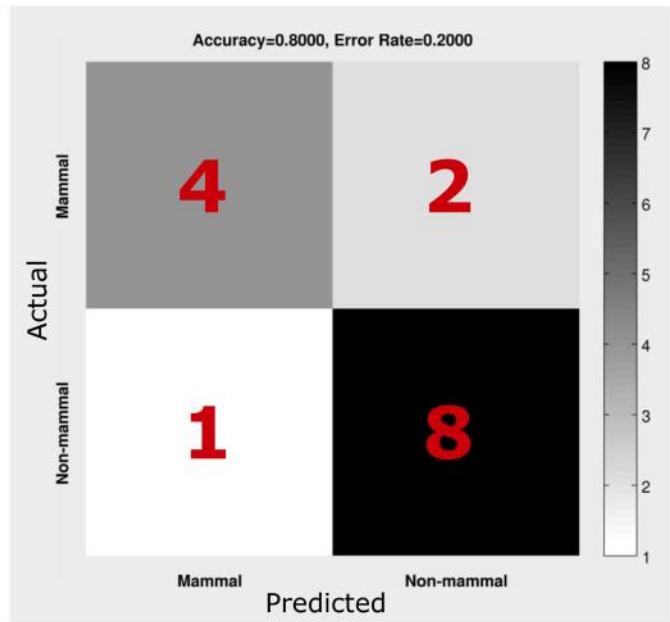
(Number of correctly predicted observations divided by the total number of observations)

$$\frac{4 + 8}{4 + 2 + 1 + 8} = 80\%$$

- Error rate**

(Number of in-correctly predicted observations divided by the total number of observations)

$$\frac{2 + 1}{4 + 2 + 1 + 8} = 20\%$$



Evaluating a regression model

Compute average loss per observation:

$$E = \frac{1}{N} \sum_{i=1}^N L(y_i, f(\mathbf{x}_i))$$

Where we either use L_1 or L_2 (Euclidean) loss

$$L_1(y_i, \hat{y}_i) = |y_i - \hat{y}_i|, \quad L_2(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

(Compare these to p -norms)

Example: Iris data

The iris data set

- **Three 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

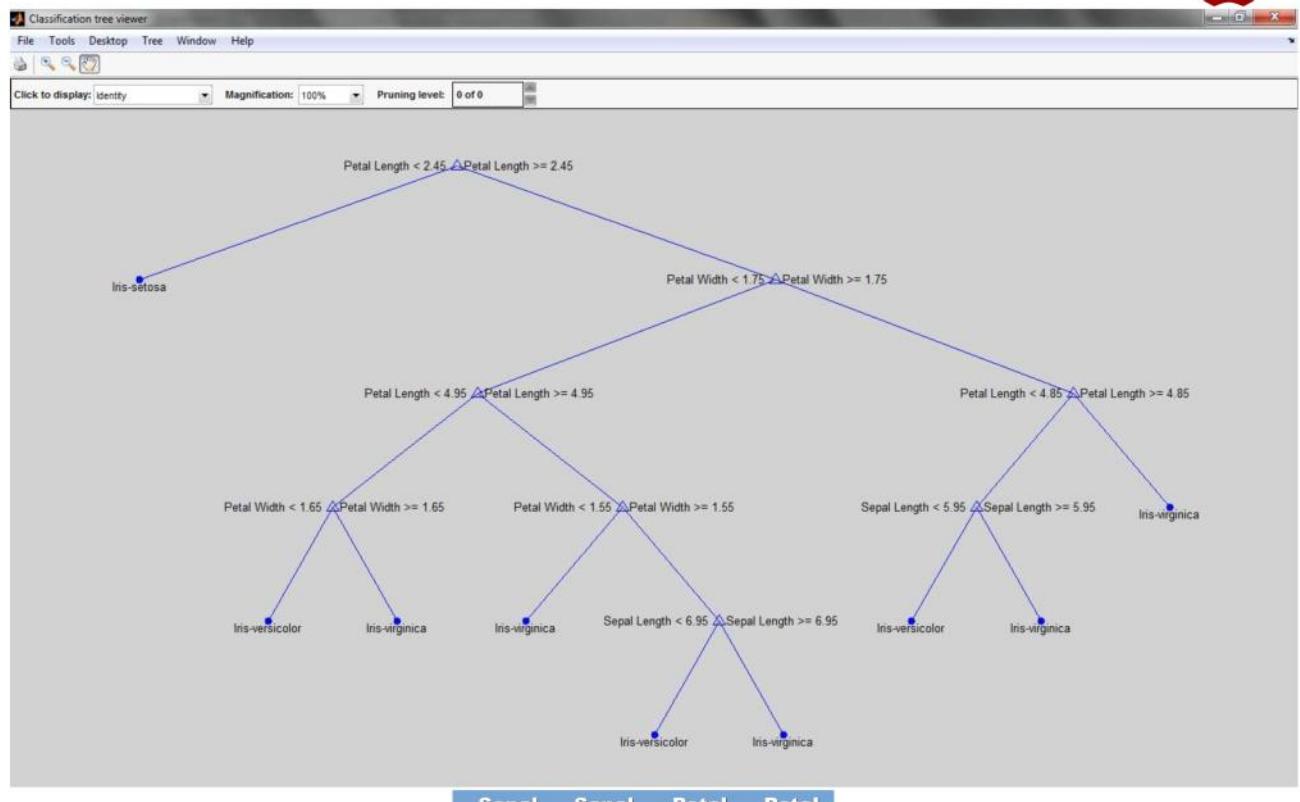
- Class of flower

- Iris Setosa
- Iris Versicolour
- Iris Virginica



Flower 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

$X^{\text{Observation} \times \text{Attribute}}$

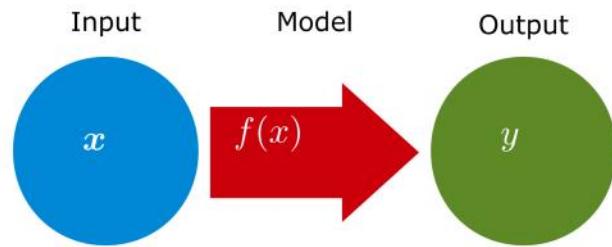


What would the following iris flower be classified as?

Sepal Length	Sepal Width	Petal Length	Petal Width
4.0	3.5	3.0	2.0

Lecture 5 28 September, 2021

Supervised learning



- **Mapping between domains**

- Classification: Discrete (nominal) output
- Regression: Continuous output

Supervised learning

- **Data**

- Inputs and outputs $\{\mathbf{x}_n, y_n\}_{n=1}^N$

- **Model**

- Function that maps inputs to outputs
 $f(\mathbf{x})$

- **Cost function**

- Dissimilarity measure between data and model

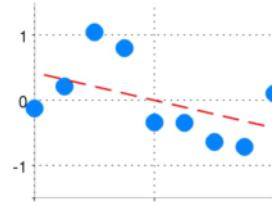
$$d(y, f(\mathbf{x}))$$

Regression

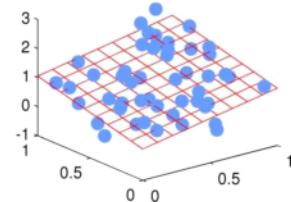
- **Definition:** Learning a function that maps a data object to a continuous-valued output
- **Why Regression?**
 - Descriptive modeling
 - Explain / understand the relation between attributes and continuous-valued output
 - Predictive modeling
 - Predict the output value of a new data object

Linear regression

- 1-dimensional inputs
 $f(x) = w_0 + w_1 x$



- 2-dimensional inputs
 $f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2$



- K-dimensional inputs
 $f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$

Linear regression

- K-dimensional inputs

$$f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_Kx_K$$

- Non-linearly transformed inputs

$$f(x) = w_0 + w_1x + w_2x^2 + \cdots + w_Kx^K$$

$$f(x) = w_0 + w_1\sin(x) + w_2\cos(x)$$

The linear regression is linear in the weights (w), not necessary in the x (we can do all kind of trasformation to the input data x, like square or cos and sin).

Linear regression

- K-dimensional inputs

$$f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_Kx_K$$

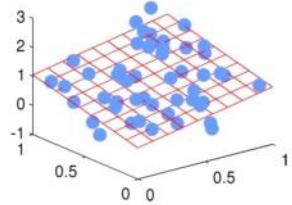
- Non-linearly transformed inputs

$$f(x) = w_0 + w_1x + w_2x^2 + \cdots + w_Kx^K$$

$$f(x) = w_0 + w_1\sin(x) + w_2\cos(x)$$

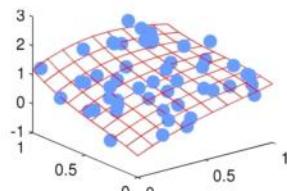
• Example

$$f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2$$



$$\begin{aligned} f(\mathbf{x}) = & w_0 + w_1x_1 + w_2x_2 \\ & + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2 \\ & + w_6x_1^3 + w_7x_1^2x_2 + w_8x_1x_2^2 + w_9x_2^3 \end{aligned}$$

Non linear regression:
Adding terms of higher grade



Vector notation

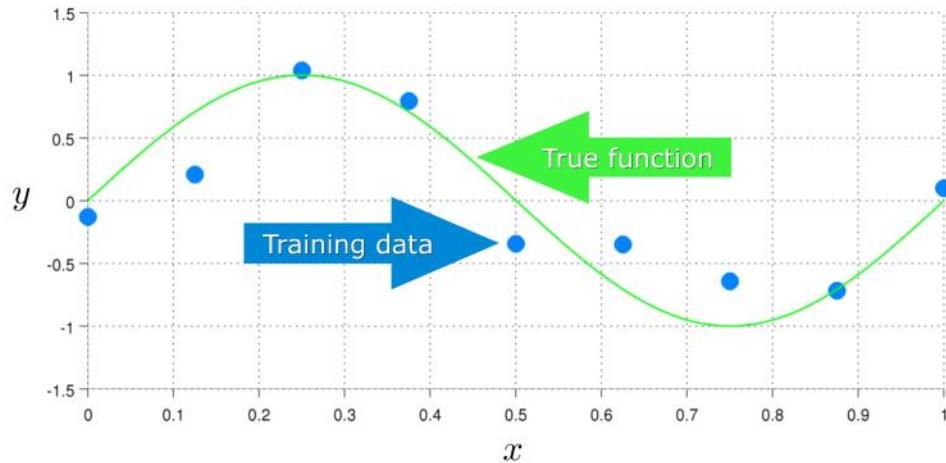
- The linear model can be written compactly using vector notation

$$\begin{aligned}f(\mathbf{x}) &= w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K \\&= \sum_{k=0}^K w_k x_k = \boxed{\mathbf{x}^\top \mathbf{w}} \quad = \mathbf{w} \text{ dot } \mathbf{x} \quad (\text{dot product})\end{aligned}$$

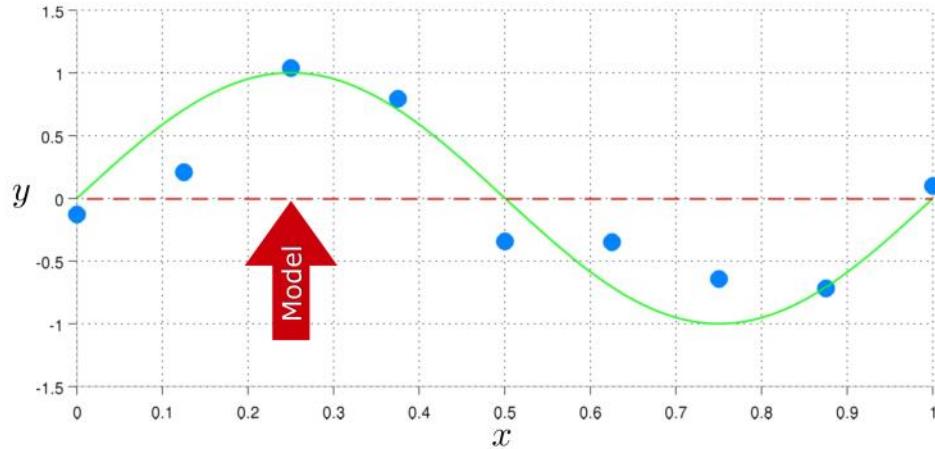
- where $x_0 = 1$

Linear regression

We want to learn the TRUE function



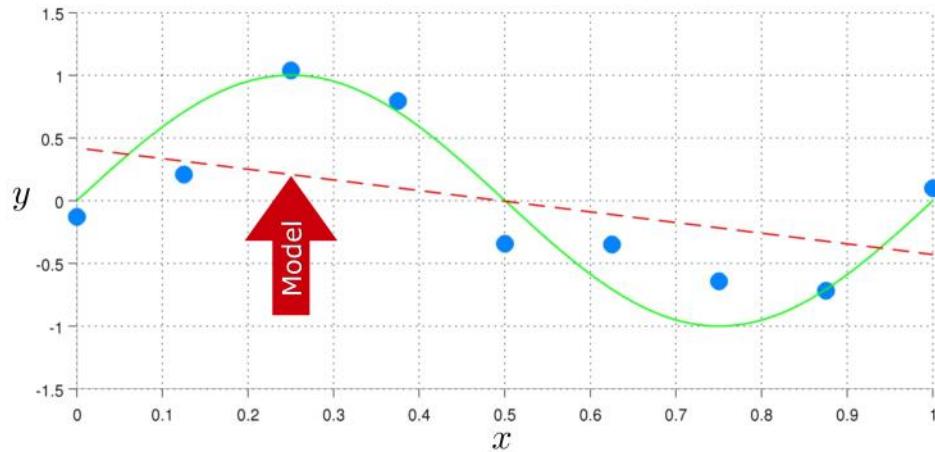
Linear regression



Model

$$f(x) = w_0$$

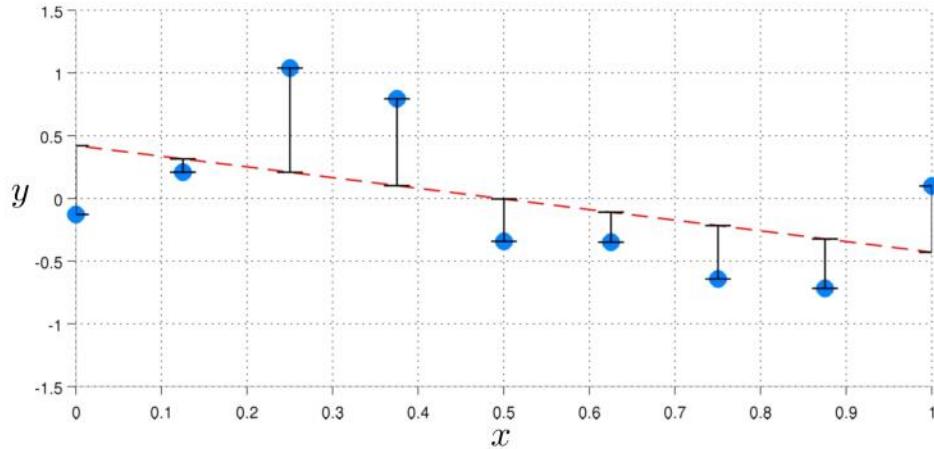
Linear regression



Model

$$f(x) = w_0 + w_1 x$$

Residual error

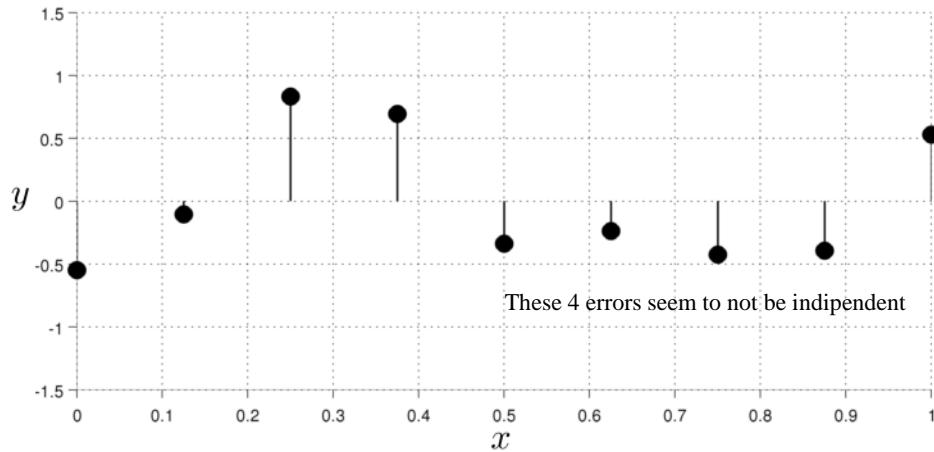


Model

$$f(x) = w_0 + w_1x$$

Residual error

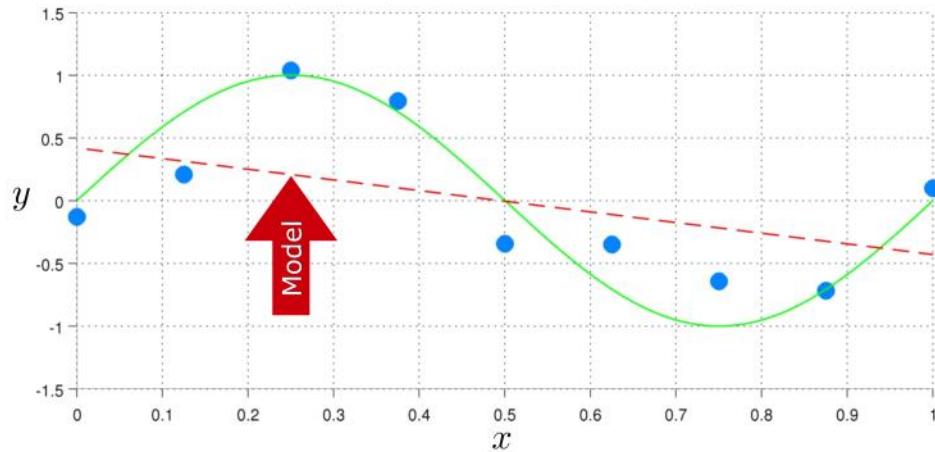
The residual is defined as the prediction – the training data



Model

$$f(x) = w_0 + w_1 x$$

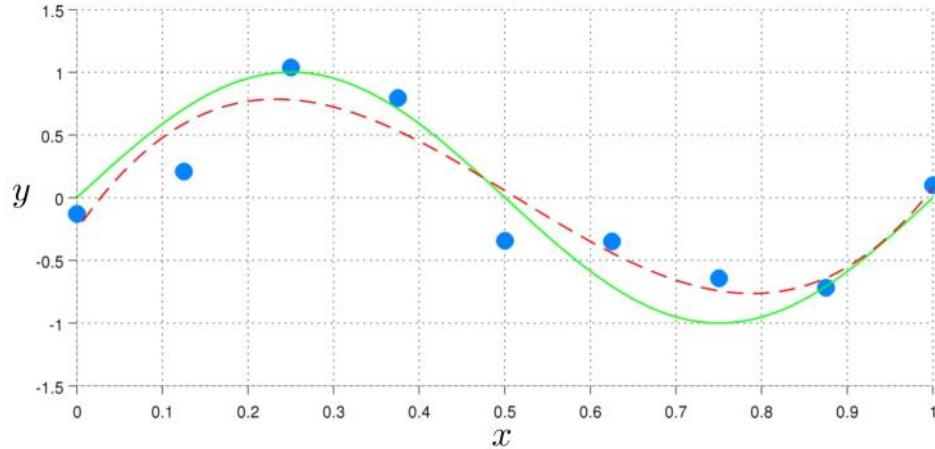
Linear regression



Model

$$f(x) = w_0 + w_1x$$

Linear regression



Model

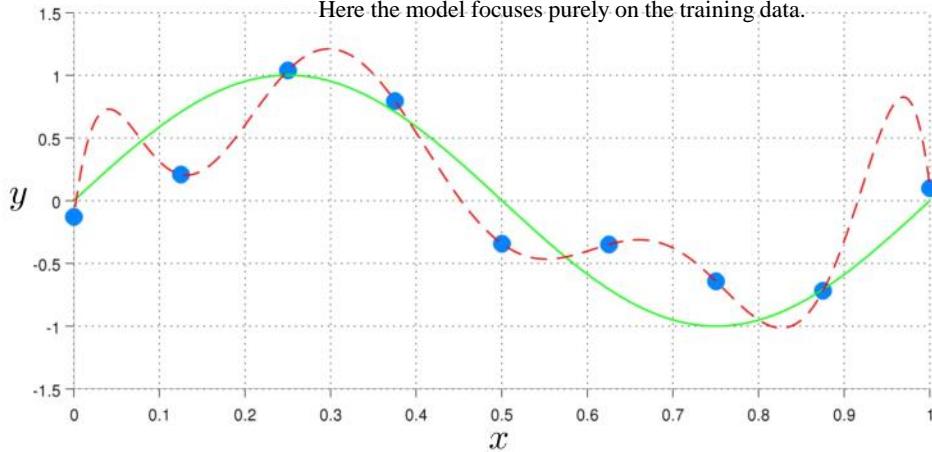
$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

Linear regression

Here we have overfitting: because we have 9 points and 9 degrees of freedom -> It means that the solution is unique and the residual error is 0.

If we have overfitting, then if we make a prediction, our model is really good when the new observation is close to one of the 9 point, but it's bad otherwise.

Here the model focuses purely on the training data.



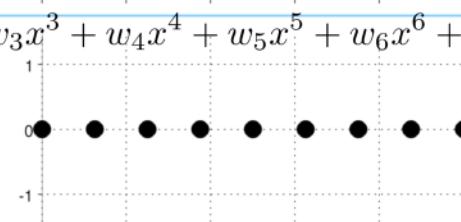
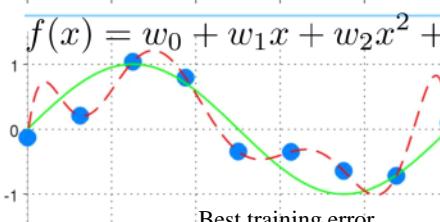
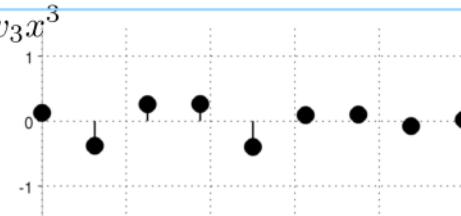
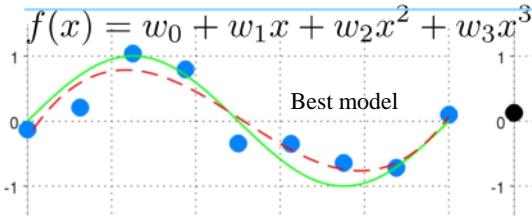
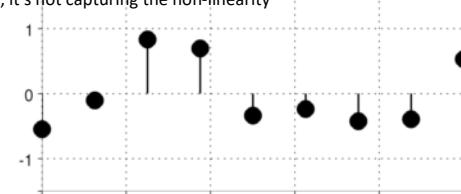
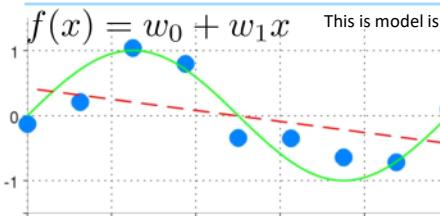
Model

$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5 + w_6 x^6 + w_7 x^7 + w_8 x^8$$



Model order

- Which model order
 - Gives the best fit?
 - Do you think is most "correct"?



Learning

How do we deduct the parameters w??

Setup: We got some data (\mathbf{y}, \mathbf{X}) . We have a way to define a function

$$f(\mathbf{x}; \mathbf{w}) = \tilde{\mathbf{x}}^T \mathbf{w} = w_0 + \sum_{k=1}^M x_k w_k$$

Learning

Setup: We got some data (\mathbf{y}, \mathbf{X}) . We have a way to define a function

$$f(\mathbf{x}; \mathbf{w}) = \tilde{\mathbf{x}}^T \mathbf{w} = w_0 + \sum_{k=1}^M x_k w_k$$

- **Question:** How do we learn the correct \mathbf{w} ?

Learning

Setup: We got some data (\mathbf{y}, \mathbf{X}) . We have a way to define a function

$$f(\mathbf{x}; \mathbf{w}) = \tilde{\mathbf{x}}^T \mathbf{w} = w_0 + \sum_{k=1}^M x_k w_k$$

- **Question:** How do we learn the correct \mathbf{w} ?

- Answer: For each observation, assume:

$$y_i = f(\mathbf{x}_i; \mathbf{w}) + \varepsilon_i$$

where ε_i is a normally distributed noise term $\mathcal{N}(\varepsilon_i | \mu = 0, \sigma^2)$. Recall:
The noise normal distributed is a choice

$$\mathcal{N}(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Learning

Setup: We got some data (\mathbf{y}, \mathbf{X}) . We have a way to define a function

$$f(\mathbf{x}; \mathbf{w}) = \tilde{\mathbf{x}}^T \mathbf{w} = w_0 + \sum_{k=1}^M x_k w_k$$

- **Question:** How do we learn the correct \mathbf{w} ?

- Answer: For each observation, assume:

$$y_i = f(\mathbf{x}_i; \mathbf{w}) + \varepsilon_i \rightarrow \mathbf{f}_i = \mathbf{Y}_i - f(\mathbf{X}_i; \mathbf{w})$$

where ε_i is a normally distributed noise term $\mathcal{N}(\varepsilon_i | \mu = 0, \sigma^2)$. Recall:

$$\mathcal{N}(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- This means that

$$p(y_i | \mathbf{x}_i, \mathbf{w}) = N(Y_i | f(\mathbf{x}_i; \mathbf{w}), \sigma^2)$$

This is also going to be normally distributed

Recall from last time: Maximum A Posteriori (MAP) learning

- Consider some data $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_N$ and $\mathbf{y} = y_1, \dots, y_N$

- Suppose \mathbf{x}_i relates to y_i by some parameters \mathbf{w}

- Assume**

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \mathbf{w}), \quad p(\mathbf{w}|\mathbf{X}) = p(\mathbf{w})$$

- Then

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})} = \frac{\prod_{i=1}^N p(y_i|\mathbf{w}, \mathbf{x}_i)p(\mathbf{w})}{p(\mathbf{X}|\mathbf{y})}$$

Normal distribution

- And maximizing: $\mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{y})$ is equivalent to

$$\text{Minimize: } \mathbf{w}^* = \arg \min_{\mathbf{w}} E(\mathbf{w})$$

$$E(\mathbf{w}) = \frac{1}{N} \left[- \sum_{i=1}^N \log p(y_i|\mathbf{x}_i, \mathbf{w}) - \log p(\mathbf{w}) \right]$$

Back to the linear model

$$p(y_i | \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \tilde{\mathbf{x}}_i^\top \mathbf{w})^2}{2\sigma^2}}$$

Optimal $\mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{w} | \mathbf{X}, \mathbf{y})$ found as $\mathbf{w}^* = \arg \min_{\mathbf{w}} E(\mathbf{w})$

$$E(\mathbf{w}) = \frac{1}{N} \left[- \sum_{i=1}^N \log p(y_i | \mathbf{x}_i, \mathbf{w}) - \log p(\mathbf{w}) \right]$$

By assuming a constant (flat prior) we can ignore we obtain (Máximo Likelihood learning)

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{N} \sum_{i=1}^N \frac{(y_i - \tilde{\mathbf{x}}_i^\top \mathbf{w})^2}{2\sigma^2} = \frac{1}{N} \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \tilde{\mathbf{x}}_i^\top \mathbf{w})^2 \propto \|\mathbf{y} - \tilde{\mathbf{X}}\mathbf{w}\|^2 \\ \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} &= 2\tilde{\mathbf{X}}^\top (\mathbf{y} - \tilde{\mathbf{X}}\mathbf{w}) = 0 \\ \Rightarrow \mathbf{w}^* &= (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y} \end{aligned}$$

} Kind of final cost
It's basically a mean square error

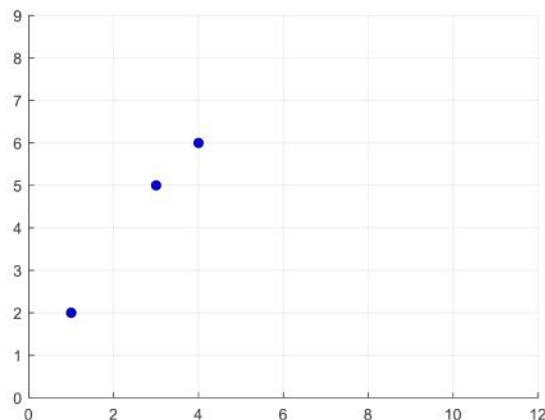
Quiz 2, The linear model

Suppose you observe three points:

$$(x, y) = (1, 2), (3, 5), (4, 6)$$

Knowing what you have learned so far, you first bring these points to the standard format:

$$\mathbf{X} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 2 \\ 5 \\ 6 \end{bmatrix}$$



You wish to train a linear model of the form $y = ax + b$ on this dataset. What is $\mathbf{w} = \begin{bmatrix} b \\ a \end{bmatrix}$? Then, compute the prediction of the model at $x = 5$? (the prediction is given as: $y = \tilde{\mathbf{x}}^\top \mathbf{w}^*$)

- A. 6.5 How to get the X_tilda?
- B. 7
- C. 7.5
- D. 8
- E. Don't know.

$$\begin{aligned} \mathbf{X}_{\text{tilda}} &= [1 \ 1 \ 1 \ 3 \ 1 \ 4] & \mathbf{Y} &= \mathbf{X}_{\text{tilda}} * \mathbf{w} \\ &&& \\ \mathbf{W} &= [0.71 \ 1.36] & \mathbf{X}_4 &= [1 \ 5] \end{aligned}$$

$$\text{Recall } \mathbf{w}^* = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y}$$

Logistic regression

- Assume we are given (\mathbf{X}, \mathbf{y}) , but assume y is *binary*: $y_i = 0, 1$
- An idea is to use the Bernoulli distribution

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \text{Bernoulli}(y_i|\theta_i) = \theta_i^{y_i} (1 - \theta_i)^{1-y_i}$$

Where θ_i depends on \mathbf{w} and \mathbf{x}_i .

- **Problem:** θ_i must belong to the unit interval, but $f(\mathbf{x}_i, \mathbf{w}) = \tilde{\mathbf{x}}_i^\top \mathbf{w}$ won't
- **Solution:** Assume

$$\theta_i = \sigma(f(\mathbf{x}, \mathbf{w})), \quad \text{where } \sigma(z) = \frac{1}{1 + e^{-z}} \text{ is the logistic sigmoid}$$

$0 < .. < 1$

Then

$$-\log p(y_i|\mathbf{x}_i, \mathbf{w}) =$$

Recall from 10 minutes ago: Maximum A Posteriori (MAP) learning

- Consider some data $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_N$ and $\mathbf{y} = y_1, \dots, y_N$

- **Assume**

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \mathbf{w}), \quad p(\mathbf{w}|\mathbf{X}) = p(\mathbf{w})$$

- Then

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{\prod_{i=1}^N p(\mathbf{y}_i|\mathbf{w}, \mathbf{x}_i)p(\mathbf{w})}{p(\mathbf{X}|\mathbf{y})}$$

- Maximizing: $\mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{y})$ is equivalent to $\mathbf{w}^* = \arg \min_{\mathbf{w}} E(\mathbf{w})$

$$E(\mathbf{w}) = \frac{1}{N} \left[- \sum_{i=1}^N \log p(y_i|\mathbf{x}_i, \mathbf{w}) - \log p(\mathbf{w}) \right]$$

- By assuming a constant (flat prior) we can ignore we obtain (Máximo Likelihood learning)

$$E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N [-y_i \log(\theta_i) - (1 - y_i) \log(1 - \theta_i)], \quad \theta_i = \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}) = \frac{1}{1 + e^{-\tilde{\mathbf{x}}_i^\top \mathbf{w}}}$$

Quiz 3, Logistic regression

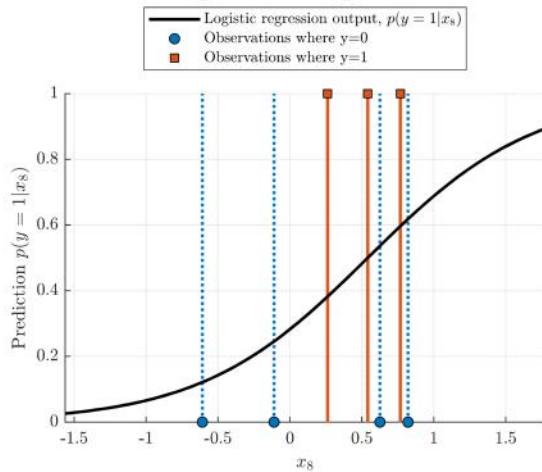


Figure 1: Output of a logistic regression classifier trained on 7 observations from the dataset.

Consider the Avila Bible dataset. We are particularly interested in predicting whether a bible copy was written by copyist 1, and we therefore wish to train a logistic regression classifier to distinguish between copyist one vs. copyist two and three.

To simplify the setup further, we select just 7

observations and train a logistic regression classifier using only the feature x_8 as input (as usual, we apply a simple feature transformation to the inputs to add a constant feature in the first coordinate to handle the intercept term). To be consistent with the lecture notes, we label the output as $y = 0$ (corresponding to copyist one) and $y = 1$ (corresponding to copyist two and three).

In Figure 1 is shown the predicted output probability an observation belongs to the positive class, $p(y = 1|x_8)$. What are the weights?

A. $\begin{bmatrix} -0.93 \\ 1.72 \end{bmatrix}$

B. $\begin{bmatrix} -2.82 \\ 0.0 \end{bmatrix}$

C. $\begin{bmatrix} 1.36 \\ 0.4 \end{bmatrix}$

D. $\begin{bmatrix} -0.65 \\ 0.0 \end{bmatrix}$

E. Don't know.

General linear model

$$\text{Linear regr.: } E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \|y_i - \tilde{\mathbf{x}}_i^\top \mathbf{w}\|^2$$

$$\text{Logistic regr.: } E(\mathbf{w}) = \frac{-1}{N} \sum_{i=1}^N [y_i \log(\theta_i) + (1-y_i) \log(1-\theta_i)], \quad \theta_i = \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w})$$

$$\text{GLM } E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N d(y_i, g(\tilde{\mathbf{x}}_i^\top \mathbf{w}))$$

General linear model

Linear reg.: $E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \|y_i - \tilde{\mathbf{x}}_i^\top \mathbf{w}\|^2$

Logistic reg.: $E(\mathbf{w}) = \frac{-1}{N} \sum_{i=1}^N [y_i \log(\theta_i) + (1-y_i) \log(1-\theta_i)], \quad \theta_i = \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w})$

GLM $E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N d(y_i, g(\tilde{\mathbf{x}}_i^\top \mathbf{w}))$ It's the generalization: the previous 2 are a specific case of this One, where a well defined cost function is used

We call d the cost function and g the link function. In our examples:

Lin.reg. : $d(y, z) = \|y - z\|^2, \quad z = g(\tilde{\mathbf{x}}_i^\top \mathbf{w}) = \tilde{\mathbf{x}}_i^\top \mathbf{w}$

Log.reg. : $d(y, z) = -y \log z - (1-y) \log(1-z), \quad z = g(\tilde{\mathbf{x}}_i^\top \mathbf{w}) = \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w})$

Resources

<http://www2.imm.dtu.dk> Our interactive regression demo

(<http://www2.imm.dtu.dk/courses/02450/DemoComplexityRegression.html>)

Lecture 6

5. oktober 2021 13:02



02450: Introduction to Machine Learning and Data Mining

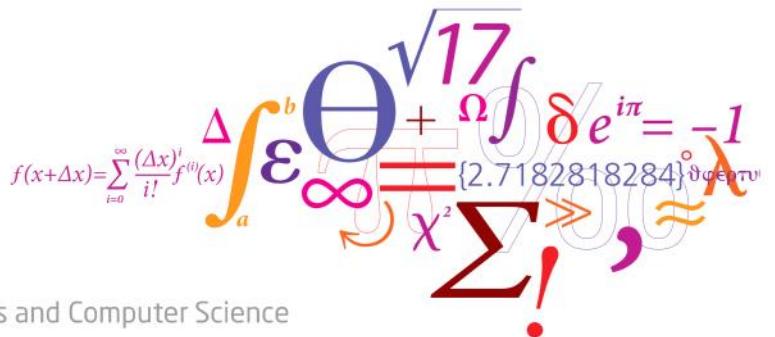
Overfitting, cross-validation and Nearest Neighbor

Tommy Sonne Alstrøm

DTU Compute, Technical University of Denmark (DTU)

DTU Compute

Department of Applied Mathematics and Computer Science

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$


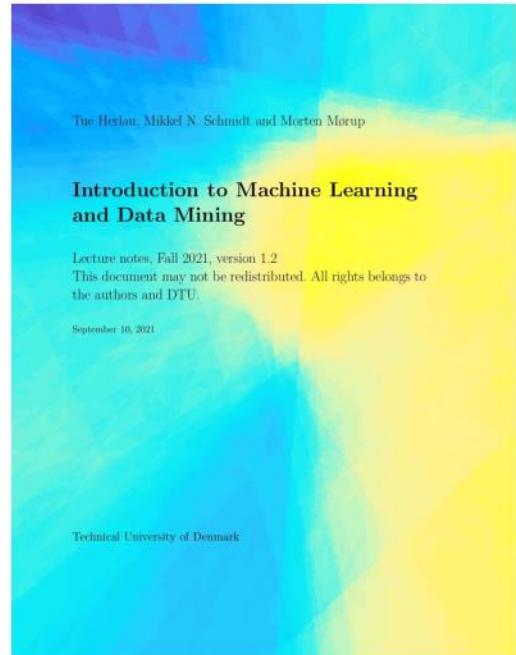
Today

Feedback Groups of the day:

Anders Dan Hansen, Anders Helbo, Andreas Hürtig, Andreas Jacobsen, Andreas Krogsgaard Holme, August Høgsted, Benjamin Fester Henningsen, Betina Mee Hansen, Bjarni Pétur Hinriksson, Carina Nørgaard Kock Holst, Christian Højdevang, Christopher Sonne Hansen, Chunjun Huang, Colin Hoffmann, Cornelius Valentin Hasselris, David Immanuel Hartel, Elias Høstmælingen, Eline Hareide, Fabian Depenau Bjørnholt Jacobsen, Frederik Bingen Jacobsen, Freja Andersen Jelling Holm, Halla María Hjartardóttir, Hans Christian Munter Hansen, Hussein Abukar Hussein, Ida Kofoed Hvidsten, Johannes Ikas, Jonas Christian Henriksen, Jonas Hoffmann, Jonas Løvenhardt Henriksen, Juliane Hermary, Kasper Brun Hansen, Kristín Sóley Ingvars dóttir, Lauge Thode Hermansen, Liv Isa Christina Hyllinge, Magnus Waldemar Hoff Harder, Maja Jønck Hjuler, Maria Hendrikx, Marie Hoel, Martin Hanna Hoffmann, Massimo Hansen, Michael Alexander Harborg, Natasha Hougaard, Nicholas Alan Hill, Niels Raunkjær Holm, Oskar William Rosenstand Hibbert, Philip Johannes Fog Helsted, Roza Ibrahim Hasso, Shania Hau, Stefan Ignat, Tobias Høyrup Hemmingsen, Torine Reed Herstad, Verena Vanessa Irmng-Pedersen, Victor Alexander Hansen, Victoria Charlotte Ipsen, Weijun Huang, Yi Huang, Yun Huang

2 DTU Compute

Reading material: Chapter 10, Chapter 12



Lecture Schedule

1 Introduction

31 August: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

7 September: C2, C3

3 Measures of similarity, summary statistics and probabilities

14 September: C4, C5

4 Probability densities and data visualization

21 September: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

28 September: C8, C9

6 Overfitting, cross-validation and Nearest Neighbor

5 October: C10, C12 (Project 1 due before 13:00)

7 Performance evaluation, Bayes, and Naive Bayes

12 October: C11, C13

8 Artificial Neural Networks and Bias/Variance

26 October: C14, C15

9 AUC and ensemble methods

2 November: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 November: C18

11 Mixture models and density estimation

16 November: C19, C20 (Project 2 due before 13:00)

12 Association mining

23 November: C21

Recap

13 Recap and discussion of the exam

30 November: C1-C21

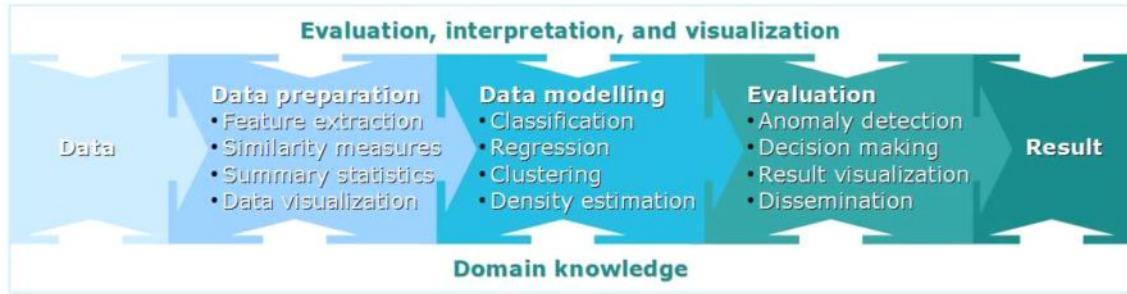
Online help: Forum on DTU Learn

Videos of lectures: <https://video.dtu.dk>

Streaming of lectures: Zoom (link on DTU Learn)

Lecture 6

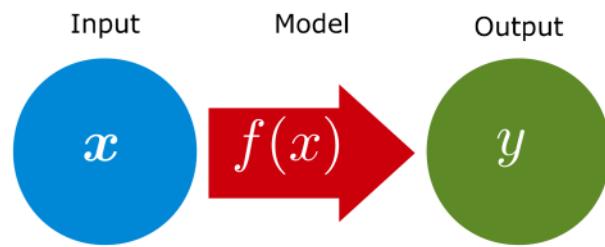
5 October, 2021



Learning Objectives

- Explain the difference between training, test and generalization error
- Explain how cross-validation can be used for (i) performance evaluation (ii) model selection
- Apply forward and backward selection
- Explain how K-Nearest Neighbours can be used to classify data

Supervised learning



- **Mapping between domains**

- Classification: Discrete output
- Regression: Continuous output

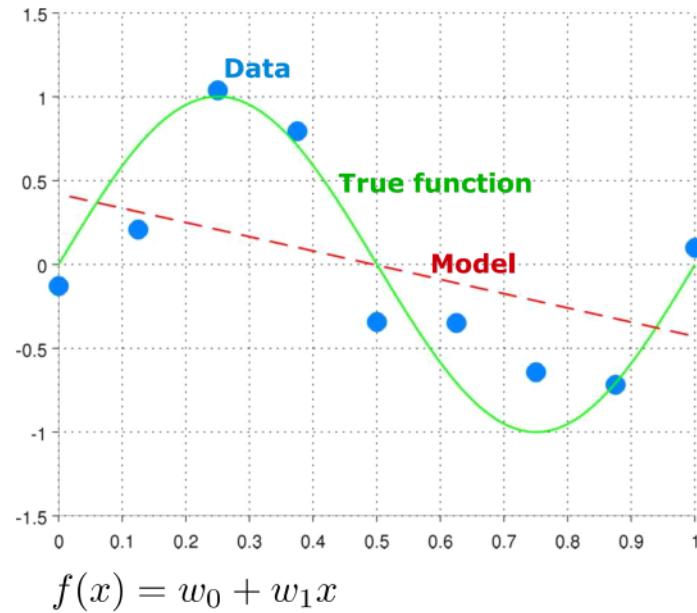
Roadmap

- Introduce errors
 - Training error
 - Test error
 - Generalization error
- Introduce cross-validation
 - **Basic cross validation** for **performance evaluation**
 - **Cross-validation** for **model selection**
 - **Two-level cross-validation** for **model selection and performance evaluation**
- Nearest Neighbor methods

Why are there “multiple models”?

Example: Linear regression

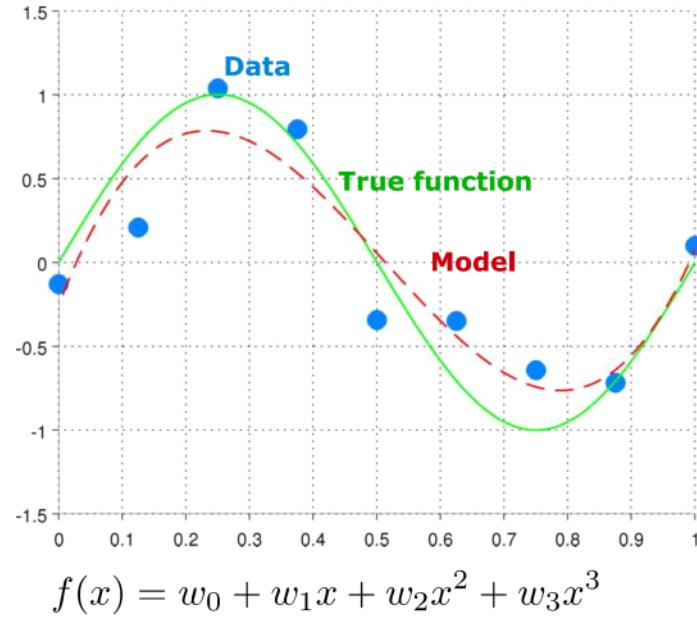
- Bad fit
- **Too simple model**



Why are there “multiple models”?

Example: Linear regression

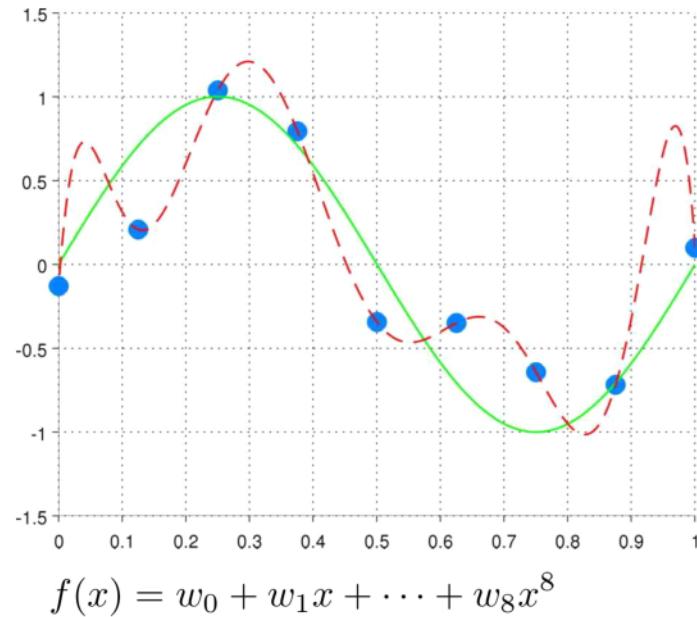
- Reasonable fit
- **Reasonable model**



Why are there “multiple models”?

Example: Linear regression

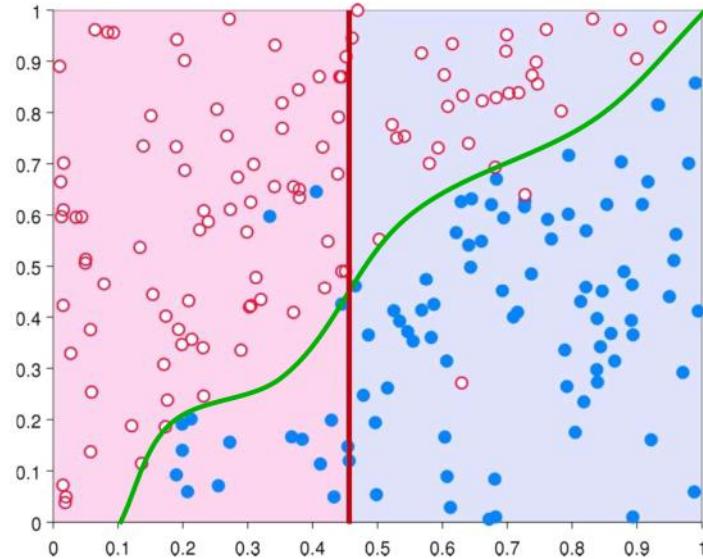
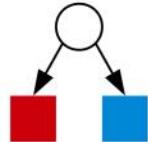
- Perfect fit
- **Too complex model**



Why are there “multiple models”?

Example: Classification tree

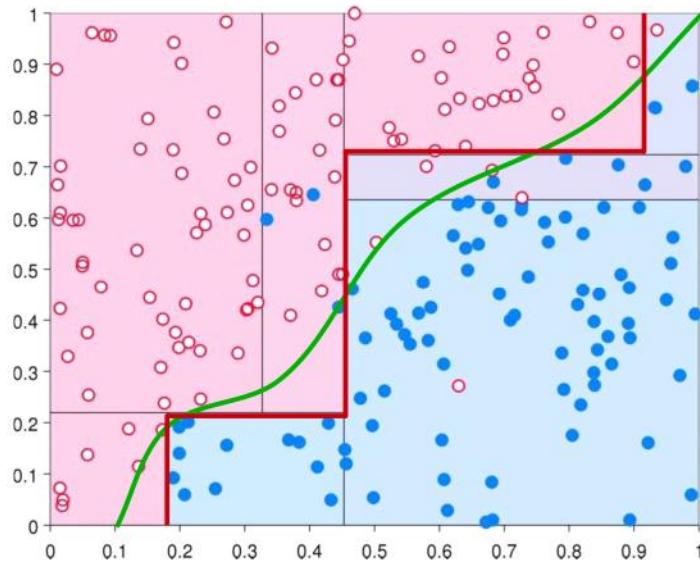
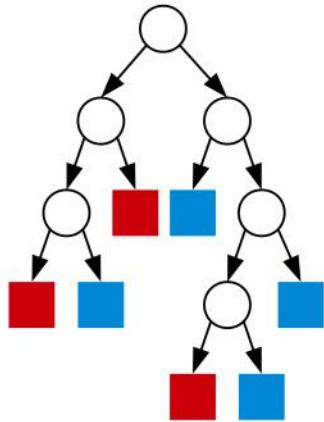
- Bad fit
- **Too simple model**



Why are there “multiple models”?

Example: Classification tree

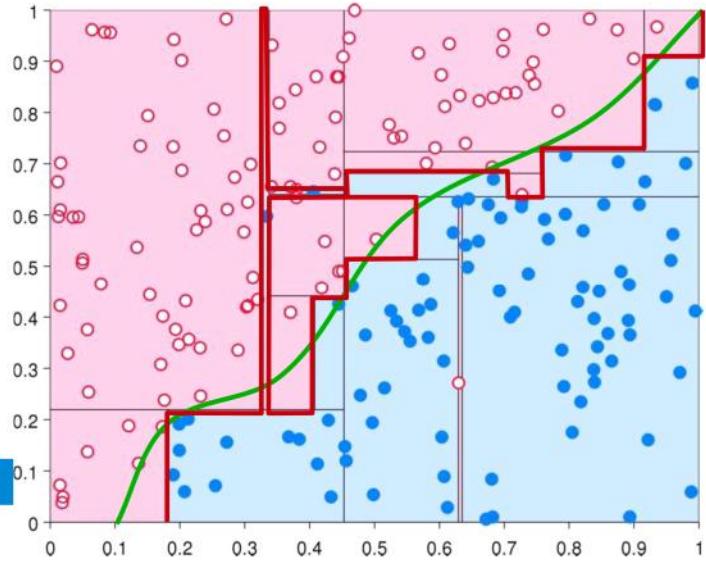
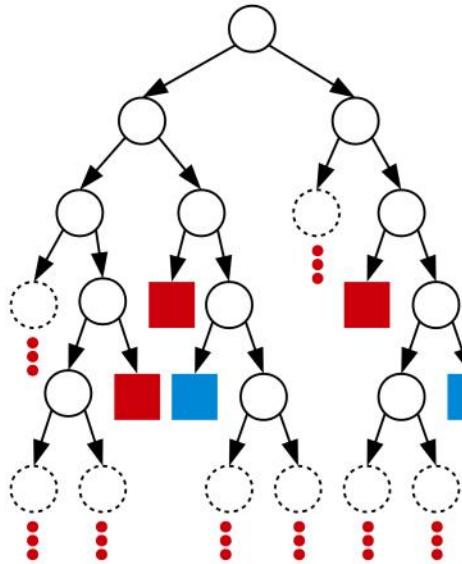
- Reasonable fit
- **Reasonable model**



Why are there “multiple models”?

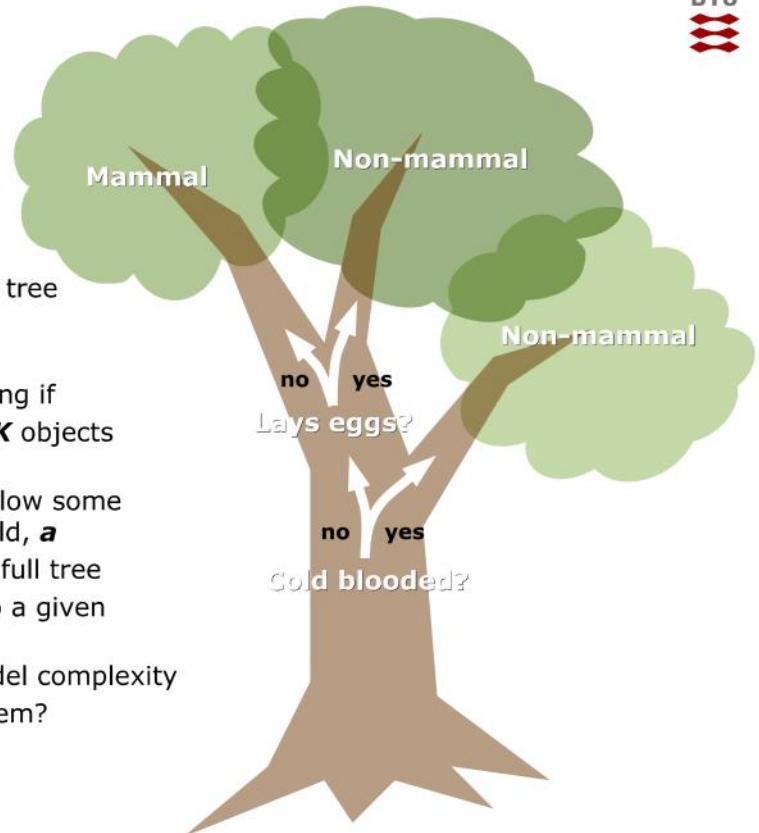
Example: Classification tree

- Perfect fit
- **Too complex model**

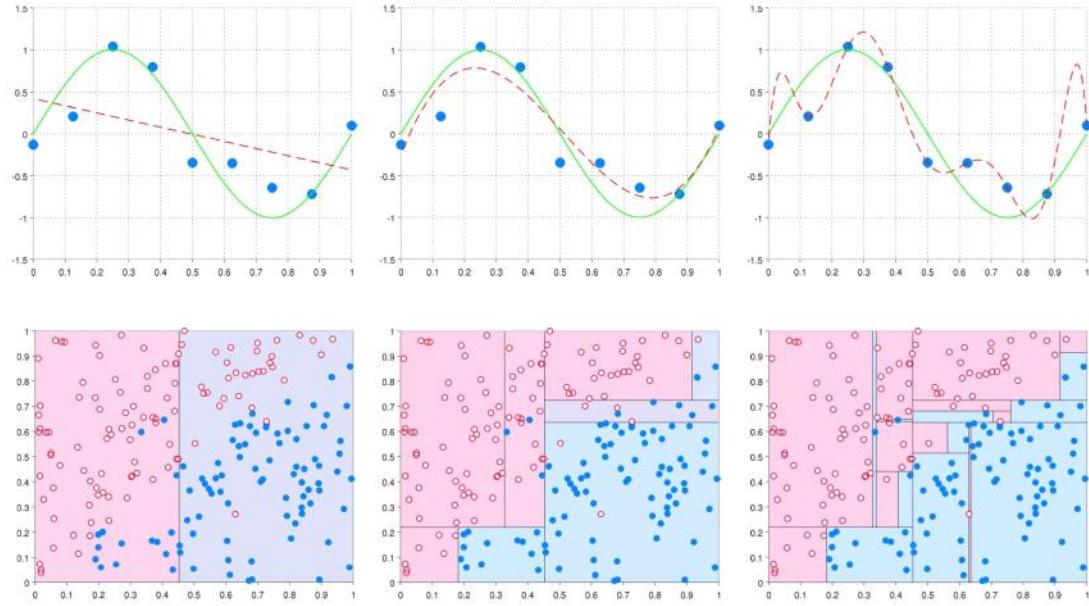


Decision trees

- Hunts algorithm
 - Continue splitting until each node is pure
 - Results in a very complex tree (overfitting)
- **Control complexity**
 - **Pre-pruning:** Stop splitting if
 - There is less than K objects on the branch
 - Impurity gain is below some predefined threshold, a
 - **Post-pruning:** Generate full tree
 - Cut off branches to a given pruning level, c
- K , a , and/or c determine model complexity
 - How should we choose them?

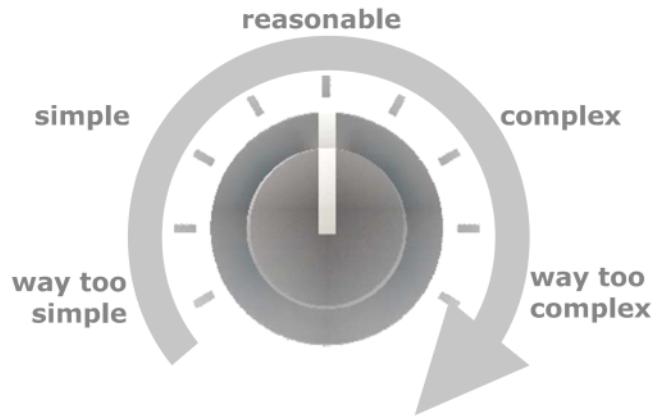


Model overfitting



Control the model complexity

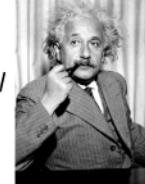
- Find **parameter** or **mechanism** in model that controls complexity



Lex Parsimoniae, Law of parsimony



Given two models with same predictive performance, the simpler model is preferred over the more complex model
- William of Ockham (1288-1347)
(paraphrased)



"Everything should be made as simple as possible, but not simpler" - Einstein

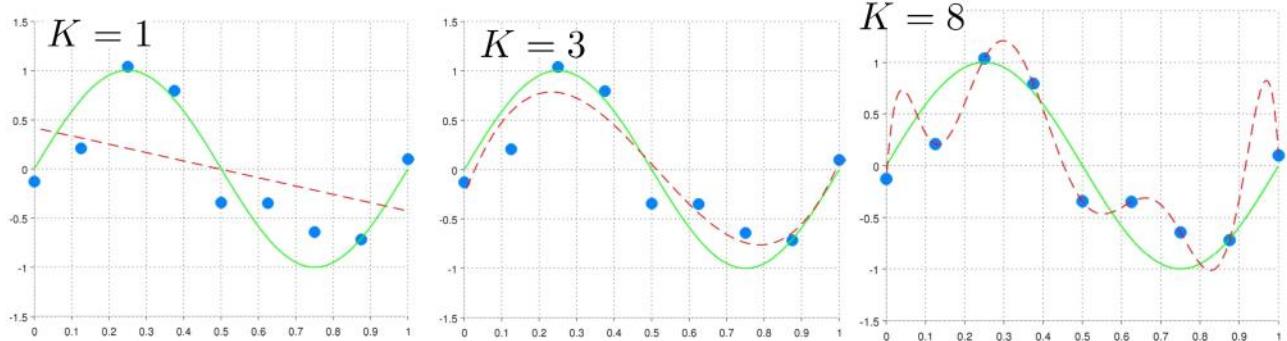
https://commons.wikimedia.org/wiki/File:William_of_Ockham.png

Linear regression

- Linear regression on non-linearly transformed inputs (polynomials)

$$f(x) = w_0 + w_1x + \cdots + w_8x^8$$

– **Control complexity:** Choose a suitable value for K



Solution:
Assess model performance correctly and select best model

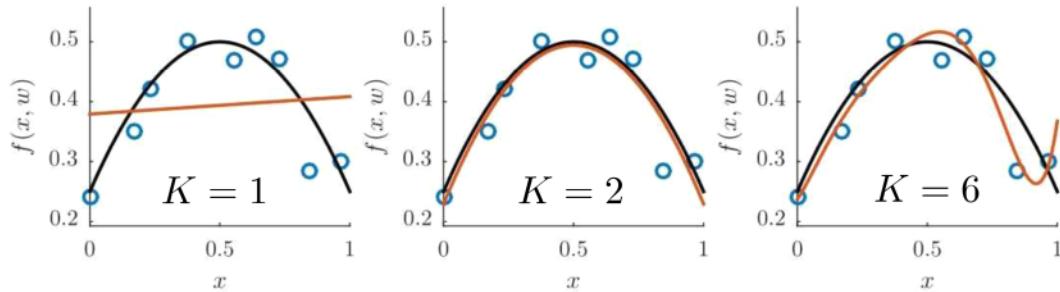
Training error

- Suppose we train 3 models on a dataset of 9 observations

$\mathcal{M}_1 = \{\text{1'st order polynomial}\}$

$\mathcal{M}_2 = \{\text{2'nd order polynomial}\}$

$\mathcal{M}_3 = \{\text{6'th order polynomial}\}$

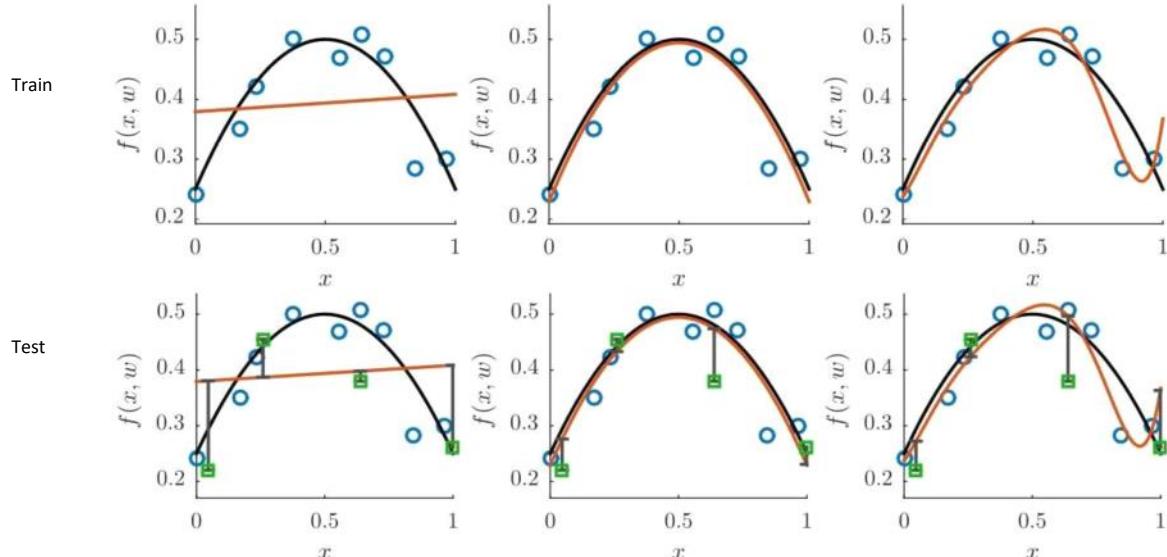


$$E_{\mathcal{M}_k}^{\text{train}} = \frac{1}{N^{\text{train}}} \sum_{i \in \mathcal{D}^{\text{train}}} (y_i - f_{\mathcal{M}_k}(x_i, \mathbf{w}))^2.$$

This is the LOST function $L(y_i, f_m(x_i))$

Test error error

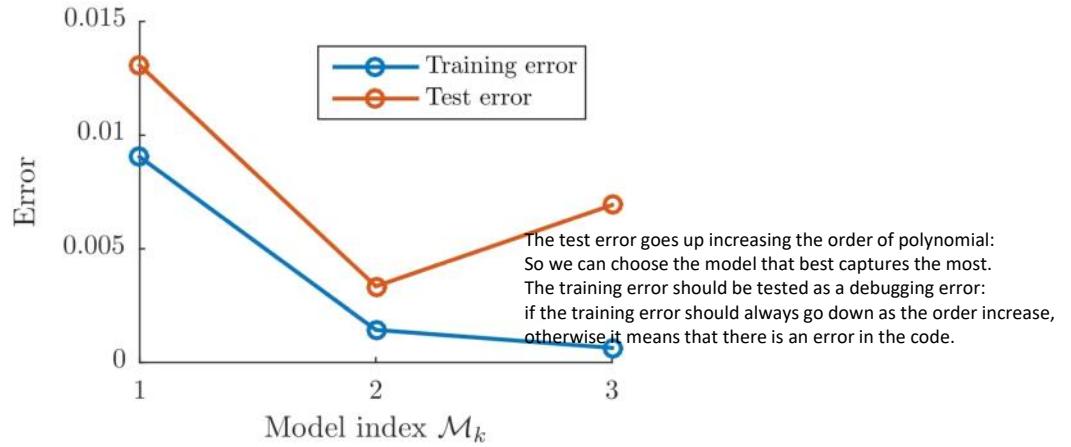
- Test error is obtained by testing the trained models on new data



$$E_{\mathcal{M}_k}^{\text{train}} = \frac{1}{N^{\text{train}}} \sum_{i \in \mathcal{D}^{\text{train}}} (y_i - f_{\mathcal{M}_k}(x_i, \mathbf{w}))^2.$$

$$E_{\mathcal{M}_k}^{\text{test}} = \frac{1}{N^{\text{test}}} \sum_{i \in \mathcal{D}^{\text{test}}} (y_i - f_{\mathcal{M}_k}(x_i, \mathbf{w}))^2$$

Overfitting

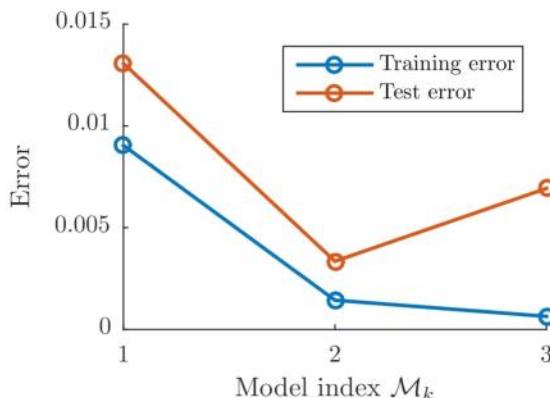


- **Overfitting** is that the training error usually decreases for overly complex models while the test error increases
- Test error is the more true error
- **Never, ever validate a model on the same data it was trained upon**

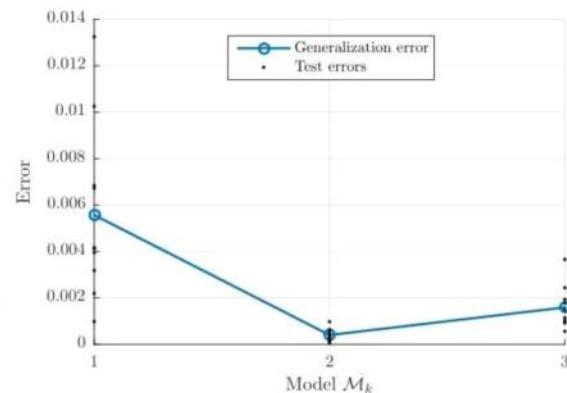
Generalization error

- The generalization error is the test error evaluated over an infinitely large test set
- The generalization error is the "true performance" of the trained model
 - Train model \mathcal{M} on the available dataset \mathcal{D} to get prediction rule $f_{\mathcal{M}}$
 - Compute $E_{\mathcal{M}}^{\text{gen}} = \mathbb{E}_{(\mathbf{x}, \mathbf{y})} [L(\mathbf{y}, f_{\mathcal{M}}(\mathbf{x}))]$ expectation over the loss function: $\int \int L(Y, f_{\mathcal{M}}(x)) \rho(x, y) dxdy$
- If we somehow had many test sets $\mathcal{D}_1, \dots, \mathcal{D}_K$

$$E_{\mathcal{M}}^{\text{gen}} \approx \frac{1}{K} \sum_{k=1}^K E_{\mathcal{M}, \mathcal{D}_k}^{\text{test}}$$



20 DTU Compute



Lecture 6 5 October, 2021

Basic cross-validation

- **Purpose:** Estimate the generalization error

Basic cross-validation

- **Purpose:** Estimate the generalization error
- 3 variants:

– **Holdout:** Partitions dataset in two (training, test), approximate the generalization error based on the generated test set

$$\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{test}}$$

$$E_{\mathcal{M}}^{\text{gen}} \approx E_{\mathcal{M}}^{\text{test}}$$



Basic cross-validation

- **Purpose:** Estimate the generalization error
- 3 variants:

– **Holdout:** Partitions dataset in two (training, test), approximate the generalization error based on the generated test set

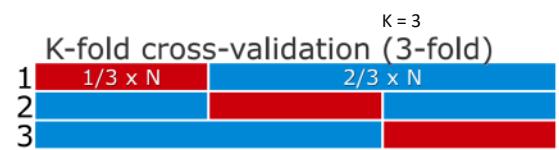
$$\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{test}}$$

$$E_{\mathcal{M}}^{\text{gen}} \approx E_{\mathcal{M}}^{\text{test}}$$

– **K-fold:** Partitions dataset in K parts. Each part is a test set and the other K-1 training sets

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_K$$

$$E_{\mathcal{M}}^{\text{gen}} \approx \frac{1}{K} \sum_{k=1}^K E_{\mathcal{M},k}^{\text{test}}$$



Basic cross-validation

- **Purpose:** Estimate the generalization error
- 3 variants:

– **Holdout:** Partitions dataset in two (training, test), approximate the generalization error based on the generated test set

$$\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{test}}$$

$$E_{\mathcal{M}}^{\text{gen}} \approx E_{\mathcal{M}}^{\text{test}}$$

– **K-fold:** Partitions dataset in K parts. Each part is a test set and the other K-1 training sets

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_K$$

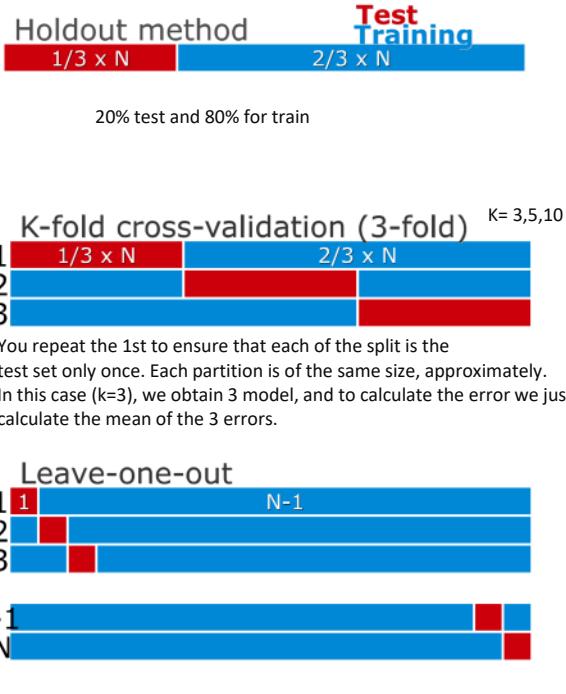
$$E_{\mathcal{M}}^{\text{gen}} \approx \frac{1}{K} \sum_{k=1}^K E_{\mathcal{M},k}^{\text{test}}$$

– **Leave-one-out:** Partitions dataset into N parts. Let each observation be a test set and the other N-1 training sets (K-fold with K=N)

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_N$$

$$E_{\mathcal{M}}^{\text{gen}} \approx \frac{1}{N} \sum_{k=1}^N E_{\mathcal{M},k}^{\text{test}}$$

24 DTU Compute



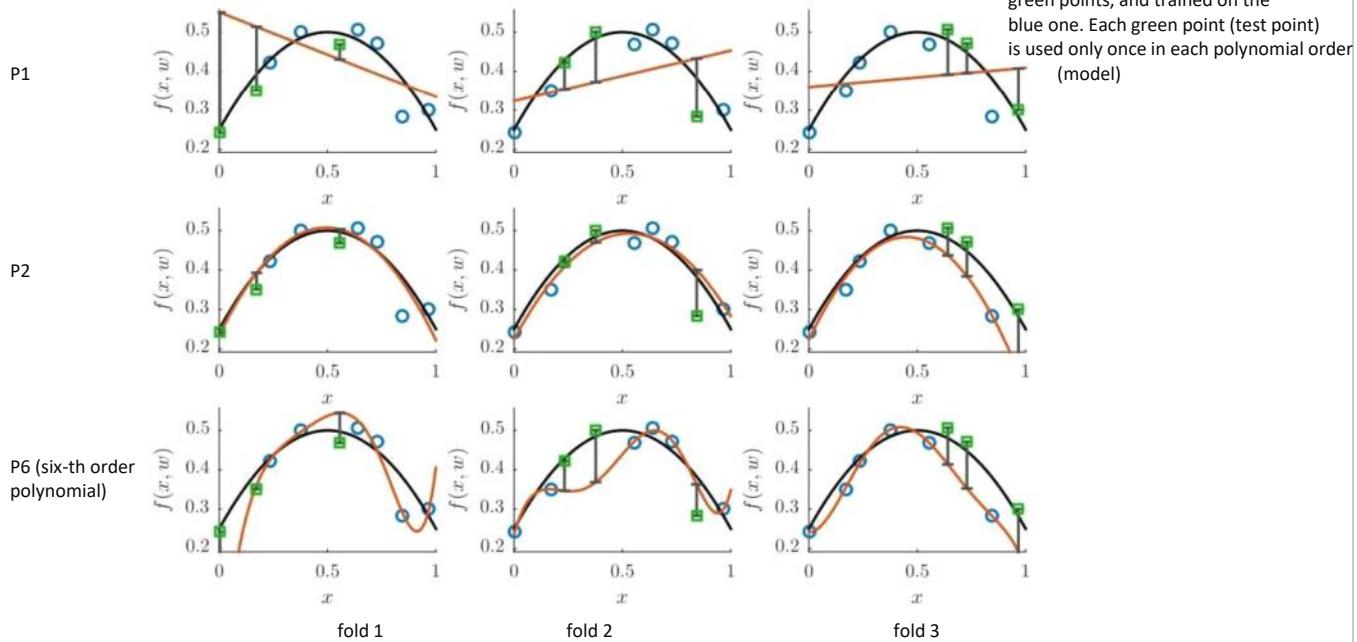
Lecture 6 5 October, 2021

Cross-validation (1-layer)

- K=3 fold cross-validation for the three Linear-regression models

Vertically: The three models

Horizontally: The three cross-validation folds



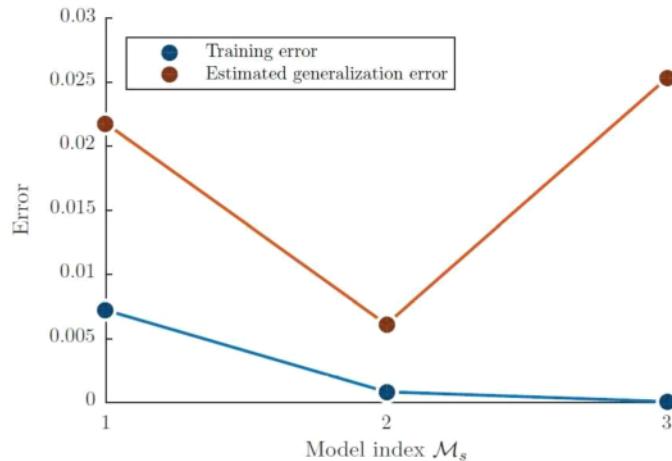
Cross-validation for model selection (1-layer)

- Purpose: Select the best of S models
- The idea:

- For each model, estimate the cross-validation error $\hat{E}_{\mathcal{M}_1}^{\text{gen}}, \dots, \hat{E}_{\mathcal{M}_S}^{\text{gen}}$ using basic cross-validation.
- Select the optimal model \mathcal{M}_{s^*} as that with the lowest error:

$$s^* = \arg \min_s \hat{E}_{\mathcal{M}_s}^{\text{gen}}$$

you compute the generalization error for all the models, and than you choose the one with the lowest one.



Cross-validation (1-layer)

- K-fold cross-validation for model selection, the algorithm

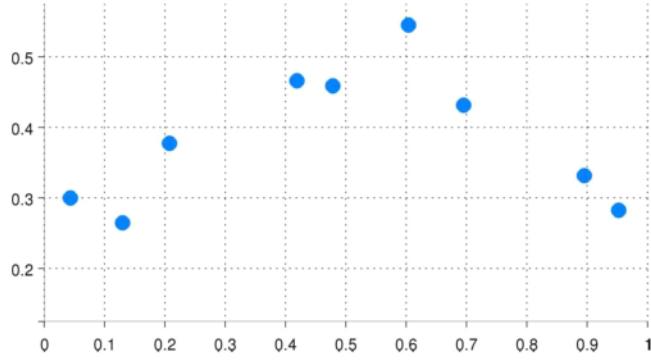
Algorithm 4: *K*-fold cross-validation for model selection

Require: K , the number of folds in the cross-validation loop
Require: $\mathcal{M}_1, \dots, \mathcal{M}_S$. The S different models to select between
Ensure: \mathcal{M}_{s^*} the optimal model suggested by cross-validation

```
for  $k = 1, \dots, K$  splits do
    Let  $\mathcal{D}_k^{\text{train}}, \mathcal{D}_k^{\text{test}}$  the  $k$ 'th split of  $\mathcal{D}$ 
    for  $s = 1, \dots, S$  models do
        Train model  $\mathcal{M}_s$  on the data  $\mathcal{D}_k^{\text{train}}$ 
        Let  $E_{\mathcal{M}_s, k}^{\text{test}}$  be the test error of the model  $\mathcal{M}_s$  when it is tested on  $\mathcal{D}_k^{\text{test}}$ 
    end for
end for
For each  $s$  compute:  $\hat{E}_{\mathcal{M}_s}^{\text{gen}} = \sum_{k=1}^K \frac{N_k^{\text{test}}}{N} E_{\mathcal{M}_s, k}^{\text{test}}$ 
Select the optimal model:  $s^* = \arg \min_s \hat{E}_{\mathcal{M}_s}^{\text{gen}}$ 
 $\mathcal{M}_{s^*}$  is now the optimal model suggested by cross-validation
```

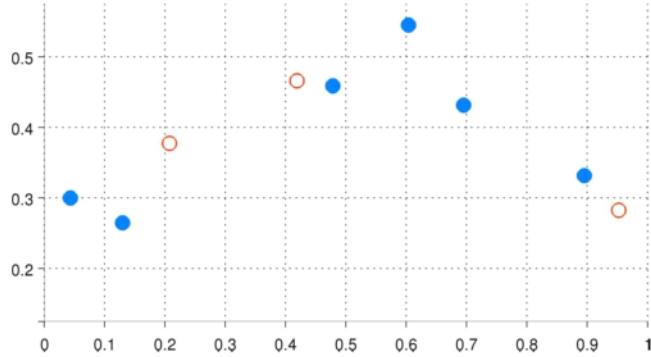
Holdout method

- Randomly choose a subset of data points to be in a **test set**
 - For example choose 1/3 of the points
- The rest is the **training set**



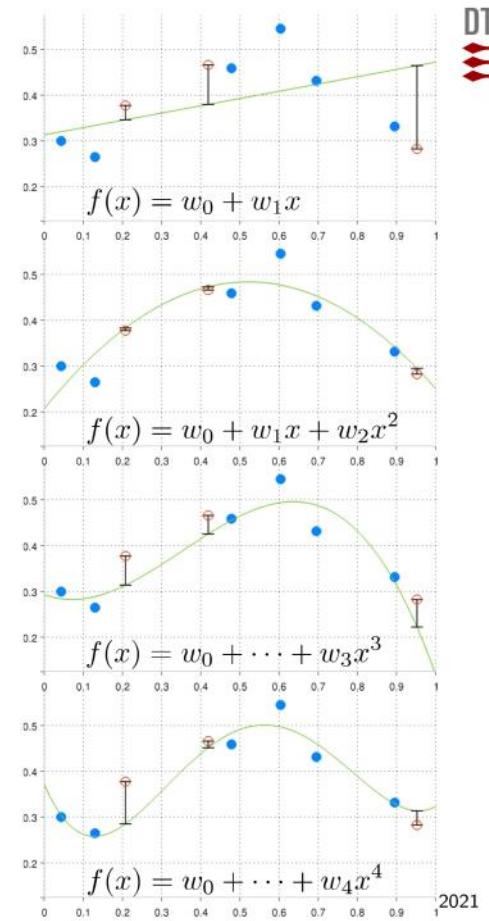
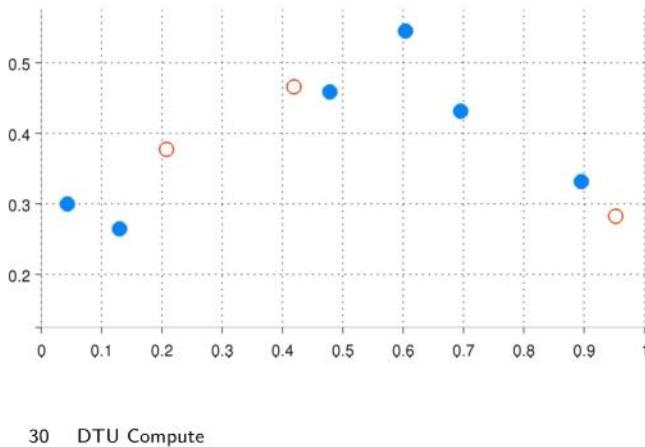
Holdout method

- Randomly choose a subset of data point to be in a **test set**
 - For example choose 1/3 of the points
- The rest is the **training set**



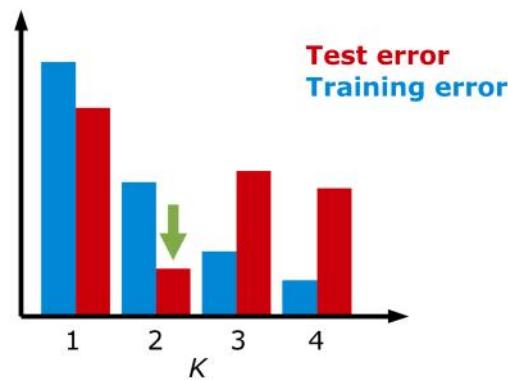
Holdout method

- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- Choose the model with lowest **test error**

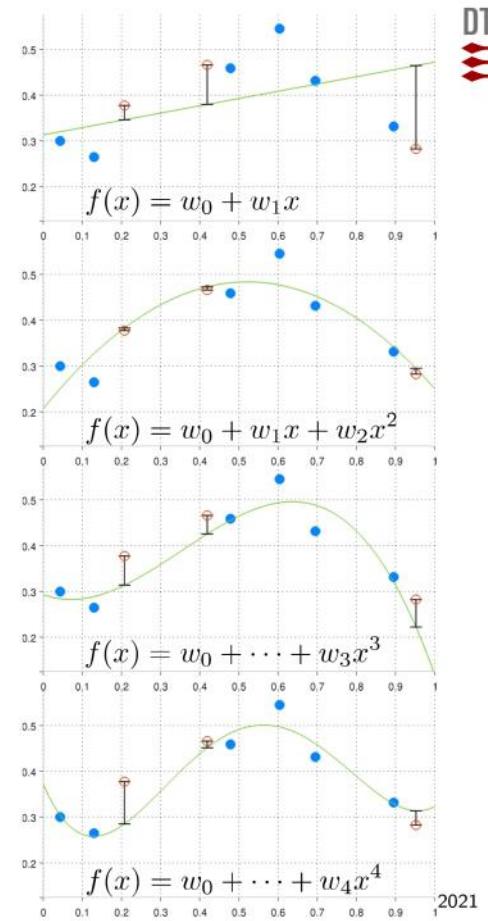


Holdout method

- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- Choose the model with lowest **test error**



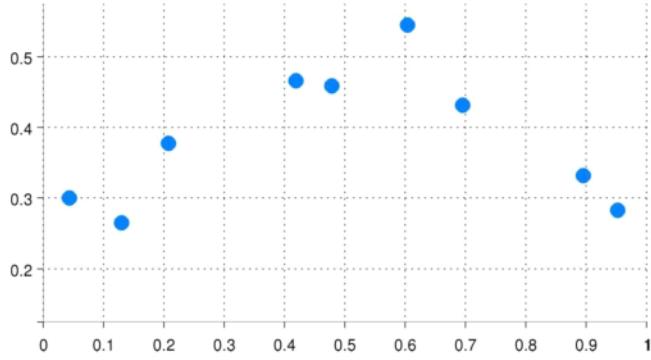
31 DTU Compute



2021

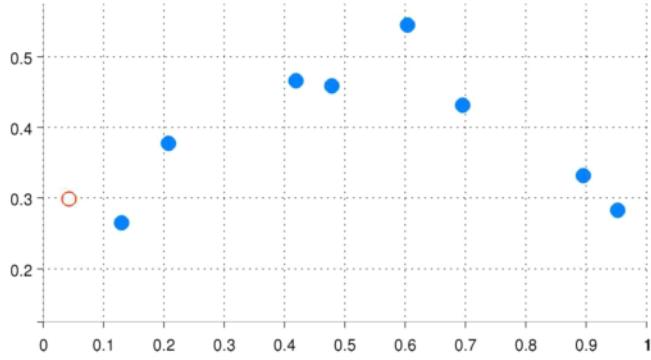
Leave-one-out

- Choose the first data point as a **test set**
- The rest is the **training set**



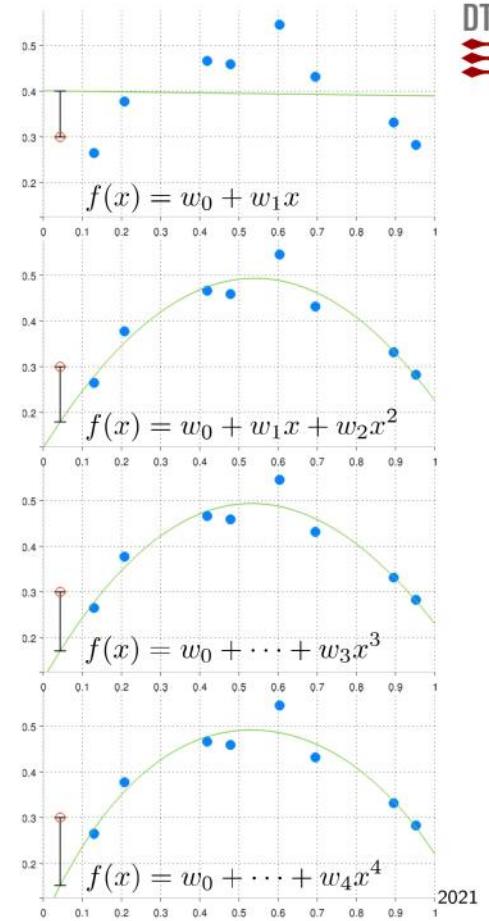
Leave-one-out

- Choose the first data point as a **test set**
- The rest is the **training set**

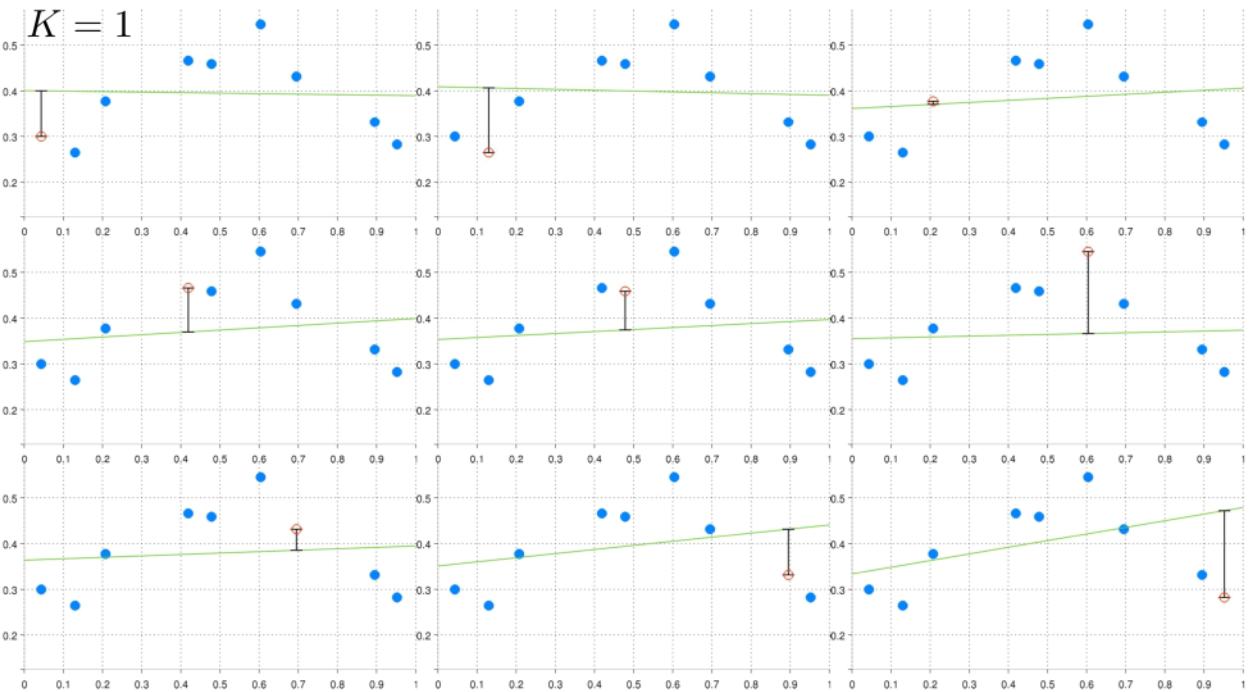


Leave-one-out

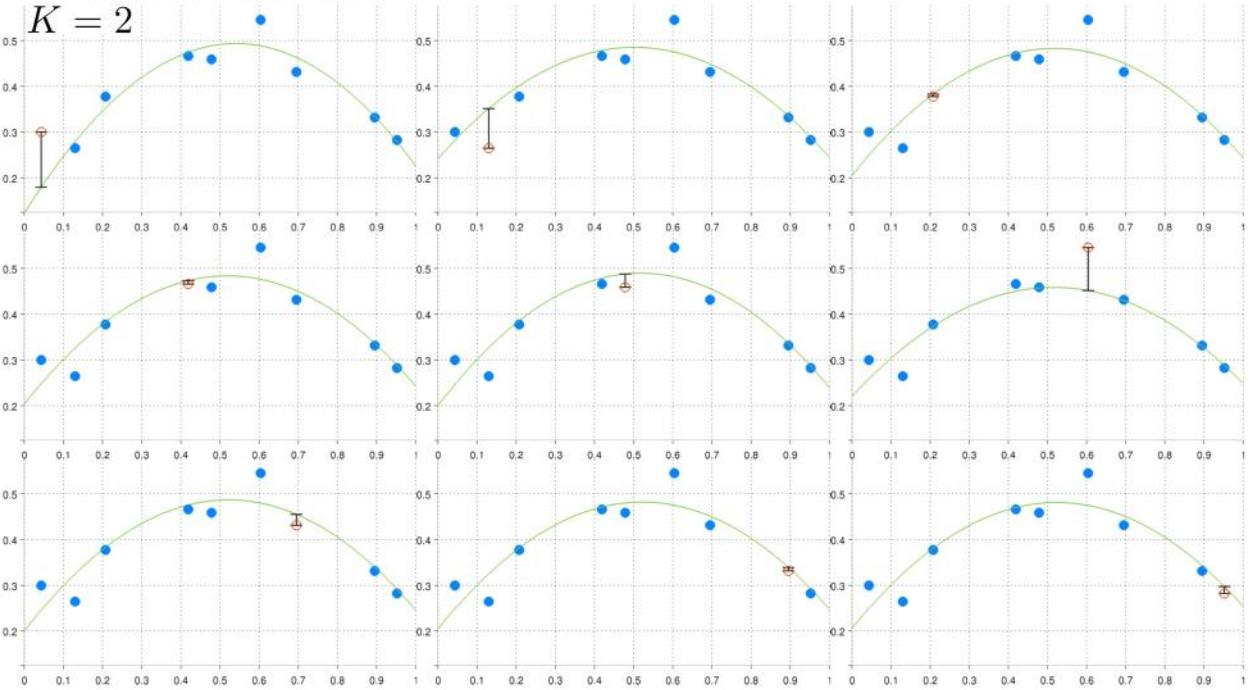
- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- **Repeat for all data points**
 - All data points get to be test set
 - Compute **average test error**



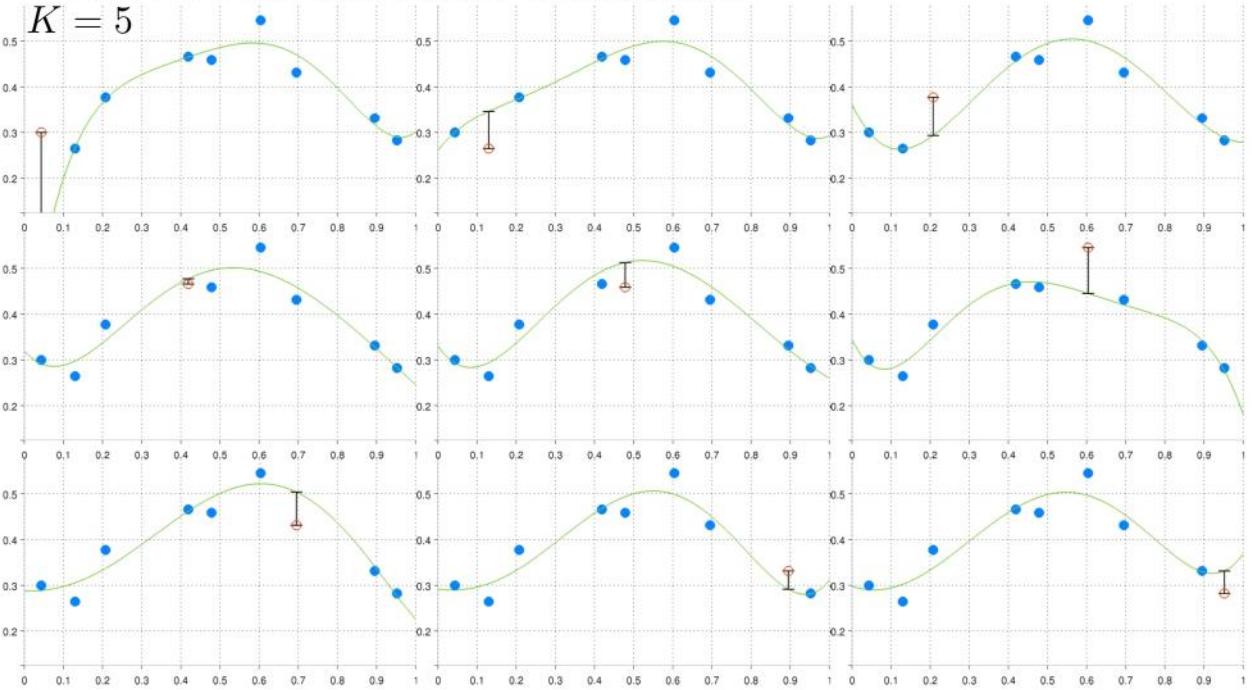
Leave-one-out



Leave-one-out

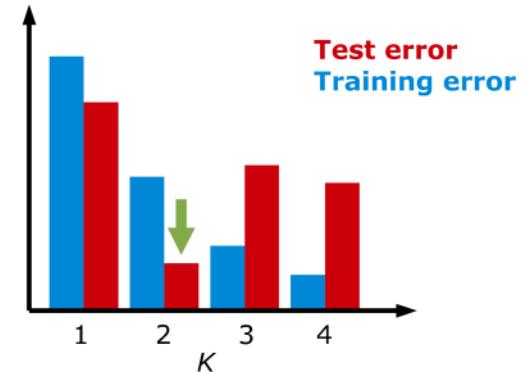


Leave-one-out cross-validation



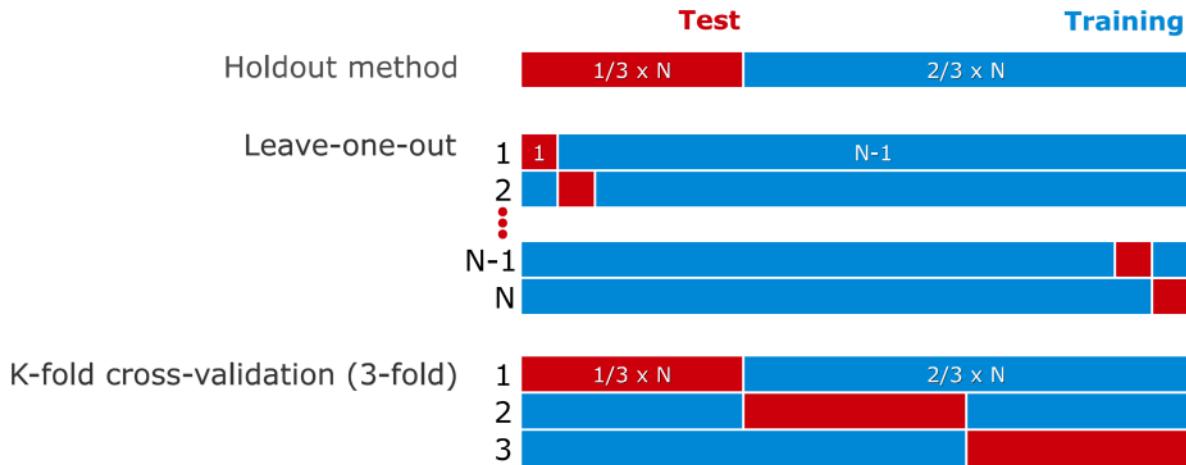
Leave-one-out

- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- **Repeat for all data points**
 - All data points get to be test set
 - Compute **average test error**



So which method should I choose?

- Holdout is computationally least intensive, but not very data efficient
- Leave-one-out is very computationally intensive, and the estimates are highly correlated
- Recommendation: K -fold for $K = 5, 10$ or holdout if problem very large.



Quiz 1, Cross validation (Spring 2012)

Feature(s)	Training MSE	Test MSE
A	2.0	2.2
B	1.8	1.9
C	1.6	1.7
D	1.9	2.1
A and B	1.7	2.0
A and C	1.3	1.8
A and D	1.4	1.5
B and C	1.5	1.6
B and D	1.7	1.8
C and D	1.5	2.0
A and B and C	1.2	2.1
A and B and D	1.1	2.0
A and C and D	1.0	2.3
B and C and D	1.2	2.5
A and B and C and D	0.9	2.8

Consider a neural network regression problem with four attributes denoted A, B, C and D. A neural network with two hidden units is trained using different combinations of the attributes. The neural network is trained on 50% of the data and tested on the remaining 50% of the data using the hold-out method. In the table is given the training and test performance of the neural network for the different combinations of attributes. Which of the following statements is *incorrect*

- A. Hold-out 50% of the data is more computationally efficient than 5 fold cross-validation.
- B. Leave one-out cross-validation gives a poor estimate of the generalization error as only one observation is part of the test set at a time.
- C. The size of the training set in 10 fold cross-validation is larger than the size of the training set in 5 fold cross-validation.
- D. Not all observations are used for testing using the hold-out method.
- E. Don't know.

Forward selection

- Suppose we want to do linear regression
- As usual, we have M attributes

$$f(x) = w_0$$

$$f(x) = w_0 + w_1x_1 + w_2x_{27} + w_3x_{88}$$

$$f(x) = w_0 + w_1x_{19} + w_2x_{76}$$

$$f(x) = w_0 + w_1x_{19} + w_2x_{76} + w_3x_{88}$$

$$f(x) = w_0 + w_1x_1 + w_2x_{27} + w_3x_{19}$$

$$f(x) = w_0 + w_1x_{27} + w_2x_{88}$$

$$x_1, x_2, \dots, x_M$$

⋮

- We can control model complexity by using a subset of attributes
 - Large subset: Complex model; hard to interpret
 - Small subset: Too simple model
- In general, we can construct 2^M models; often far too many
- Sequential feature selection allow us to efficiently search the model space

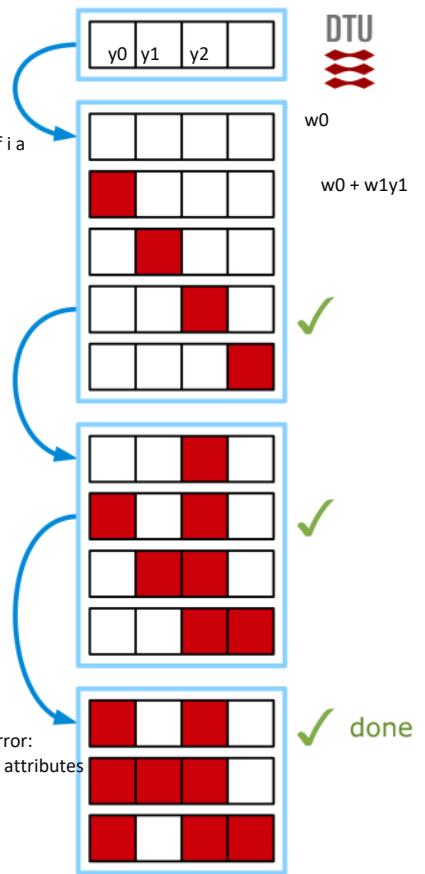
it's a greedy approach, which is good if the feature are not correlated, to select the best ones.

Sequential feature selection

starting from model w_0 : no features:
you then ask how well will my model perform if i add features 1? -> w_1y_1

Forward selection

- Start with no features
- Compute **cross-validation error** for
 - Current feature subset
 - All subsets equal to the current + one added feature
- Choose best subset
- Repeat until no further improvement

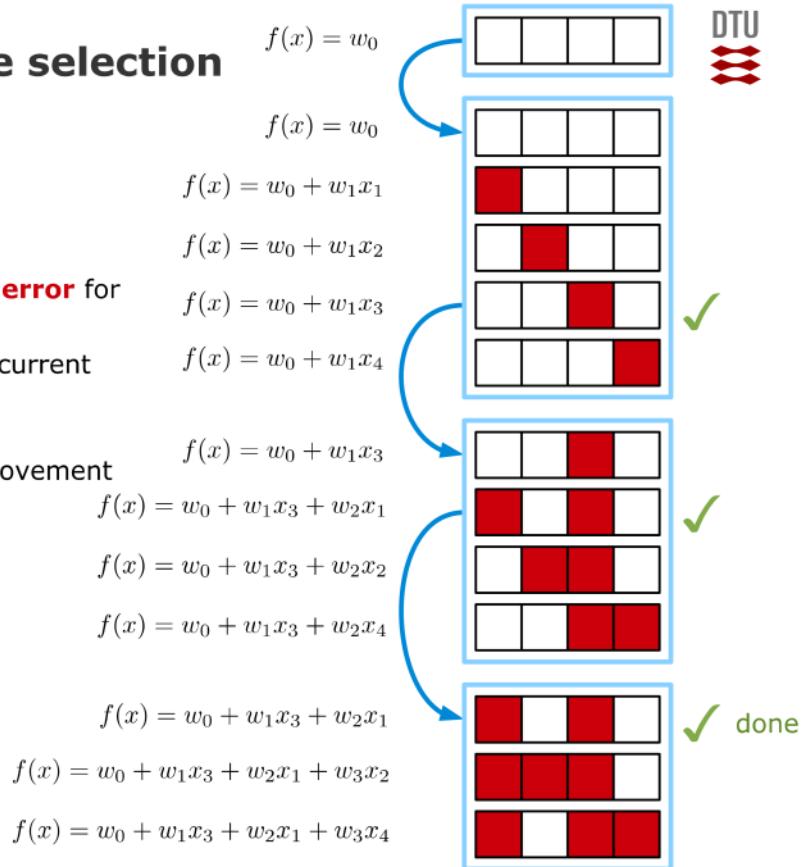


Sequential feature selection



Forward selection

- Start with no features
- Compute **cross-validation error** for
 - Current feature subset
 - All subsets equal to the current + one added feature
- Choose best subset
- Repeat until no further improvement

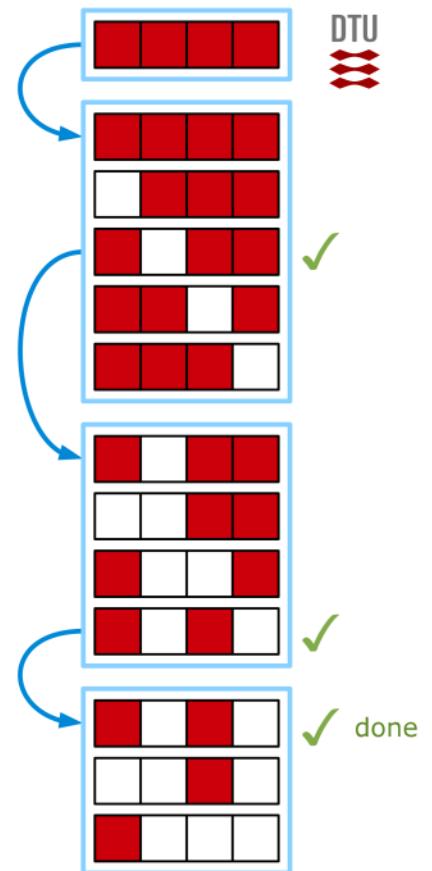


Sequential feature selection

we can also go backward, starting from the testing error instead of the training error

Backward selection

- Start with all features
- Compute **cross-validation error** for
 - Current feature subset
 - All subsets equal to the current - one removed feature
- Choose best subset
- Repeat until no further improvement



Quiz 2, Forward selection (Spring 2012)

Feature(s)	Training MSE	Test MSE
A	2.0	2.2
B	1.8	1.9
C	1.6	1.7
D	1.9	2.1
A and B	1.7	2.0
A and C	1.3	1.8
A and D	1.4	1.5
B and C	1.5	1.6
B and D	1.7	1.8
C and D	1.5	2.0
A and B and C	1.2	2.1
A and B and D	1.1	2.0
A and C and D	1.0	2.3
B and C and D	1.2	2.5
A and B and C and D	0.9	2.8

Consider a neural network regression problem with four attributes denoted A, B, C and D where a neural network with two hidden units is trained using different combinations of the attributes. The table gives the training and test performance of the neural network for different combinations of attributes. Which of the following statements is *correct*?

- A. Using a forward selection strategy feature B and C would be selected as the optimal model.
- B. Using a forward selection strategy features A and D would be selected as the optimal model.
- C. Using a forward selection strategy features A and C and D would be selected as the optimal model.
- D. Using a forward selection strategy features A and B and C would be selected as the optimal model.
- E. Don't know.

A correct answer:

you choose the lowest error on test data between the single attributes and then you look for the best combination that include that attributes and the others.

What could be problem with this 1-level cross-validation?

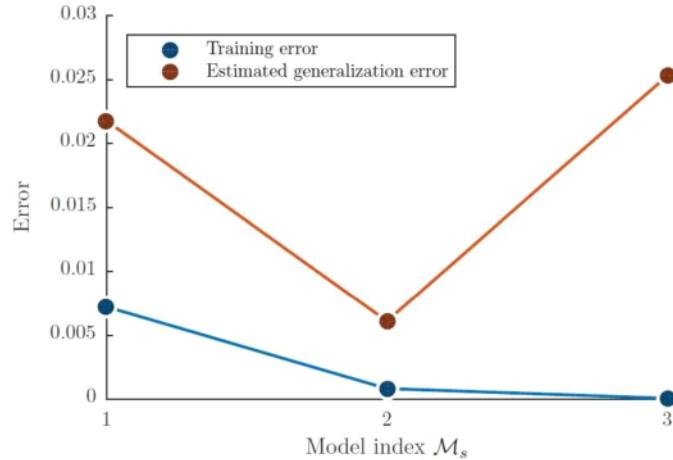
if we now have picked a model, how good we expect the generalization error to be compared to the one obtained from the test? -> it's the best scenario, so the generalization error results to be over-estimated



A problem with 1 level cross-validation?

- For each model, estimate the cross-validation error $\hat{E}_{\mathcal{M}_1}^{\text{gen}}, \dots, \hat{E}_{\mathcal{M}_S}^{\text{gen}}$ using basic cross-validation.
- Select the optimal model \mathcal{M}_{s^*} as that with the lowest error:

$$s^* = \arg \min_s \hat{E}_{\mathcal{M}_s}^{\text{gen}}$$



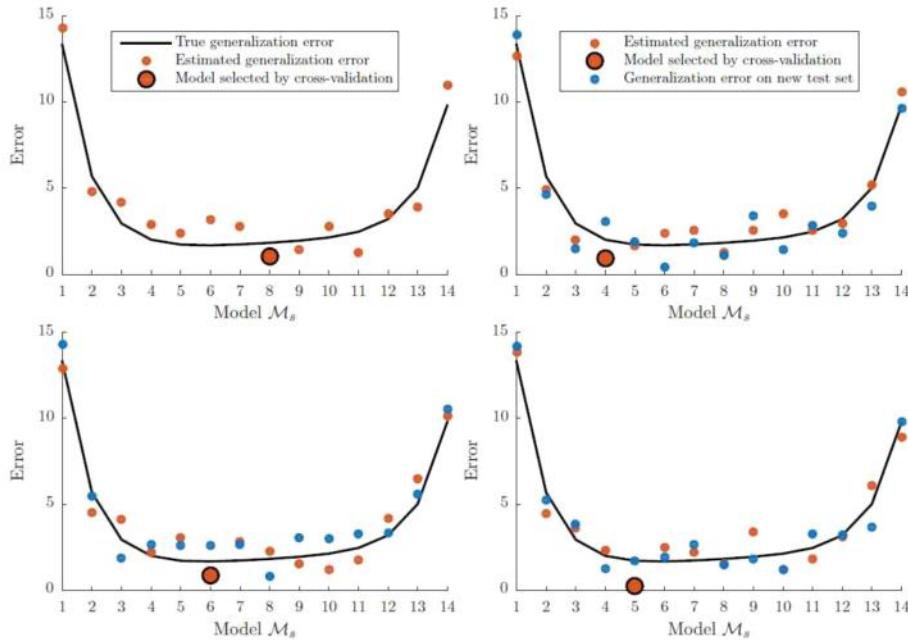
- Is the generalization error the selected model ($k=2$) about 0.007?

We have 14 models, and we have the true gen. error (black line) -> we will pick the model with the lowest gen. error (the point rounded in black). -> the model wil not performed as desired:
 how can we decide the generalization error? -> cross-validation again on the generalization error.
 we get better estimate of the ge. error -> TWO LAYER CROSS-VALIDATION



A problem with 1 level cross-validation?

- Same as before, just with more models. Is the error of the red dot a fair estimate of the generalization error?



Two-layer cross-validation

- **Purpose:** Select optimal model and estimate generalization error of optimal model

Two-layer cross-validation

- **Purpose:** Select optimal model and estimate generalization error of optimal model
- **How?**
 - Recall "**one layer cross-validation for model selection**"
 - This method returns a model (the best model)
 - We can consider "**one-layer cross-validation for model selection**" as a single model

Two-layer cross-validation

- **Purpose:** Select optimal model and estimate generalization error of optimal model
- **How?**
 - Recall "**one layer cross-validation for model selection**"
 - This method returns a model (the best model)
 - We can consider "**one-layer cross-validation for model selection**" as a single model
- **Recall:**
 - "**Basic cross-validation for performance evaluation**" estimates the generalization error of a model

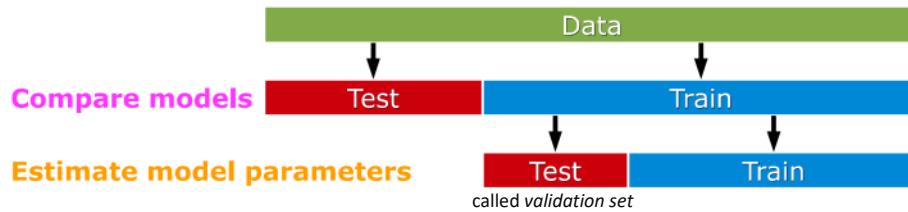
We have our data.

- 1- split data into train and test
- 2- further, split the train into test and train again



Two-layer cross-validation

- **Purpose:** Select optimal model and estimate generalization error of optimal model
- **How?**
 - Recall "**one layer cross-validation for model selection**"
 - This method returns a model (the best model)
 - We can consider "**one-layer cross-validation for model selection**" as a single model
- **Recall:**
 - "**Basic cross-validation for performance evaluation**" estimates the generalization error of a model
- **Idea:** Apply "**basic cross-validation for performance evaluation**" on the "**one-layer cross-validation for model selection**"-model to estimate its generalization error



Cross-validation (2-layer)

- Two-layer cross-validation, the algorithm

Algorithm 5: Two-level cross-validation

$K=10$ 10
Require: K_1, K_2 , folds in outer, and inner cross-validation loop respectively

Require: $\mathcal{M}_1, \dots, \mathcal{M}_S$: The S different models to cross-validate

Ensure: \hat{E}^{gen} , the estimate of the generalization error

for $i = 1, \dots, K_1$ do

 Outer cross-validation loop. First make the outer split into K_1 folds
 Let $\mathcal{D}_i^{\text{par}}, \mathcal{D}_i^{\text{test}}$ be the i 'th split of \mathcal{D}

 for $j = 1, \dots, K_2$ do

 Inner cross-validation loop. Use cross-validation to select optimal model

 Let $\mathcal{D}_j^{\text{train}}, \mathcal{D}_j^{\text{val}}$ be the j 'th split of $\mathcal{D}_i^{\text{par}}$

 for $s = 1, \dots, S$ do

 Train \mathcal{M}_s on $\mathcal{D}_j^{\text{train}}$

 Let $E_{\mathcal{M}_s, j}^{\text{val}}$ be the validation error of the model \mathcal{M}_s when it is tested on $\mathcal{D}_j^{\text{val}}$

 end for

 end for

 For each s compute: $\hat{E}_s^{\text{gen}} = \sum_{j=1}^{K_2} \frac{|\mathcal{D}_j^{\text{val}}|}{|\mathcal{D}_i^{\text{par}}|} E_{\mathcal{M}_s, j}^{\text{val}}$

 Select the optimal model $\mathcal{M}^* = \mathcal{M}_{s^*}$ where $s^* = \arg \min_s \hat{E}_s^{\text{gen}}$

 Train \mathcal{M}^* on $\mathcal{D}_i^{\text{par}}$

 Let E_i^{test} be the test error of the model \mathcal{M}^* when it is tested on $\mathcal{D}_i^{\text{test}}$

end for

Compute the estimate of the generalization error: $\hat{E}^{\text{gen}} = \sum_{i=1}^{K_1} \frac{|\mathcal{D}_i^{\text{test}}|}{N} E_i^{\text{test}}$

Quiz 3, two-level cross-validation (Spring 2016)

Consider a classification tree model applied to a dataset of $N = 1000$ observations. Suppose we wish to both select the optimal pruning level and estimate the generalization error of the classification tree model by cross-validation. To simplify the problem, we only consider 3 possible pruning levels:

3, 4, 5.

We opt for a two-level cross-validation strategy in which we use an inner loop of $K_2 = 5$ -fold cross-validation to estimate the optimal pruning level and an outer loop of $K_1 = 10$ fold cross-validation to estimate the generalization error. That is, for each of the K_1 outer folds, the dataset is divided into

a validation set and a parameter estimation set on which K_2 -fold cross-validation is used to select the optimal pruning level for this outer fold.

How many models do we *train* using 2-level cross-validation?

A. 50

B. 150

C. 160

$$K_1(K_2 \cdot L + 1) = 10 * (5 * 3 + 1)$$

L: number of model

D. 180

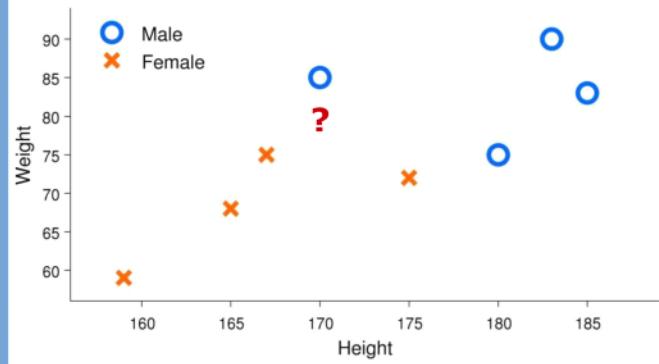
E. Don't know.

Want to identify something: look at what is most similar to it -> nearest neighbour



Classify gender based on height and weight

	Height	Weight	Gender
1	183	90	Male
2	180	75	Male
3	170	85	Male
4	185	83	Male
5	159	59	Female
6	167	75	Female
7	165	68	Female
8	175	72	Female
9	171	82	?

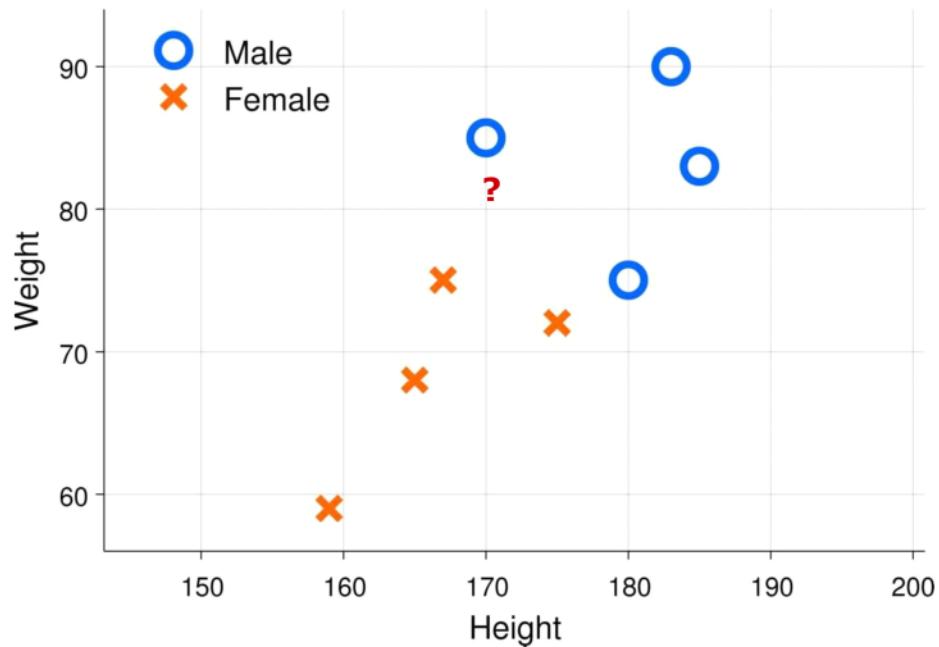


We have two features: height and weight.
Suppose we have a new person: is this a male or female? -> to what in my database it looks like?
The parameters 'K' is what impose the model complexity.



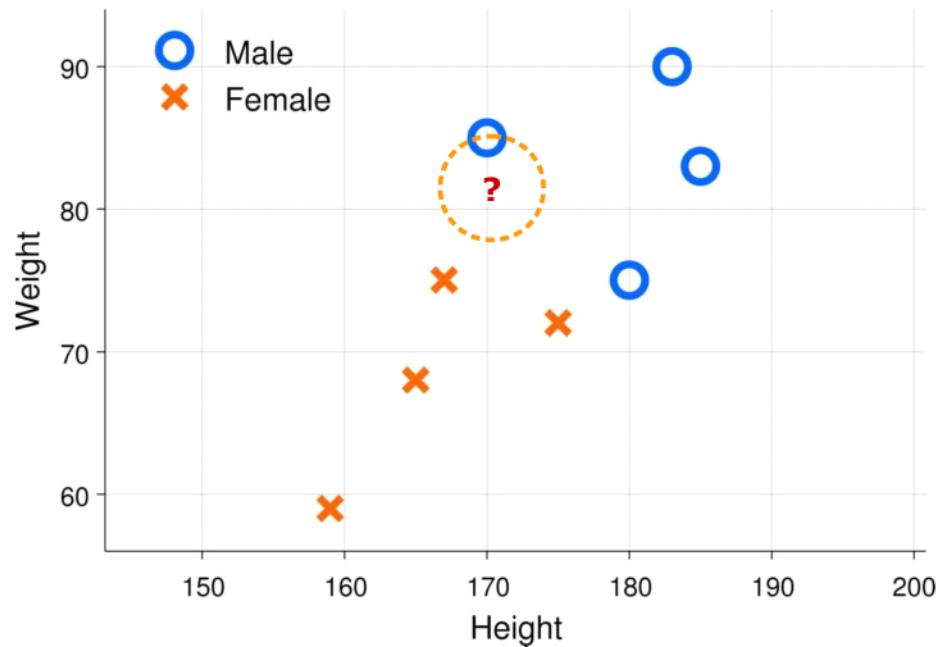
Nearest neighbor classifier

- 1 nearest neighbor



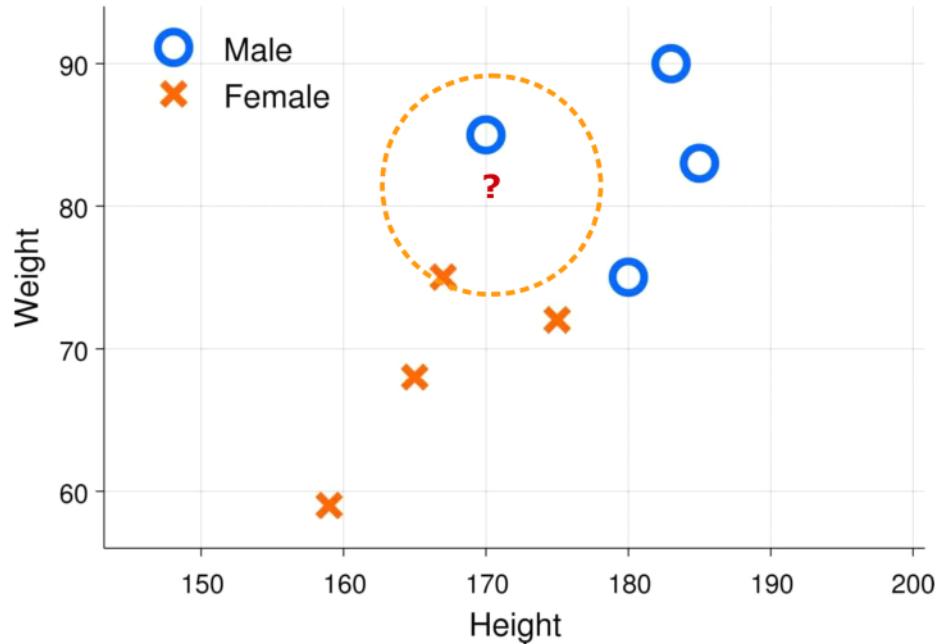
Nearest neighbor classifier

- 1 nearest neighbor



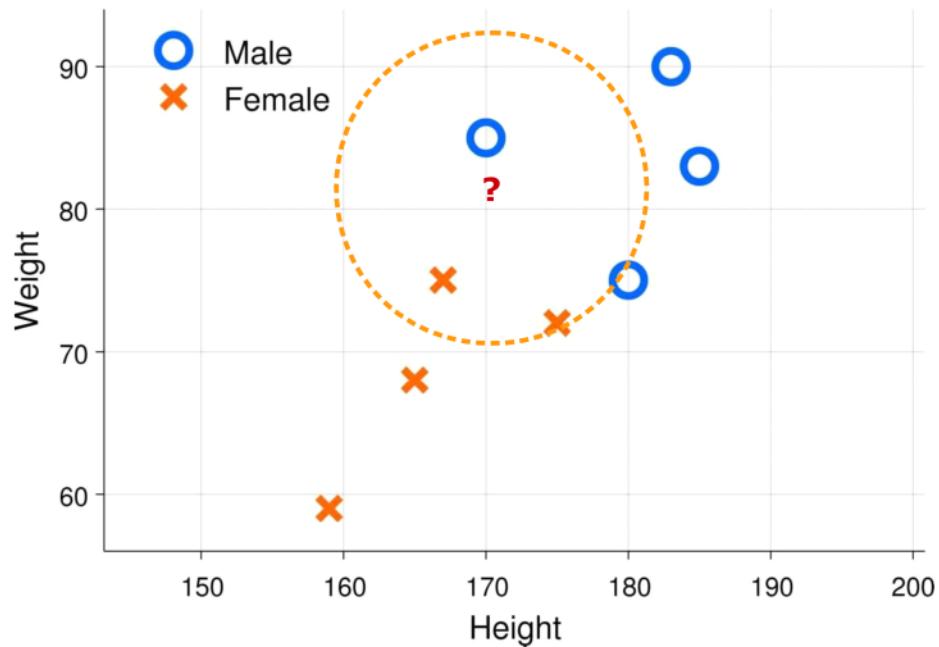
Nearest neighbor classifier

- 2 nearest neighbors



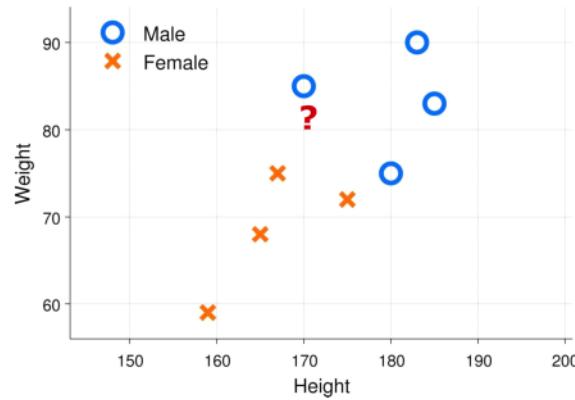
Nearest neighbor classifier

- 3 nearest neighbors

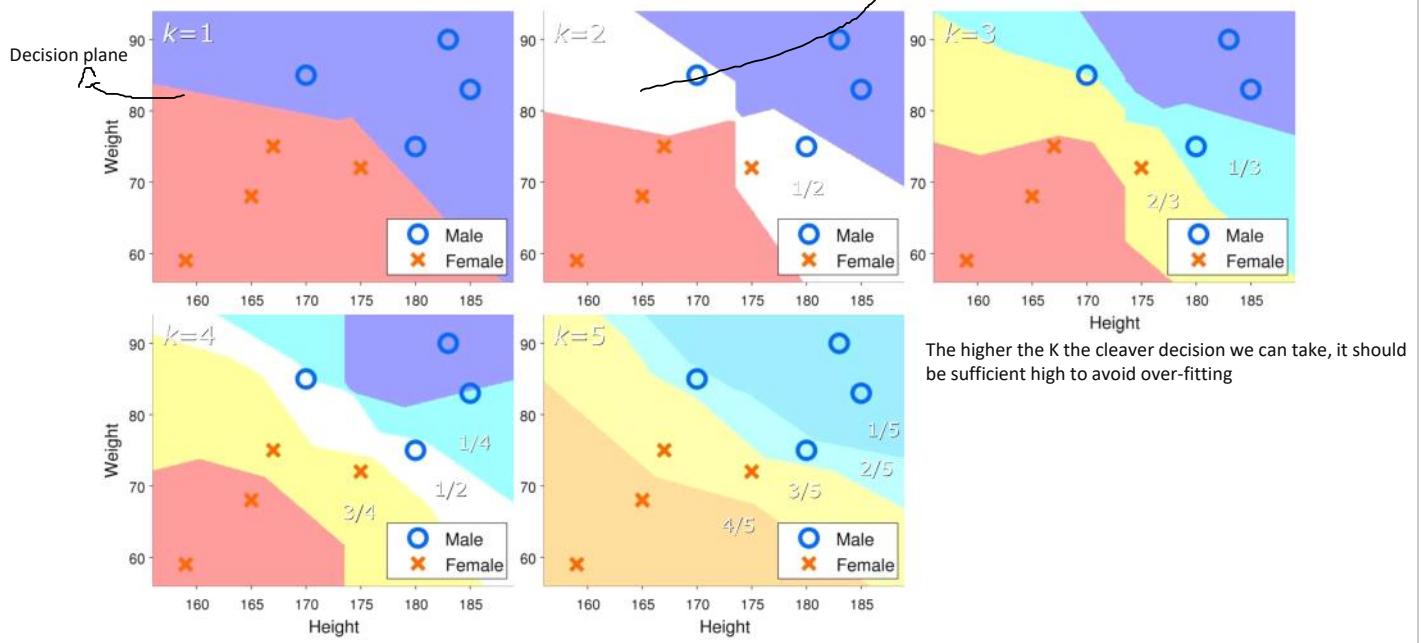


Nearest neighbor classifier

- Choose
 - The number of neighbors, k
 - A distance measure
1. Compute distance to all other data objects
 2. Find the k nearest data objects
 3. Classify according to majority of neighbors



Nearest neighbor decision surface

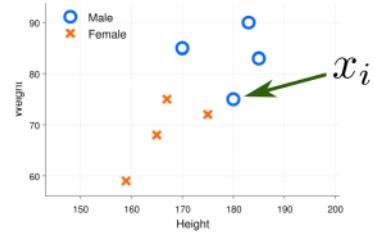


KNN with leave-one-out CV

we can do cross validation for KNN

Leave-one-out CV is convenient with KNN

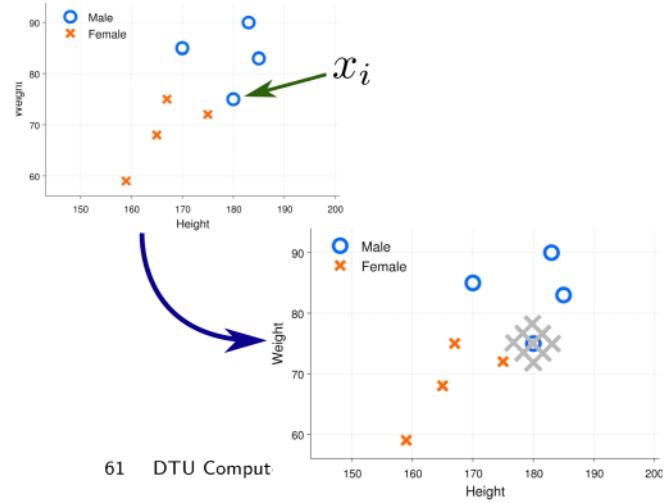
- For each observation x_i



KNN with leave-one-out CV

Leave-one-out CV is convenient with KNN

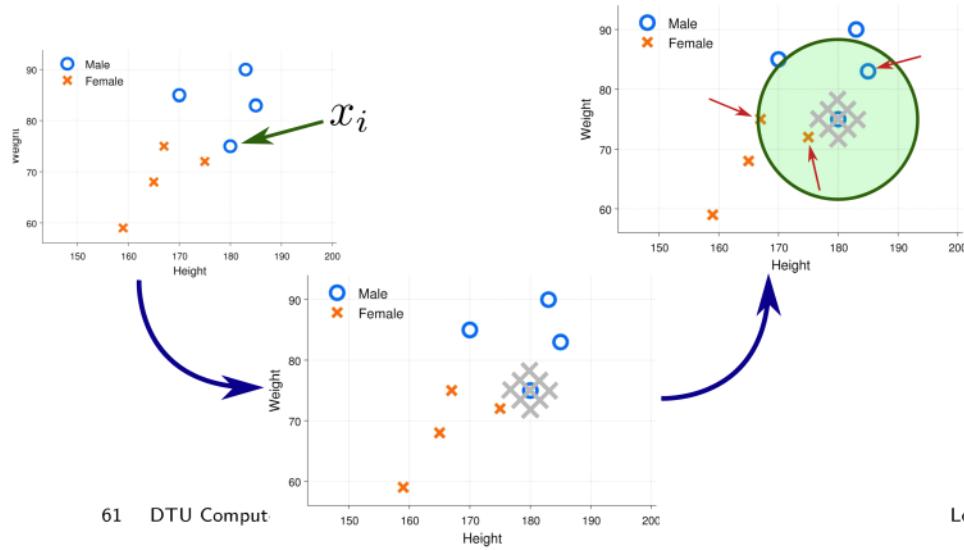
- For each observation x_i
 - Temporarily remove x_i



KNN with leave-one-out CV

Leave-one-out CV is convenient with KNN

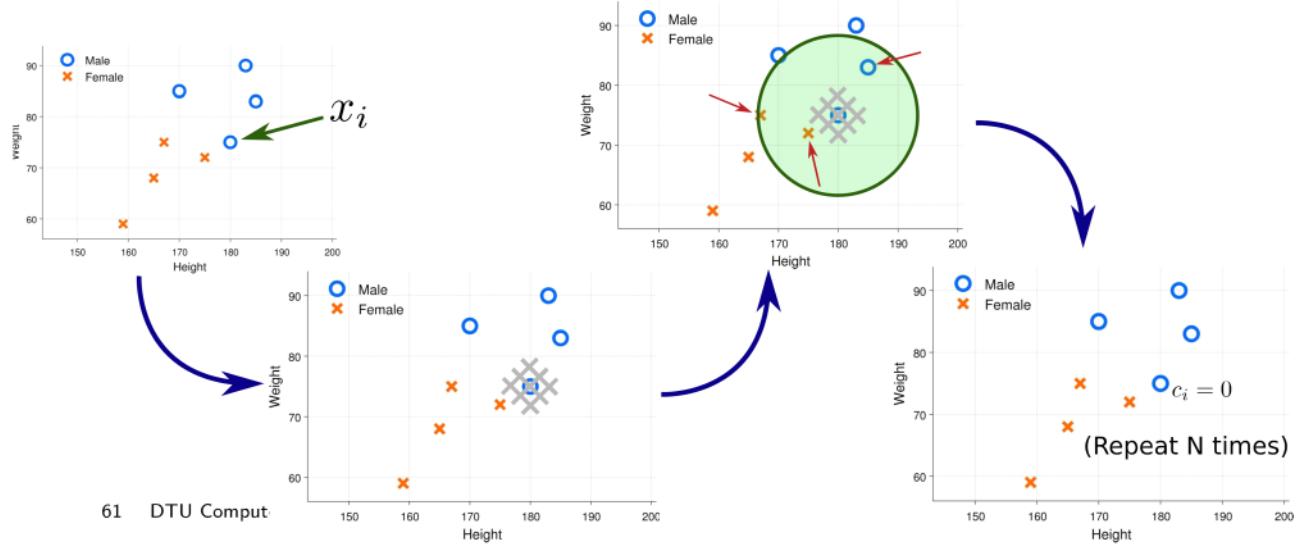
- For each observation x_i
 - Temporarily remove x_i
 - Find K nearest neighbors around x_i (not x_i itself)



KNN with leave-one-out CV

Leave-one-out CV is convenient with KNN

- For each observation x_i
 - Temporarily remove x_i
 - Find K nearest neighbors around x_i (not x_i itself)
 - Determine whether x_i is classified correctly based on this neighborhood, $c_i = 0, 1$

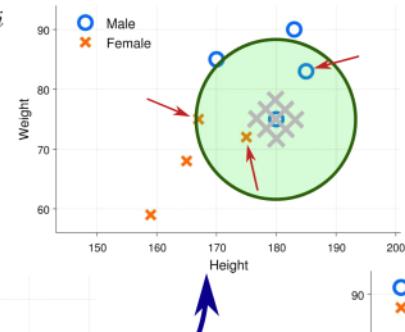
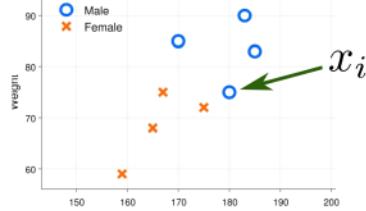


KNN with leave-one-out CV

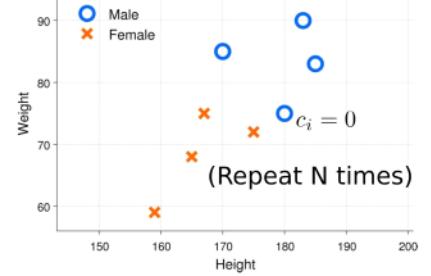
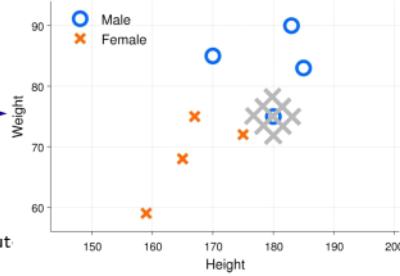
Leave-one-out CV is convenient with KNN

- For each observation x_i
 - Temporarily remove x_i
 - Find K nearest neighbors around x_i (not x_i itself)
 - Determine whether x_i is classified correctly based on this neighborhood, $c_i = 0, 1$

- Compute accuracy as $\frac{1}{N} \sum_{i=1}^N c_i$



61 DTU Comput



Quiz 4, KNN (Spring 2011)

- For each observation x_i
 - Temporarily remove x_i
 - Find K nearest neighbors around x_i (not x_i itself)
 - Determine whether x_i is classified correctly based on this neighborhood

	Diabetic				Normal			
	D1	D2	D3	D4	N1	N2	N3	N4
Diabetic	D1	0	58.5	51.6	18.1	38.0	52.5	71.7
	D2	58.5	0	32.1	72.6	50.5	65.0	13.2
	D3	51.6	32.1	0	60.5	28.4	32.9	45.3
	D4	18.1	72.6	60.5	0	45.9	60.4	79.8
Normal	N1	38.0	50.5	28.4	45.9	0	17.5	63.7
	N2	52.5	65.0	32.9	60.4	17.5	0	78.2
	N3	71.7	12.2	45.3	79.8	63.7	78.2	0
	N4	50.7	63.8	56.3	56.8	50.7	57.2	71.0

The figure shows the distance between the first four diabetic (D1–D4) and normal (N1–N4) women. What are the number of misclassified observations for leave-

one-out cross validation based on 3-nearest neighbor classification when only considering the 8 observations (i.e., D1–D4 and N1–N4) in the figure?

- A. None of the observations will be misclassified.
- B. 2 of the observations will be misclassified.
- C. 6 of the observations will be misclassified.
- D. All of the observations will be misclassified.

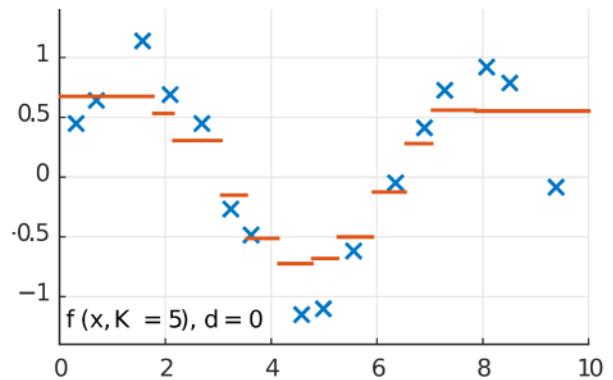
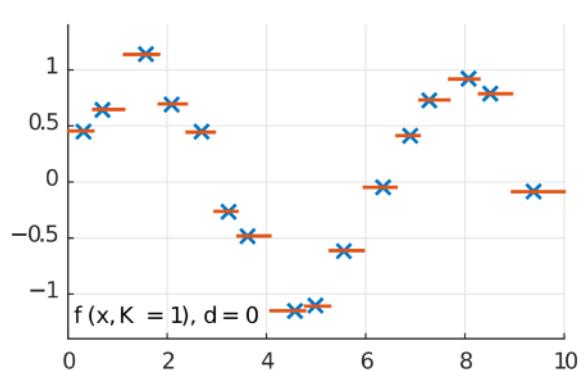
Take D1:
we find the 3 nearest number 

if more NN are about *Normal* then it's misclassified
And viceversa for the blue point

KNN Regression

- Given a training set \mathbf{X}, \mathbf{y}
- For a test observation x predict the average y -value in the neighbourhood

$$\hat{y} = f(\mathbf{x}, K) = \frac{1}{K} \sum_{i \in N_{\mathbf{X}}(\mathbf{x}, K)} y_i$$



Resources

<https://towardsdatascience.com> Alternative introduction to cross-validation

(<https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85>)

02450: Introduction to Machine Learning and Data Mining

Performance evaluation, Bayes, and Naive Bayes

Tommy Sonne Alstrøm

DTU Compute, Technical University of Denmark (DTU)

DTU Compute
Department of Applied Mathematics and Computer Science

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$
$$\Delta \int_a^b \Theta^{\sqrt{17}} + \Omega \int \delta e^{i\pi} = -1$$
$$\varepsilon \in \{2.7182818284\} \text{ is } e^1$$
$$\infty \approx \chi^2 \sum \gg, \approx !$$

Today

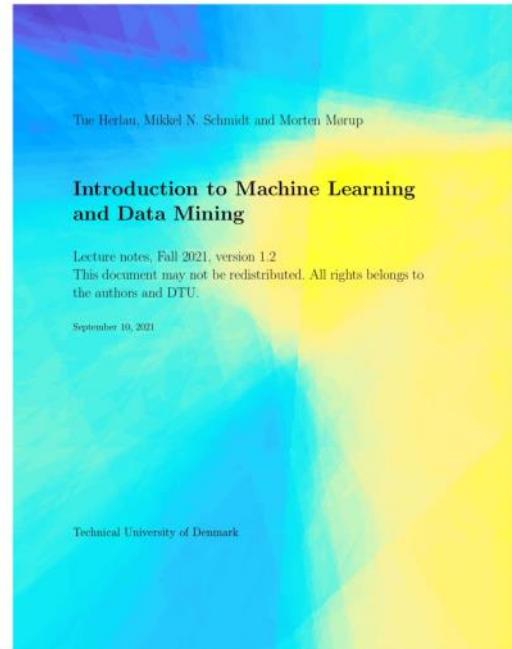
Feedback Groups of the day:

Adrian Valentin Kragh-Hillers, Albert Kaare Klug,
 Alexander Torp Jørgensen, Alland Karim, Anders Bo
 Jensen, Andrea Kristine Gloppen Johnsen, Andreas
 Holm Jørgensen, Andreas Lyndrup Jensen, Angel
 Ricardo Jara Jimenez, Anton Faber Klausen,
 Benjamin Simon Stenbjerg Jepsen, Benjamin
 Starostka Jakobsen, Bernhard Jonsson, Bertrand
 Kouoi, Bjarti Kristiansen, Christian Møller Jensen,
 Christina Naja Tsvilling Jensen, Christopher Jung,
 Clara Regine Hoeg Kold, Daniel Kirchner, Danuta
 Kondarewicz, Dimosthenis Karafylas, Dmitri
 Khlebutin, Eigil Albæk Knudsen, Filip Wolf
 Kristiansen, Frederik Tinus Jeppesen, Fredrik Christer
 Johansson, Kalle Emil Leander Johansen, Karoline
 Helene Baarsøe Jørgensen, Kenneth Sylvest Knudsen,
 Konstantin Krasnopjorov, Maja Brøndtoft Klerk,
 Maks Suppras Kjeldsen, Marcus Hornhawer Thorstén
 Jepsen, Marcus Lari Jørgensen, Marcus Rosenkilde
 Jensen, Marie Juhl Jørgensen, Mathias Bang
 Kristensen, Mathias Lund Jensen, Oliver Andreas
 Munch Jørgensen, Paul Jurisch, Pernille Bachmann
 Jessen, Sana Ullah Khan, Sara Hunfjord Josepsdottir,
 Siddhi Yash Jain, Sika Nuunu Slobodziuk Kristensen,
 Simon Stampe Jensen, Sindri Jonsson, Søren Bønning
 Jakobsen, Thomas Høj Jørgensen, Tobias Ovdal
 Eiberg Jørgensen, Tóra Kristín Jónsdóttir, Ulrika
 Woulhøj Jakobsen, Viktor Guldborg Johnsen,
 Vilhjalmur Kari Jansson, Yuhao Jiang, Zsolt Kovacs

2 DTU Compute

Reading material:

Chapter 11, Chapter 13



Lecture 7 12 October, 2021

Lecture Schedule

1 Introduction

31 August: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

7 September: C2, C3

3 Measures of similarity, summary statistics and probabilities

14 September: C4, C5

4 Probability densities and data visualization

21 September: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

28 September: C8, C9

6 Overfitting, cross-validation and Nearest Neighbor

5 October: C10, C12 (Project 1 due before 13:00)

7 Performance evaluation, Bayes, and Naive Bayes

12 October: C11, C13

8 Artificial Neural Networks and Bias/Variance

26 October: C14, C15

9 AUC and ensemble methods

2 November: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 November: C18

11 Mixture models and density estimation

16 November: C19, C20 (Project 2 due before 13:00)

12 Association mining

23 November: C21

Recap

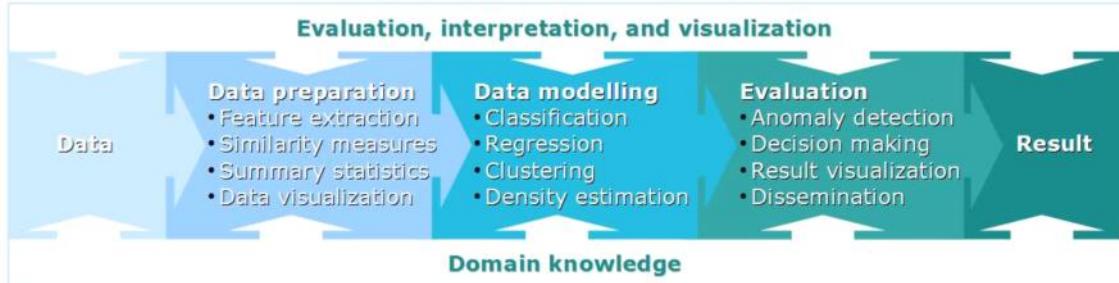
13 Recap and discussion of the exam

30 November: C1-C21

3 DTU Compute

Online help: Forum on DTU Learn
 Videos of lectures: <https://video.dtu.dk>
 Streaming of lectures: Zoom (link on DTU Learn)

Lecture 7 12 October, 2021



Learning Objectives

- Understand the two different evaluation setups
- Apply appropriate statistical tests to evaluate and compare models
- Account for the assumptions made in Naïve Bayes
- Apply Bayes Theorem to obtain the class posterior likelihood

Statistical testing

- A social media company wish to know if a new ad-placement method increases the click-through rate
- How many customers are likely click adds next month?
- How well can a neural network model learn to distinguish between diseased/non-diseased X-rays?
- Should I recommend my neural network model over a competing method?

All involve induction beyond the dataset

Statistical testing

Tests **can** provide:

- An objective way to choose between methods
- A quantification of model performance which takes uncertainty into account

Statistical testing

Tests **can** provide:

- An objective way to choose between methods
- A quantification of model performance which takes uncertainty into account

Tests **do not** provide

- Certain conclusions (Model *A* is better than *B*)
- A black-box recipe

Statistical testing

Tests **can** provide:

- An objective way to choose between methods
- A quantification of model performance which takes uncertainty into account

Tests **do not** provide

- Certain conclusions (Model *A* is better than *B*)
- A black-box recipe

Use statistical tests to aid your interpretation of your results not as an argument in itself

Outline

- What is our overall **objective**? What conclusions do we want?
- What tools do we have available?
- What specific test should I use? (classification, regression, etc.)

The objective and evaluation criteria

Our Guiding principle: create model and predict how well it will perform in the future with new data

- Models are compared based on how well they **generalize to future data**
- Suppose we have data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ and two models $\mathcal{M}_A, \mathcal{M}_B$
- Training on \mathcal{D} , we obtain prediction rules

$$f_{\mathcal{D},A} : \mathbf{x} \rightarrow y, \quad \text{and} \quad f_{\mathcal{D},B} : \mathbf{x} \rightarrow y.$$

The objective and evaluation criteria

- Models are compared based on how well they **generalize to future data**
- Suppose we have data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ and two models $\mathcal{M}_A, \mathcal{M}_B$
- Training on \mathcal{D} , we obtain prediction rules

$$f_{\mathcal{D},A} : \mathbf{x} \rightarrow y, \quad \text{and} \quad f_{\mathcal{D},B} : \mathbf{x} \rightarrow y.$$

- Compared by the **difference in generalization error**:

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}} \quad z_{\mathcal{D}} \text{ will result to be a random variable -> we want to perform some statistics}$$

$$E_{\mathcal{D},A}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},A}(\mathbf{x}), y) d\mathbf{x} dy, \quad E_{\mathcal{D},B}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},B}(\mathbf{x}), y) d\mathbf{x} dy.$$

The objective and evaluation criteria

- Models are compared based on how well they **generalize to future data**
- Suppose we have data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ and two models $\mathcal{M}_A, \mathcal{M}_B$
- Training on \mathcal{D} , we obtain prediction rules

$$f_{\mathcal{D},A} : \mathbf{x} \rightarrow y, \quad \text{and} \quad f_{\mathcal{D},B} : \mathbf{x} \rightarrow y.$$

- Compared by the **difference in generalization error**:

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$

$$E_{\mathcal{D},A}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},A}(\mathbf{x}), y) d\mathbf{x} dy, \quad E_{\mathcal{D},B}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},B}(\mathbf{x}), y) d\mathbf{x} dy.$$

These two models are dependant on the train data $f_{\mathcal{D}}$
 $L(f_{\mathcal{D}})$ = is the loss function on the train data: how much information I lose.

- If $z_{\mathcal{D}} < 0$, it means **that \mathcal{M}_A is better than \mathcal{M}_B ...**

The objective and evaluation criteria

- Models are compared based on how well they **generalize to future data**
- Suppose we have data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ and two models $\mathcal{M}_A, \mathcal{M}_B$
- Training on \mathcal{D} , we obtain prediction rules

$$f_{\mathcal{D},A} : \mathbf{x} \rightarrow y, \quad \text{and} \quad f_{\mathcal{D},B} : \mathbf{x} \rightarrow y.$$

- Compared by the **difference in generalization error**:

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$
$$E_{\mathcal{D},A}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},A}(\mathbf{x}), y) d\mathbf{x} dy, \quad E_{\mathcal{D},B}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},B}(\mathbf{x}), y) d\mathbf{x} dy.$$

- If $z_{\mathcal{D}} < 0$, it means **that \mathcal{M}_A is better than \mathcal{M}_Bwhen trained on \mathcal{D}**

The objective and evaluation criteria

- Models are compared based on how well they **generalize to future data**
- Suppose we have data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ and two models $\mathcal{M}_A, \mathcal{M}_B$
- Training on \mathcal{D} , we obtain prediction rules

$$f_{\mathcal{D},A} : \mathbf{x} \rightarrow y, \quad \text{and} \quad f_{\mathcal{D},B} : \mathbf{x} \rightarrow y.$$

- Compared by the **difference in generalization error**:

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$

$$E_{\mathcal{D},A}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},A}(\mathbf{x}), y) d\mathbf{x} dy, \quad E_{\mathcal{D},B}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},B}(\mathbf{x}), y) d\mathbf{x} dy.$$

- If $z_{\mathcal{D}} < 0$, it means **that \mathcal{M}_A is better than \mathcal{M}_Bwhen trained on \mathcal{D}**

Setup I Statistical tests of performance considering the **specific** training set \mathcal{D}

An alternative objective

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$

- If you prove $z_{\mathcal{D}} < 0$, you can't know if this is true for an independent dataset \mathcal{D}'

An alternative objective

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$

- If you prove $z_{\mathcal{D}} < 0$, you can't know if this is true for an independent dataset \mathcal{D}'
- Therefore, the conclusion is not independently reproducible

An alternative objective

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$

- If you prove $z_{\mathcal{D}} < 0$, you can't know if this is true for an independent dataset \mathcal{D}'
- Therefore, the conclusion is not independently reproducible
- To overcome this, test if \mathcal{M}_A is better than \mathcal{M}_B when averaging over dataset

$$\begin{aligned} z &= \mathbb{E}_{\mathcal{D}}[z_{\mathcal{D}}] < 0 \\ E^{\text{gen}} &= \int \left[\int L(f_{\mathcal{D}}(\mathbf{x}), y) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \right] p(\mathcal{D}) d\mathcal{D} \end{aligned}$$

An alternative objective

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$

- If you prove $z_{\mathcal{D}} < 0$, you can't know if this is true for an independent dataset \mathcal{D}'
- Therefore, the conclusion is not independently reproducible
- To overcome this, test if \mathcal{M}_A is better than \mathcal{M}_B when averaging over dataset

$$z = \mathbb{E}_{\mathcal{D}}[z_{\mathcal{D}}] < 0$$
$$E^{\text{gen}} = \int \left[\int L(f_{\mathcal{D}}(\mathbf{x}), y) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \right] p(\mathcal{D}) d\mathcal{D}$$

- If $z < 0$, it means \mathcal{M}_A is better than \mathcal{M}_B ... using a typical training set

Setup II Statistical tests of performance considering a dataset of size N

Choices, choices

Setup I Statistical tests of performance considering the **specific** training set \mathcal{D} ?

Setup II *Statistical tests of performance considering a **dataset** of size N*

Which to choose fundamentally depends on what you want to conclude

Choices, choices

With setup I, we can only say that our model is better in our dataset, but not that the model is better in general. This is setup II.

Setup I Statistical tests of performance considering the **specific** training set \mathcal{D} ?

This two are different:

Setup II *Statistical tests of performance considering a **dataset** of size N*

Which to choose fundamentally depends on what you want to conclude

- Setup II is a more general (impressive) conclusion
- Setup II is probably what we want in science
- Setup II requires (a lot of) cross-validation
- If you have a single train/test split, use setup I

We will consider **setup I** here

Statistical tasks and tools



Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

Statistical tasks and tools

Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

Hypothesis testing Determine **whether there is an effect** by choosing between $H_0 : z = 0$ vs. $H_1 : z \neq 0$

Statistical tasks and tools

Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

Hypothesis testing Determine whether there is an effect by choosing between $H_0 : z = 0$ vs. $H_1 : z \neq 0$

Estimation Determine a likely value $z \approx \hat{z}$ and an interval $[z_L, z_U]$ that likely contains z

Statistical tasks and tools

Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

Hypothesis testing Determine whether there is an effect by choosing between $H_0 : z = 0$ vs. $H_1 : z \neq 0$

Estimation Determine a likely value $z \approx \hat{z}$ and an interval $[z_L, z_U]$ that likely contains z

- Evidence against H_0 is measured by a **p-value** (low p is evidence for an effect $z \neq 0$)

Statistical tasks and tools

Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

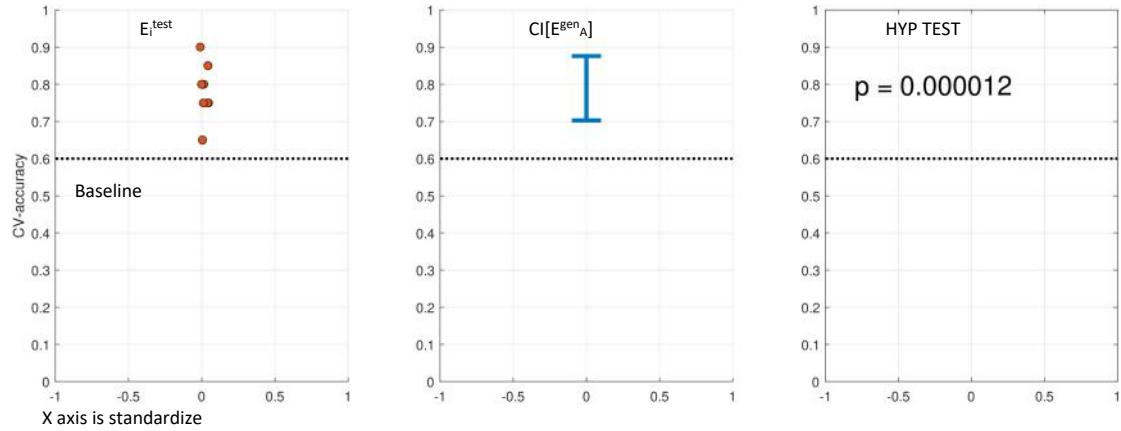
Hypothesis testing Determine whether there is an effect by choosing between $H_0 : z = 0$ vs. $H_1 : z \neq 0$

Estimation Determine a likely value $z \approx \hat{z}$ and an interval $[z_L, z_U]$ that likely contains z

- Evidence against H_0 is measured by a **p-value** (low p is evidence for an effect $z \neq 0$)
- Estimation of $[z_L, z_U]$ done using an **α -confidence interval** (lower α means a more conservative, wider, interval)

Choosing the right tool

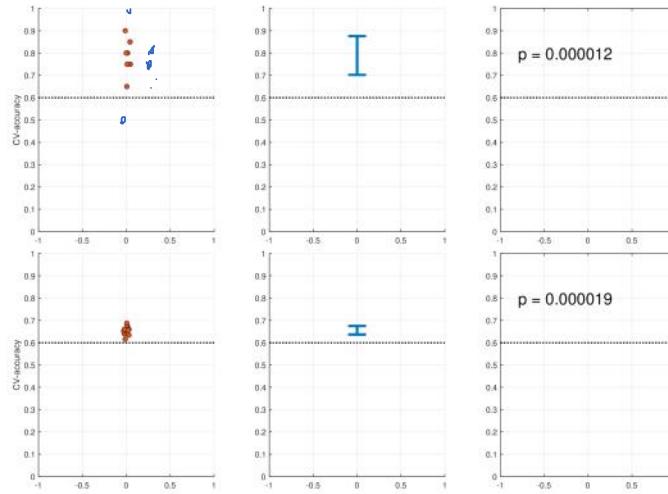
- Consider binary classification using $N = 200$ samples
- We estimate test error using $K = 10$ -fold CV (10 test-error estimates)
- Question: Is accuracy $E_A^{\text{gen}} \approx \frac{1}{K} \sum_{k=1}^K E_i^{\text{test}}$ greater than baseline θ_0 ?
- (Baseline classify everything as maximum class, accuracy 60%)



Which tool to use

- Top: $N = 200$ sample example
- Bottom: Harder problem using $N = 2000$ samples

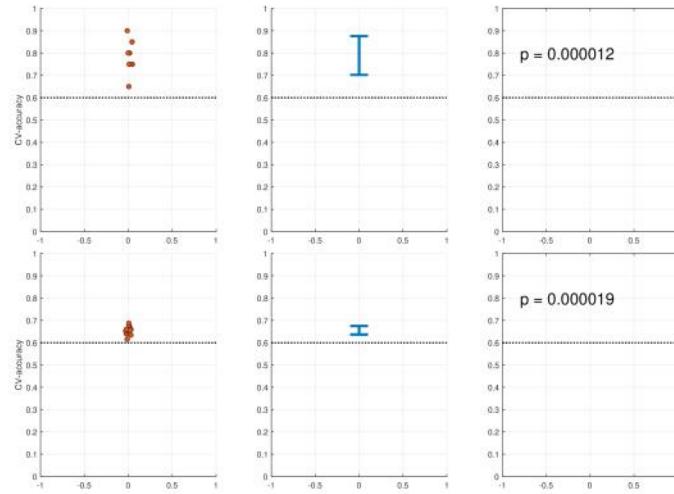
From the first plot we can say that the area of interest point we want to investigate, like the ones in the limit



The confidence interval says if our models work better than the random model. These two are a function of how many samples we collect

Which tool to use

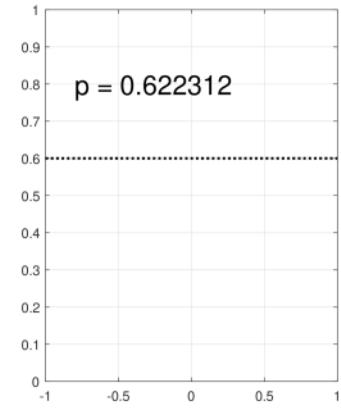
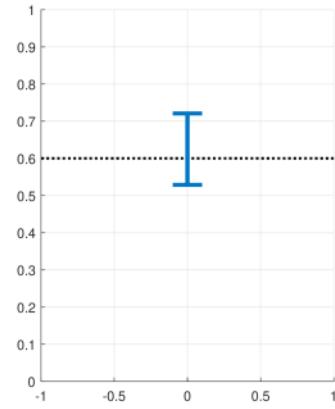
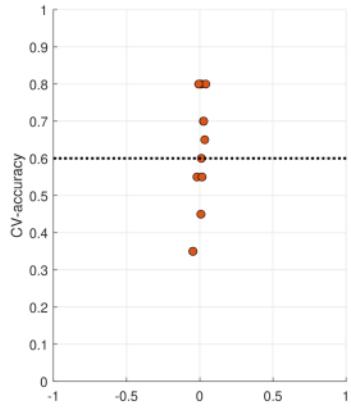
- Top: $N = 200$ sample example
- Bottom: Harder problem using $N = 2000$ samples



- p -value primarily measure of sample size (not **effect size!**)
- Which do **you** think are more informative?

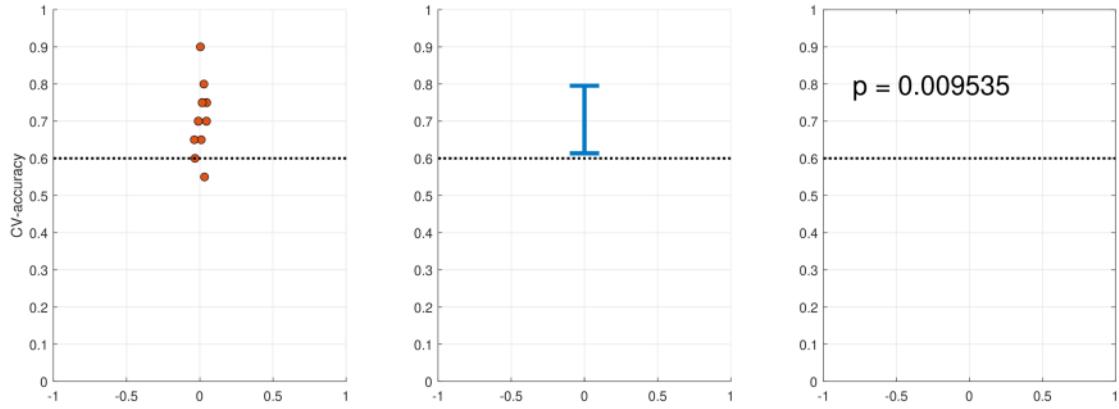
Variability

- New problem using $N = 200$ samples. Is there an effect?

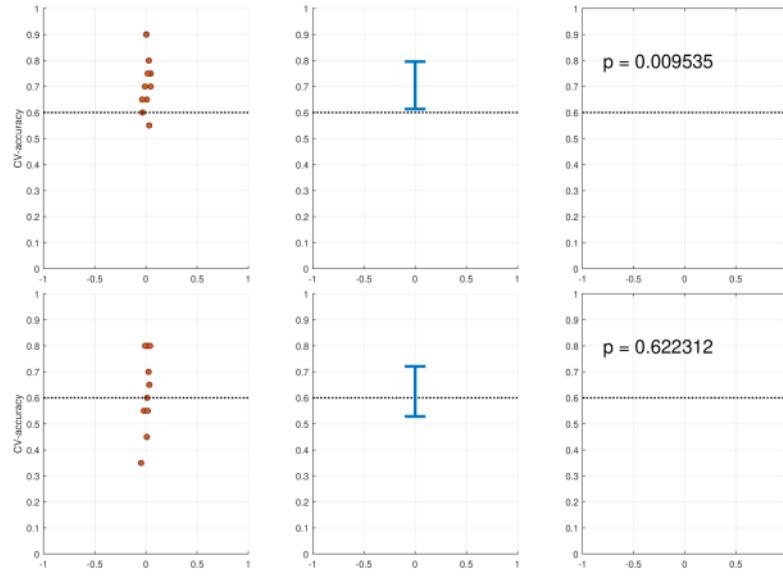


Variability

- Another problem using $N = 200$ samples. Is there an effect?

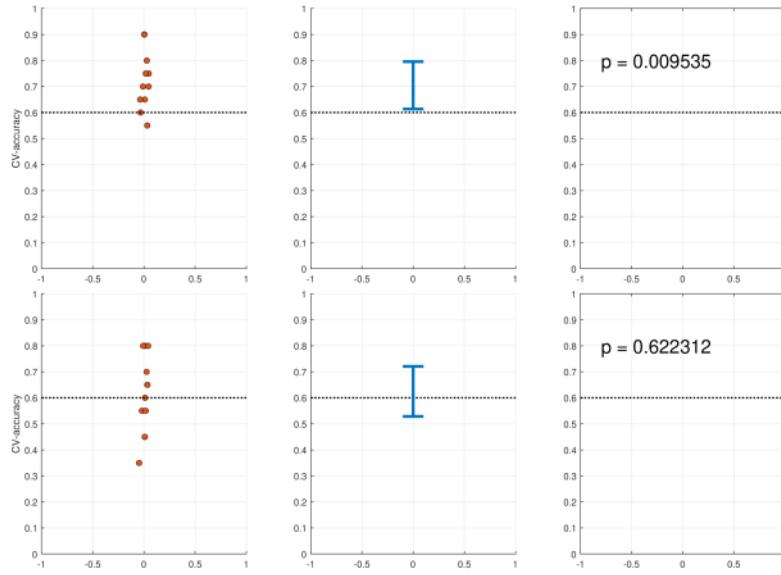


The nasty bit



The same underline process has been run multiple times
Sometimes we can be lucky or not -> statistics.

The nasty bit



- Only difference is random variability in dataset
- Low p -value does **not necessarily** mean reproducible
 - Training many models will lead to false positives
 - Statistics **will not** fix unclear results; probably just lead to false positives

Connecting objective to numbers

- We want to draw conclusions about the difference in performance:

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$

$$E_{\mathcal{D},A}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},A}(\mathbf{x}), y) d\mathbf{x} dy, \quad E_{\mathcal{D},B}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},B}(\mathbf{x}), y) d\mathbf{x} dy.$$

- This can be estimated as

$$\hat{z}_{\mathcal{D}} = \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} [L(f_{\mathcal{D},A}(\mathbf{x}_i), y_i) - L(f_{\mathcal{D},B}(\mathbf{x}_i), y_i)]$$

The i refers to the index of
the point of both models

$$= \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} z_i, \quad \text{where: } z_i = L(f_{\mathcal{D},A}(\mathbf{x}_i), y_i) - L(f_{\mathcal{D},B}(\mathbf{x}_i), y_i).$$

Looks like the mean value of the losses on all the points

Abstracting to a statistical question

Consider data as the n numbers

$$D = (z_1, \dots, z_n). \quad (1)$$

General form of the problem: Draw conclusions about

$$\theta = E_{A,D}^{\text{gen}} - E_{B,D}^{\text{gen}}$$

Based on the estimate:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n z_i. \quad \begin{matrix} \text{We want to do our statistical consideration on} \\ \text{Theta_hat (estimated variable)} \end{matrix} \quad (2)$$

Statistical tools: Parameter

- Assume z_i is a realization of a random variable Z_i

Because we have done the estimation of z_d (from integral to The estimation (it is impossible to calculate the integral)

- It has density

$$p(Z_i = z_i | \theta) = p_\theta(z_i) \quad (\text{definition})$$

- Density of all dataset

$$p_\theta(D) = \prod_{i=1}^n p_\theta(z_i). \quad (3)$$

Given Hypothesis: All the z_i are independent and identically distributed

Statistical tools: Parameter

- Assume z_i is a realization of a random variable Z_i

- It has density

$$p(Z_i = z_i | \theta) = p_\theta(z_i)$$

- Density of all dataset

$$p_\theta(D) = \prod_{i=1}^n p_\theta(z_i). \quad (3)$$

- Returning to our goals:

- **estimating plausible ranges of θ**
- **hypothesis testing such as whether θ takes a particular value**

- Let's look at the statistical tools to accomplish this

Statistical tools: Statistic and estimator

Statistic A statistic is a function of the data D and will be denoted t .
For instance, the mean and variance are both statistics:

$$t_0(D) = \frac{1}{n} \sum_{i=1}^n Z_i, \text{ or } t_1(D) = \frac{1}{n} \sum_{i=1}^n (Z_i - t_0(D))^2.$$

Statistical tools: Statistic and estimator

Statistic A statistic is a function of the data D and will be denoted t .

For instance, the mean and variance are both statistics:

$$t_0(D) = \frac{1}{n} \sum_{i=1}^n Z_i, \text{ or } t_1(D) = \frac{1}{n} \sum_{i=1}^n (Z_i - t_0(D))^2.$$

Estimator An estimator is a statistic t of D such that $t(D)$ is close to θ .

In the examples we will consider the mean

$$t_0(D) = \frac{1}{n} \sum_{i=1}^n Z_i$$

Statistical tools: Confidence interval

- A **confidence interval** (CI) is an interval $[\theta_L, \theta_U]$ which likely contains θ

Statistical tools: Confidence interval

- A **confidence interval** (CI) is an interval $[\theta_L, \theta_U]$ which likely contains θ
- The CI is a function of the data D . θ_L and θ_U are two statistics and for a concrete dataset the interval is computed to be

$$[\theta_L(D), \theta_U(D)]. \quad (4)$$

Statistical tools: Confidence interval

The CI means that if we take infinite measurements, then the 95% of times our value(that could be the mean) is inside the CI

- A **confidence interval** (CI) is an interval $[\theta_L, \theta_U]$ which likely contains θ
- The CI is a function of the data D . θ_L and θ_U are two statistics and for a concrete dataset the interval is computed to be

$$[\theta_L(D), \theta_U(D)]. \quad (4)$$

- With probability $1 - \alpha$, the true value θ should fall within the confidence interval $[\theta_L(D), \theta_U(D)]$ as we randomize over different datasets

$$P_\theta(\theta \in [\theta_L, \theta_U]) = 1 - \alpha. \quad (5)$$

Statistical tools: Null hypothesis testing and p -value

- Determining whether a **null hypothesis** H_0 about the parameters is true or false

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0$$

- Intuitively, if H_0 is true, the data should behave in a certain way

- **We test if the data is implausible assuming H_0**

Statistical tools: Null hypothesis testing and p -value

- Determining whether a **null hypothesis** H_0 about the parameters is true or false

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0$$

- Intuitively, if H_0 is true, the data should behave in a certain way

- **We test if the data is implausible assuming H_0**

- Specifically, let t be a statistic, for our purpose

$$t(D) = \frac{1}{n} \sum_{i=1}^n Z_i$$

On our dataset it has a particular value $t_0 = \frac{1}{n} \sum_{i=1}^n z_i$

- We can compute the density $t(D)$ takes a particular value given H_0 is true:

$$p(t(D) = t | H_0) = p_{\theta=\theta_0}(t(D) = t)$$

Statistical tools: Null hypothesis testing and p -value

- Determining whether a **null hypothesis** H_0 about the parameters is true or false

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0$$

- Intuitively, if H_0 is true, the data should behave in a certain way

- We test if the data is implausible assuming H_0

- Specifically, let t be a statistic, for our purpose

$$t(D) = \frac{1}{n} \sum_{i=1}^n Z_i$$

On our dataset it has a particular value $t_0 = \frac{1}{n} \sum_{i=1}^n z_i$

- We can compute the density $t(D)$ takes a particular value given H_0 is true:

$$p(t(D) = t_0 | H_0) = p_{\theta=\theta_0}(t(D) = t_0)$$

- p -value is the chance $t(D)$ is at least as extreme as what we actually observed:

$$p\text{-value} : \quad p = P(t(D) > |t_0| | H_0) = P_{\theta=\theta_0}(t(D) \geq |t_0|). \quad (6)$$

Setup I: Fixed training set

Suppose we carry out cross-validation to obtain:

$$(\mathcal{D}_1^{\text{train}}, \mathcal{D}_1^{\text{test}}), \dots, (\mathcal{D}_K^{\text{train}}, \mathcal{D}_K^{\text{test}}). \quad (7)$$

We collect these into (paired) vectors of predictions and true values:

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \\ \vdots \\ \hat{\mathbf{y}}_K \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1^{\text{test}} \\ \mathbf{y}_2^{\text{test}} \\ \vdots \\ \mathbf{y}_K^{\text{test}} \end{bmatrix}. \quad (8)$$

Evaluation of a single classifier

- Define:

$$c_i = \begin{cases} 1 & \text{if } \hat{y}_i = y_i \\ 0 & \text{if otherwise.} \end{cases} \quad \begin{matrix} \text{Correct} \\ \text{wrong} \end{matrix}$$

- Number of accurate guesses:

$$m = \sum_{i=1}^n c_i.$$

Evaluation of a single classifier

- Define:

$$c_i = \begin{cases} 1 & \text{if } \hat{y}_i = y_i \\ 0 & \text{if otherwise.} \end{cases}$$

- Number of accurate guesses:

$$m = \sum_{i=1}^n c_i.$$

- Let the chance the classifier is correct be θ . Then, from [Lecture 4](#), we know

$$p(\theta|m, n) = \text{Beta}(\theta|a, b), \quad a = m + \frac{1}{2}, \quad \text{and } b = n - m + \frac{1}{2}. \quad (9)$$

Theta is the accuracy

Evaluating a single classifier (Jeffreys interval)

The probability that our model accuracy is a random number -> beta distribution

- If m is the number of accurate guesses, then

$$p(\theta|m, n) = \text{Beta}(\theta|a, b), \quad a = m + \frac{1}{2}, \text{ and } b = n - m + \frac{1}{2}.$$

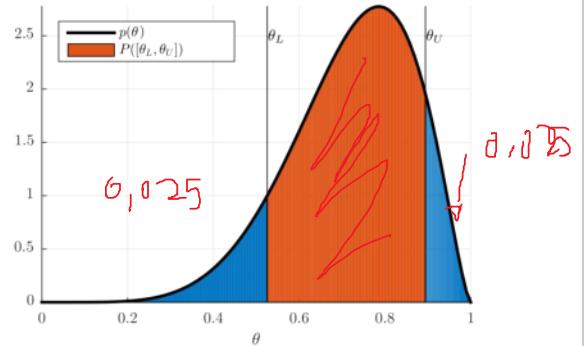
- The $1 - \alpha$ confidence interval is given as $[\theta_L, \theta_U]$: If we choose a smaller alpha then the confidence interval become wider -> choose a CI that is meaningful for the application

$$\theta_L = \text{cdf}_B^{-1} \left(\frac{\alpha}{2} | a, b \right) \text{ if } m > 0 \text{ otherwise } \theta_L = 0$$

$$\theta_U = \text{cdf}_B^{-1} \left(1 - \frac{\alpha}{2} | a, b \right) \text{ if } m < n \text{ otherwise } \theta_U = 1$$

$$\hat{\theta} = \mathbb{E}[\theta] = \frac{a}{a+b}$$

$\alpha = 0.05$



Suppose we have 14 observations, 8 of which are correct, with alpha = 0.05 -> find the CI.
 $\Theta = 8/14, N = 14, m = 8$

Comparing two classifiers

- Assume we have predictions from both classifiers:

$$\hat{\mathbf{y}}^A = \hat{y}_1^A, \dots, \hat{y}_n^A, \quad \hat{\mathbf{y}}^B = \hat{y}_1^B, \dots, \hat{y}_n^B.$$

- As before, we want to know if the classifiers are correct or not:

$$c_i^A = \begin{cases} 1 & \text{if } \hat{y}_i^A = y_i \\ 0 & \text{if otherwise.} \end{cases} \quad \text{and} \quad c_i^B = \begin{cases} 1 & \text{if } \hat{y}_i^B = y_i \\ 0 & \text{if otherwise.} \end{cases}$$

Comparing two classifiers

- Assume we have predictions from both classifiers:

$$\hat{\mathbf{y}}^A = \hat{y}_1^A, \dots, \hat{y}_n^A, \quad \hat{\mathbf{y}}^B = \hat{y}_1^B, \dots, \hat{y}_n^B.$$

- As before, we want to know if the classifiers are correct or not:

$$c_i^A = \begin{cases} 1 & \text{if } \hat{y}_i^A = y_i \\ 0 & \text{if otherwise.} \end{cases} \quad \text{and} \quad c_i^B = \begin{cases} 1 & \text{if } \hat{y}_i^B = y_i \\ 0 & \text{if otherwise.} \end{cases}$$

- The relevant information is the contingency table:

Agreement: the 2 classifier agree	$n_{11} = \sum_{i=1}^n c_i^A c_i^B$	= {Both classifiers are correct}
disagreement	$n_{12} = \sum_{k=1}^n c_k^A (1 - c_k^B)$	= {A is correct, B is wrong}
agreement	$n_{21} = \sum_{k=1}^n (1 - c_k^A) c_k^B$	= {A is wrong, B is correct}
	$n_{22} = \sum_{k=1}^n (1 - c_k^A)(1 - c_k^B)$	= {Both classifiers are wrong}

Only the disagreement
are needed in order to
stabilize which
classifier is better

Comparing two classifiers: McNemar's test

- We want to compare the accuracy difference: $\theta = \theta_A - \theta_B$
- It is possible to show (approximately) Is theta_0 = 0 (H_0) ?

To calculate the confidence
Interval for comparing 2 classifier

C1

$$p(\theta|\mathbf{n}) = \frac{1}{2} \text{Beta}\left(\frac{\theta+1}{2} \mid a=f, b=g\right),$$

$$f = \frac{E_\theta + 1}{2} (Q - 1) \quad g = \frac{1 - E_\theta}{2} (Q - 1)$$

$$E_\theta = \frac{n_{12} - n_{21}}{n}, \quad Q = \frac{n^2(n+1)(E_\theta + 1)(1 - E_\theta)}{n(n_{12} + n_{21}) - (n_{12} - n_{21})^2}.$$

$$\theta_L = 2\text{cdf}_B^{-1}\left(\frac{\alpha}{2} \mid a=f, b=g\right) - 1, \quad \theta_U = 2\text{cdf}_B^{-1}\left(1 - \frac{\alpha}{2} \mid a=f, b=g\right) - 1 \quad (10)$$

- For a p -value, note that A is better than B if $n_{12} > n_{21}$
- A p -value can be obtained as:

We can calculate the probability of p -value using the following eq:

$$p = 2\text{cdf}_{\text{binom}}\left(m = \min\{n_{12}, n_{21}\} \mid \theta = \frac{1}{2}, N = n_{12} + n_{21}\right)$$

We have 2 classifiers: c1 is correct 28 times where c2 was wrong, and c2 was correct 35 times where c1 was wrong \rightarrow calculate the p -value. ($n_{12}=28, n_{21}=35=35 \rightarrow p??$)

Confidence interval for a regression model

- Use cross-validation to obtain predictions \hat{y}_i and true values y_i . Select loss

$$z_i = |\hat{y}_i - y_i| \quad \text{or} \quad z_i = (\hat{y}_i - y_i)^2 \quad (11)$$

- Estimated error is: $\hat{z} = \frac{1}{n} \sum_{i=1}^n z_i$.
- Assume each error is normally distributed (**warning!**)

$$p(D|u, \sigma^2) = \prod_{i=1}^n \mathcal{N}(z_i|u, \sigma^2)$$

- It is possible to show u follows a generalized Student's t -distribution:

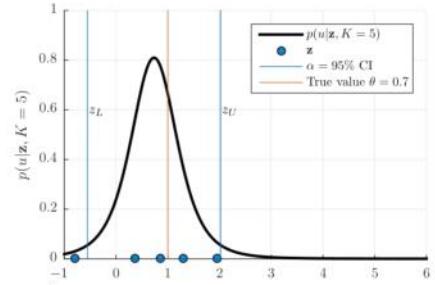
$$p(u|D) = p_T(u|\nu = n-1, \mu = \hat{z}, \sigma = \tilde{\sigma})$$

with parameters $\hat{z} = \frac{1}{n} \sum_{i=1}^n z_i$ and $\tilde{\sigma} = \sqrt{\sum_{i=1}^n \frac{(z_i - \hat{z})^2}{n(n-1)}}$.

- The Student's t -distribution has density

$$\text{Student } t\text{-distribution} \quad p_T(x|\nu, \mu, \sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2}) \sqrt{\pi\nu\sigma^2}} \left(1 + \frac{1}{\nu} \left[\frac{x-\mu}{\sigma}\right]^2\right)^{-\frac{\nu+1}{2}}.$$

Confidence interval for a regression model



- Step back: Assuming $z_i = L(y_i, \hat{y}_i)$ and

$$z_i \sim \mathcal{N}(z_i | \mu = u, \sigma^2)$$

- In this case u is the average error rate. Since we have shown:

$$p(u|D) = p_T(u|\nu = n-1, \mu = \hat{z}, \sigma = \tilde{\sigma})$$

- An approximate $1 - \alpha$ confidence interval is:

$$z_L = \text{cdf}_T^{-1} \left(\frac{\alpha}{2} \mid \nu, \hat{z}, \tilde{\sigma} \right), \quad z_U = \text{cdf}_T^{-1} \left(1 - \frac{\alpha}{2} \mid \nu, \hat{z}, \tilde{\sigma} \right). \quad (12)$$

Comparing two regression models

- Use cross-validation to obtain (paired) predictions along with true values y_i

$$\hat{y}_1^A, \dots, \hat{y}_n^A, \quad \text{and} \quad \hat{y}_1^B, \dots, \hat{y}_n^B. \quad (13)$$

- Select a loss-function to compute the per-observation losses as in

$$z_1^A, \dots, z_n^A, \quad \text{and} \quad z_1^B, \dots, z_n^B.$$

- Note that

$$\begin{aligned} z &= E_{A,D}^{\text{gen}} - E_{B,D}^{\text{gen}} \approx \hat{z} = \left(\frac{1}{n} \sum_{i=1}^n z_i^A \right) - \left(\frac{1}{n} \sum_{i=1}^n z_i^B \right) \\ &= \frac{1}{n} \sum_{i=1}^n z_i, \quad \text{where } z_i = z_i^A - z_i^B \end{aligned}$$

- Assume $z_i \sim \mathcal{N}(z_i | \mu = u, \sigma^2)$
- Compute a $1 - \alpha$ CI using methods on previous slide

Comparing two regression models: *p*-values

$$z = E_A^{\text{gen}} - E_B^{\text{gen}} \approx \hat{z} = \frac{1}{n} \sum_{i=1}^n z_i, \quad \text{where } z_i = z_i^A - z_i^B$$

- Assuming

$$z_i \sim \mathcal{N}(z_i | \mu = u, \sigma^2)$$

where u is the true difference in error function we have shown:

$$p(u|D) = p_T(u|\nu = n - 1, \mu = \hat{z}, \sigma = \tilde{\sigma})$$

- Therefore, we can test the hypothesis

$$H_0 : \text{Model } \mathcal{M}_A \text{ and } \mathcal{M}_B \text{ have the same performance, } u = 0 \quad (14)$$

$$H_1 : \text{Model } \mathcal{M}_A \text{ and } \mathcal{M}_B \text{ have different performance, } u \neq 0. \quad (15)$$

- A *p*-value can be computed as

$$p = 2\text{cdf}_T(-|\hat{z}| \mid \nu = n - 1, \mu = 0, \sigma = \tilde{\sigma}). \quad (16)$$

Which type of cross-validation?

- When using **setup 1** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed

Which type of cross-validation?

- When using **setup 1** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed
- Results will be significant with enough data

Which type of cross-validation?

- When using **setup 1** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed
- Results will be significant with enough data
 - Focus on estimated an effect size

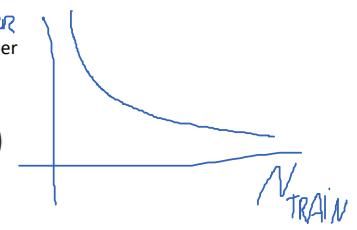
Which type of cross-validation?

- When using **setup 1** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed
- Results will be significant with enough data
 - Focus on estimated an effect size
 - Multiple-comparison problem

Which type of cross-validation?

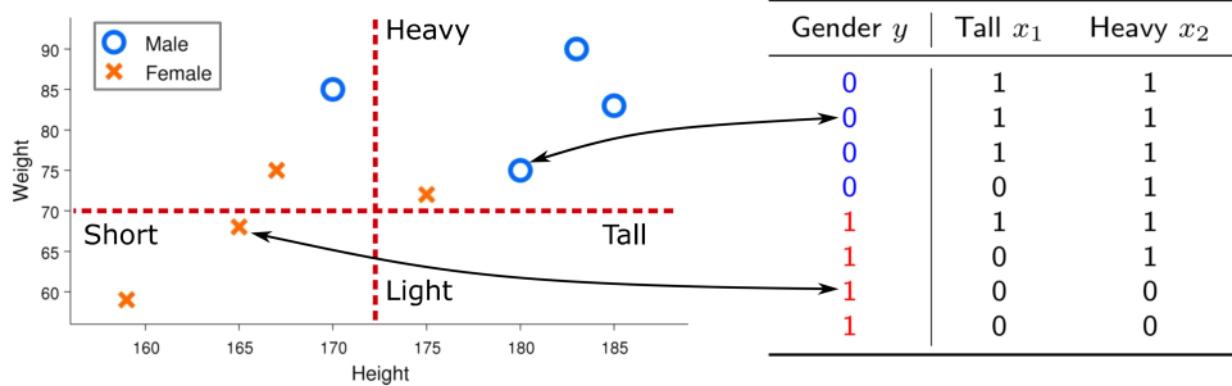
We only need to estimate the accuracy of the classifier.

The error decrease as the number of training decrease
-> better choose K large



- When using **setup I** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed
- Results will be significant with enough data
 - Focus on estimated an effect size
 - Multiple-comparison problem
 - **Transparency, availability of datasets/code, breadth of testing, self-criticism** guarantees reproducibility, not a sophisticated test
- In **setup II**, correlation of training data is taken into account and K -fold is optimal
 - Your **setup I** results do not generalize beyond your training data

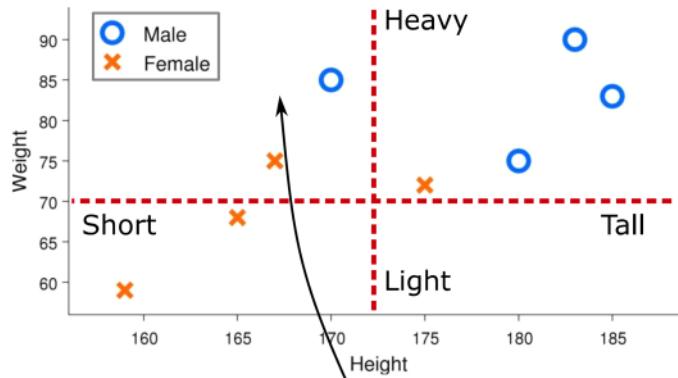
Bayes and Naive-Bayes



CLASS \downarrow
 $p(y|x_1, x_2) = \frac{p(x_1, x_2|y)p(y)}{\sum_{k=0}^1 p(x_1, x_2|y=k)p(y=k)}$
FEATURES \downarrow

Use Base rule so that we cn reverse the problem

Example 1: Normal Bayes



Gender y	Tall x_1	Heavy x_2
0	1	1
0	1	1
0	1	1
0	0	1
1	1	1
1	0	1
1	0	0
1	0	0

Probability a short, heavy person is male:

$$P(y = 0|x_1 = 0, x_2 = 1) = \frac{p(x_1 = 0, x_2 = 1|y = 0)p(y = 0)}{\sum_{k=0}^1 p(x_1 = 0, x_2 = 1|y = k)p(y = k)}$$

I need to observe the joint distribution of x and x_2 . There could be a problem in base: if multiplying all the probabilities together, so if I get 0/0 that leads to an inconclusion.

A practical problem with Bayesian classifier

- In general:

$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1, x_2, \dots, x_M|y)p(y)}{\sum_{k=0}^{K-1} p(x_1, x_2, \dots, x_M|y=k)p(y=k)}$$
$$p(x_1, \dots, x_M|y=k) = \frac{\text{Nr. obs where } y=k \text{ and we measure } x_1, \dots, x_M}{\text{Observations where } y=k}$$

A practical problem with Bayesian classifier

- In general:

$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1, x_2, \dots, x_M|y)p(y)}{\sum_{k=0}^{K-1} p(x_1, x_2, \dots, x_M|y=k)p(y=k)}$$
$$p(x_1, \dots, x_M|y=k) = \frac{\text{Nr. obs where } y=k \text{ and we measure } x_1, \dots, x_M}{\text{Observations where } y=k}$$

- Naive Bayes assumption

$$p(x_1, x_2, \dots, x_M|y) = p(x_1|y)p(x_2|y) \times \dots \times p(x_M|y)$$

A practical problem with Bayesian classifier

- In general:

$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1, x_2, \dots, x_M|y)p(y)}{\sum_{k=0}^{K-1} p(x_1, x_2, \dots, x_M|y=k)p(y=k)}$$

$$p(x_1, \dots, x_M|y=k) = \frac{\text{Nr. obs where } y=k \text{ and we measure } x_1, \dots, x_M}{\text{Observations where } y=k}$$

- Naive Bayes assumption

It is hard to get this observation: what I do here is to look at all of the features singularly, where it becomes difficult that are zero

$$p(x_1, x_2, \dots, x_M|y) = p(x_1|y)p(x_2|y) \times \dots \times p(x_M|y)$$

- Naive Bayes classifier

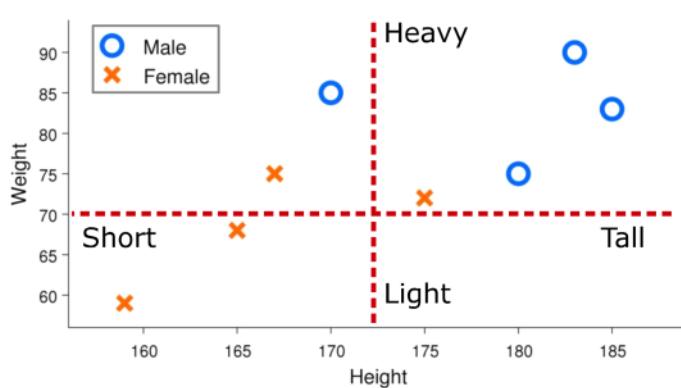
$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1, x_2, \dots, x_M|y)p(y)}{\sum_{k=0}^1 p(x_1, x_2, \dots, x_M|y=k)p(y=k)}$$

$$= \frac{p(x_1|y)p(x_2|y) \times \dots \times p(x_M|y)p(y)}{\sum_{k=0}^1 p(x_1|y=k)p(x_2|y=k) \times \dots \times p(x_M|y=k)p(y=k)}$$

Example 2:

- Naive Bayes classifier (Probability someone is a female given they are heavy and tall)

$$p(y = 1|x_1 = 1, x_2 = 1) = \frac{p(x_1|y)p(x_2|y)p(y)}{\sum_{k=0}^1 p(x_1|y=k)p(x_2|y=k)p(y=k)}$$

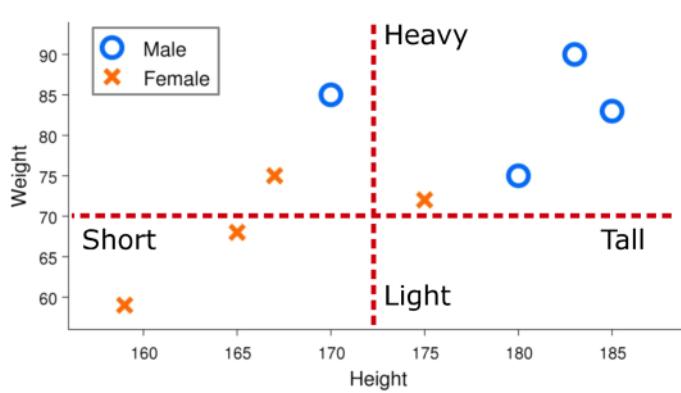


Gender y	Tall x_1	Heavy x_2
0	1	1
0	1	1
0	1	1
0	0	1
1	1	1
1	0	1
1	0	0
1	0	0

Example 2: Solution

- Naive Bayes classifier (Probability someone is a female given they are heavy and tall)

$$\begin{aligned}
 p(y = 1|x_1 = 1, x_2 = 1) &= \frac{p(x_1|y)p(x_2|y)p(y)}{\sum_{k=0}^1 p(x_1|y=k)p(x_2|y=k)p(y=k)} \\
 &= \frac{\frac{1}{4} \frac{2}{4} \frac{1}{2}}{\frac{1}{4} \frac{2}{4} \frac{1}{2} + \frac{3}{4} \frac{4}{4} \frac{1}{2}} = \frac{2}{2+12} = \frac{1}{7}
 \end{aligned}$$



Gender y	Tall x_1	Heavy x_2
0	1	1
0	1	1
0	1	1
0	0	1
1	1	1
1	0	1
1	0	0
1	0	0

Quiz 1, Naive-Bayes (Spring 2012)

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
P1	1	0	0	0	1	1	0	0	1	1
P2	1	0	1	0	0	1	1	1	0	0
P3	0	1	0	1	0	1	0	1	1	1
P4	0	1	1	1	0	0	1	0	0	0
P5	1	0	0	1	1	0	0	1	0	1
P6	1	0	1	1	1	1	1	0	1	0

Table 1: Table indicating whether 10 songs denoted S1-S10 are downloaded to 6 different phones denoted P1-P6. P1 and P2 given in red are phones that belong to females whereas P3, P4, P5, and P6 given in blue belong to males.

The phones P1 and P2 are owned by females whereas P3, P4, P5 and P6 are owned by males (this is indicated in red and blue respectively in Table 1). We would like to predict whether a phone is owned by a male based on whether or not the songs S1, S2 and S3 have been downloaded. We will therefore classify whether the phone belongs to a male or female considering only the attributes S1, S2 and S3 and the data in Table 1. We will apply a Naïve Bayes classifier that assumes independence between these attributes. Given that a phone has installed songs 1, 2 and 3 (i.e., S1=1, S2=1 and S3=1) What is the probability that the phone is owned by a male according to the Naïve Bayes classifier?

- A. 1/12
- B. 1/6
- C. 2/3
- D. 1
- E. Don't know.

$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1|y) \times \dots \times p(x_M|y)p(y)}{\sum_{k=0}^1 p(x_1|y=k) \times \dots \times p(x_M|y=k)p(y=k)}$$

Fix the problem given a robust estimation -> we put a prior in our frequency estimation



Robust estimation and non-binary data

Assume

$$p(x_1, \dots, x_M | y) = \prod_{k=1}^M p(x_k | y)$$

Defining $n_{x_j=k|y=c} = \sum_{i=1}^N \delta_{X_{ij}, k} \delta_{y_i, c}$ we have more generally:

$$\text{Binary case: } p(x_j = 1 | y = c) = \frac{n_{x_j=1|y=c} + \alpha}{N_c + 2\alpha}.$$

$$\text{Categorical case: } p(x_j = k | y = c) = \frac{n_{x_j=k|y=c} + \alpha}{N_c + K\alpha}.$$

$$\text{Continuous case: } p(x_j = x | y = c) = \mathcal{N}(x | \mu = \mu_{j|c}, \sigma^2 = (\sigma_{j|c} + \alpha)^2)$$

$$\mu_{j|c} = \mathbb{E}_{y=c}[x_j] = \frac{1}{N_c} \sum_{i=1}^N \delta_{y_i, c} X_{ij},$$

$$\sigma_{j|c} = \hat{\text{std}}_{y=c}[x_j] = \sqrt{\frac{1}{N_c - 1} \sum_{i=1}^N \delta_{y_i, c} (X_{ij} - \mu_c)^2}$$

We are doing a REGULARIZATION.
If we don't know alph how can we estimate
It ? -> we can use cross validation to figure
Out the best parameter of alpha (with 1-layer)
Or the best parameter and the generalization
error (with 2 layer)

Select these parameters using cross-validation.

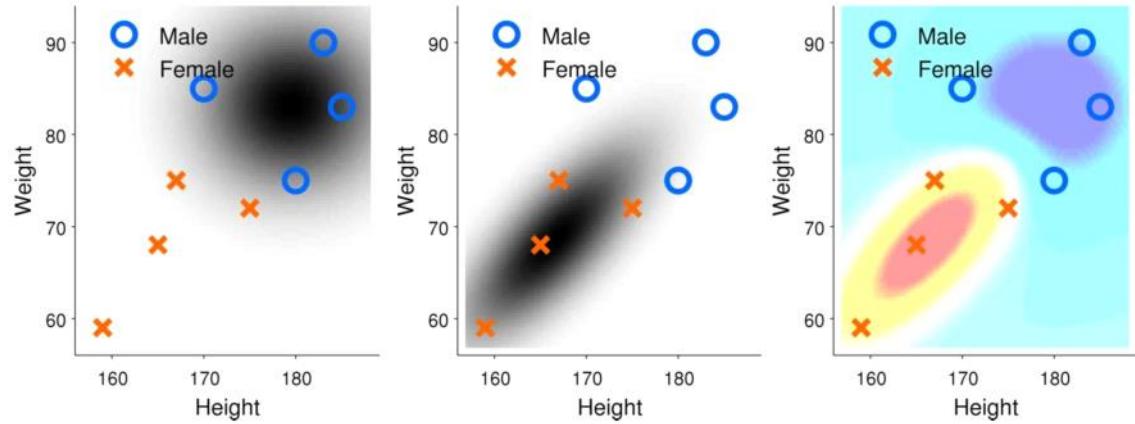
Bayesian classification by the multivariate normal distribution

Continuous density estimation

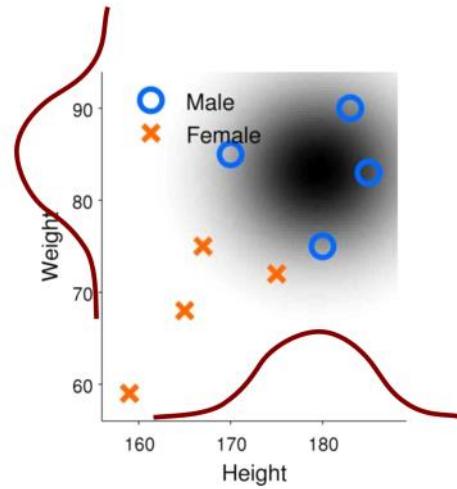
$$P(\mathbf{x}|y = c) = \frac{1}{(2\pi)^{M/2} \det(\Sigma_c)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c)\right)$$

- Fit a Normal distribution to each class
 - Compute class mean and covariance
- Classify using Bayes rule as before

$$P(y = c|\mathbf{x}) = \frac{P(\mathbf{x}|y = c)P(y = c)}{\sum_{c'} P(\mathbf{x}|y = c')P(y = c')}$$



- What does the Naive Bayes assumption of independence of the attributes correspond to in terms of the parameters of the multivariate normal distribution?



Midterm practice test

Look at the test on DTU Learn. Note the test is not part of your evaluation.

Midterm question 1

In the analysis of house prices the following attributes were collected for a house: The year the house was built (denoted YEAR), the size of the house given in square meters (denoted SIZE) the county in which the house is located (denoted LOCATION). Which statement about the three attributes is correct?

- A. YEAR is ratio, SIZE is interval and LOCATION is nominal
- B. YEAR is interval, SIZE is ratio and LOCATION is nominal
- C. YEAR is interval, SIZE is ratio and LOCATION is ordinal
- D. YEAR is interval, SIZE is ratio and LOCATION is interval
- E. Don't know.

Midterm question 2

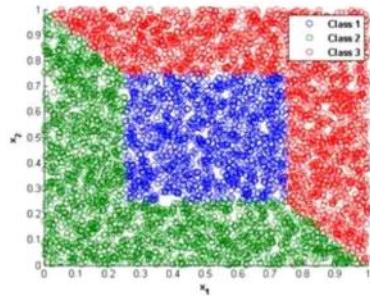
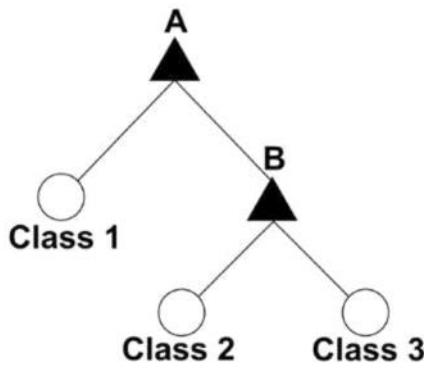


Figure 1



43 DTU Compute

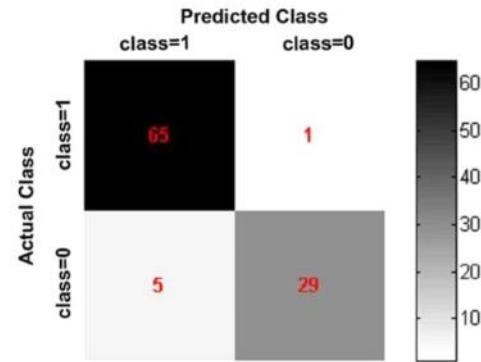
Lecture 7 12 October, 2021

Consider the classification problem given in figure 1 and the Decision Tree shown below it with two decision nodes denoted A and B . We will let $\mathbf{x}_n = (x, y)$ denote a 2-dimensional observation such that $\mathbf{x}_n - 0.5 \cdot \mathbf{1}$ denotes the subtraction of 0.5 from each of the two coordinates of \mathbf{x}_n . Which one of the following classification rules would lead to a correct classification of the data?

- A. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_1 \leq 0.25$, B: $\|\mathbf{x}_n\|_\infty \leq 1$
- B. A: $\|\mathbf{x}_n\|_1 \leq 1$, B: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq \infty$
- C. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B: $\|\mathbf{x}_n\|_\infty \leq 1$
- D. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_\infty \leq 0.25$, B: $\|\mathbf{x}_n\|_1 \leq 1$
- E. Don't know.

Midterm question 3

A classifier has the confusion matrix given in the figure below. Which statement about the classifier is correct?



- A. The Accuracy is 94% and the Error rate is 6%
- B. The Accuracy is 6% and the Error rate is 94%
- C. The Accuracy is 65% and the Error rate is 35%
- D. There is insufficient information in the confusion matrix to determine the Accuracy and Error rate.
- E. Don't know.

Midterm question 4

Which statement about crossvalidation is wrong?

- A. Cross-validation can be used to estimate the generalization error.
- B. Leave one out cross-validation is more computationally expensive than 10 fold crossvalidation.
- C. Holding out one third of the data for validation is faster but less accurate than performing 10 fold cross-validation.
- D. The same test set can be used for model selection as well as evaluation of the generalization performance of the model.
- E. Don't know.

Midterm question 5

Consider a data set of four features: A , B , C , and D that are applied in a classification algorithm. The table below shows the cross-validated Error rate when using different combinations of the features.

Feature(s)	Error rate
A	0.40
B	0.45
C	0.33
D	0.42
A and B	0.20
A and C	0.25
A and D	0.34
B and C	0.29
B and D	0.42
C and D	0.40
A and B and C	0.13
A and B and D	0.17
B and C and D	0.10
A and C and D	0.15
A and B and C and D	0.28

We will apply a forward feature selection algorithm. Which feature set will the selection algorithm choose?

- A. C
- B. B and C and D
- C. A and B
- D. A and B and C
- E. Don't know.

Midterm question 6

When training a decision tree we will use the classification error as impurity measure $I(t)$ given by $I(t) = 1 - \max_i[p(i|t)]$ where $p(i|t)$ denotes the fraction of data objects belonging to class i at a given node t . We will use Hunt's algorithm to grow the tree and recall that the purity gain is given by:

$$\Delta = I(\text{Parent}) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j)$$

where N is the total number of data objects at the parent node, k is the number of child nodes and $N(v_j)$ is the number of data objects associated with the child node, v_j . We will consider classification of Iris flowers into Iris-Setosa, Iris-Virginica and Iris-Versicolor. At a potential split we have:

- Before the split: 5 Iris-Setosa, 10 Iris-Virginica and 10 Iris Versicolor.

After the split

- 0 Iris-Setosa, 8 Iris-Virginica and 2 Iris-Versicolor in the left node.
- 5 Iris-Setosa, 2 Iris-Virginica and 8 Iris-Versicolor in the right node.

Which statement is correct?

- A. The purity gain is $\Delta = \frac{3}{5}$
- B. The purity gain is $\Delta = \frac{3}{15}$
- C. The purity gain is $\Delta = \frac{6}{25}$
- D. The purity gain is $\Delta = \frac{7}{15}$
- E. Don't know.

Midterm question 7

When people are well rested and take an exam their chance of passing the exam is 90%, however, when people are not well rested there chance of passing the exam is only 40%. On any given day 80% of people are well-rested. What is the chance that a person passing

the test is well rested?

- A. $\frac{4}{10}$
- B. $\frac{8}{10}$
- C. $\frac{9}{10}$
- D. $\frac{10}{11}$
- E. Don't know.

Midterm question 8

When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values $\sigma_1 = 4$, $\sigma_2 = 2$, $\sigma_3 = 1$, and $\sigma_4 = 0$.

Which one of the following statements is wrong?

- A. The first principal component accounts for more than 60% of the variation in the data.
- B. The third principal component accounts for less than 5% of the variation in the data.
- C. The second principal component accounts for more than 20% of the variation in the data.
- D. The data can be perfectly represented in a three dimensional sub-space.
- E. Don't know.

Midterm question 9

Consider the following sequence of numbers

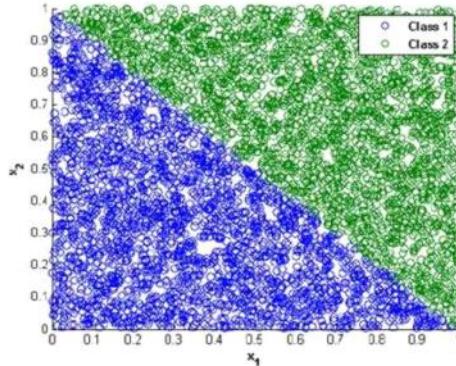
$$x = [0 \ 1 \ 1 \ 1 \ 2 \ 3 \ 4 \ 4 \ 5 \ 14].$$

What is the sum of the mean, the median and the mode of these numbers, i.e. what is the value: $y =$

$\text{mean}(x) + \text{median}(x) + \text{mode}(x)$?

- A. $y = 1$
- B. $y = 6$
- C. $y = 7$
- D. $y = 11$
- E. Don't know.

Midterm question 10



Consider the classification problem given in the figure below where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w_0 . Which one of the following

statements is **wrong**?

- A. The two classes can be perfectly separated by a logistic regression model using x_1 and x_2 as features.
- B. A decision tree with less than five nodes, all of the usual axis-aligned form $x_1 > a$ or $x_2 > b$ for different values of a, b , can perfectly separate the classes using only x_1 and x_2 as features.
- C. A logistic regression model can perfectly separate the two classes using only the feature z given by $z = x_1 + x_2$.
- D. In logistic regression the probability that each observation belong to the two classes can be derived from the logistic function.
- E. Don't know.

Resources

<https://www.youtube.com> Video explaining Naive Bayes

(<https://www.youtube.com/watch?v=8yvBqhm92xA>)

<https://machinelearningmastery.com> Statistical comparison of the cross-validation estimate of the generalization error is not a solved problem. This reference provides an overview of various issues and proposed solutions. Note no simple solution exists.

(<https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/>)



02450: Introduction to Machine Learning and Data Mining

Artificial Neural Networks and Bias/Variance

Jes Frellsen

DTU Compute, Technical University of Denmark (DTU)

DTU Compute

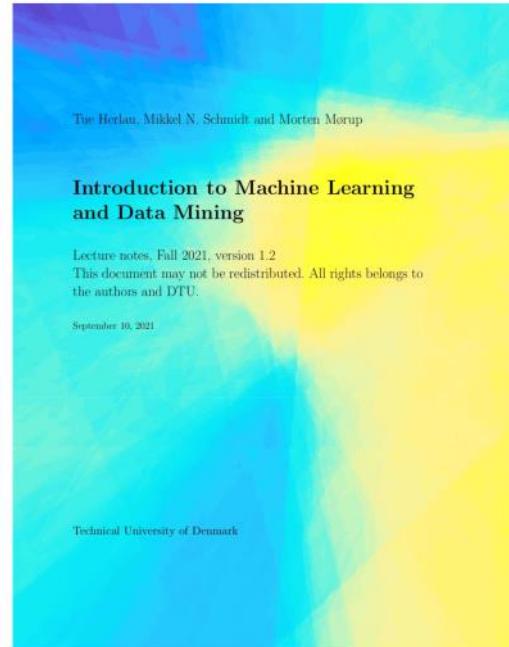
Department of Applied Mathematics and Computer Science

Today

Feedback Groups of the day:

Alexander Lambrecht, Andrea Arduin, Andreas Fabian Emiliussen Nøbbe, Aske Folkmann Musaeus, August Brogaard Tollerup, Aurelio Mottes, Carlo Antonello, Carlo Meroni, Changrun Liu, Charidimos Vratis, David Matthew Lane, Emil Block Overgaard, Felix Høgstædt Larsen, Frederik Engelsborg von Voss, Jakub Piatek Rzeznik, Joakim Bøegh Levorsen, Jonas Søeborg Nielsen, Julien Lacour, Junxuan Shi, Kristian van Kints, Lena Schlüter Nielsen, Line Sandvad Nielsen, Linnéa Haugen, Lorena Torres Lahoz, Mads Dyrved Møller, Magnus Leander Ovason, Mei Lin Verghese Law Kung Sam, Melina Siskou, Mikkel Aaby Kruse, Mikkel Amstrup Krohn, Mikkel Koefoed Lindtner, Mikkel Pedersen, Miriam Emilie Hart, Márton Ferenc Leitold, Nicholas Olesen, Nikolaos Stefanidis, Oliwia Lucyna Lasak, Patrik Kucerka, Peter Meyer Nielsen, Rebekka Steinhart, Roneet Vijay Nagale, Saxo Kilde Jessen Spiele, Siff Kasane Heike Ravn, Steen Nørkjær Larsen, Stefan Skrydstrup Pedersen, Stefanos Rodopoulos, Thor Gabriel Krøgholt-Damasceno, Victor Anton Charles Leweke, Youyang Shen, Yvet Maathuis

Reading material: Chapter 14, Chapter 15



Lecture Schedule

1 Introduction

31 August: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

7 September: C2, C3

3 Measures of similarity, summary statistics and probabilities

14 September: C4, C5

4 Probability densities and data visualization

21 September: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

28 September: C8, C9

6 Overfitting, cross-validation and Nearest Neighbor

5 October: C10, C12 (Project 1 due before 13:00)

7 Performance evaluation, Bayes, and Naive Bayes

12 October: C11, C13

8 Artificial Neural Networks and Bias/Variance

26 October: C14, C15

9 AUC and ensemble methods

2 November: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 November: C18

11 Mixture models and density estimation

16 November: C19, C20 (Project 2 due before 13:00)

12 Association mining

23 November: C21

Recap

13 Recap and discussion of the exam

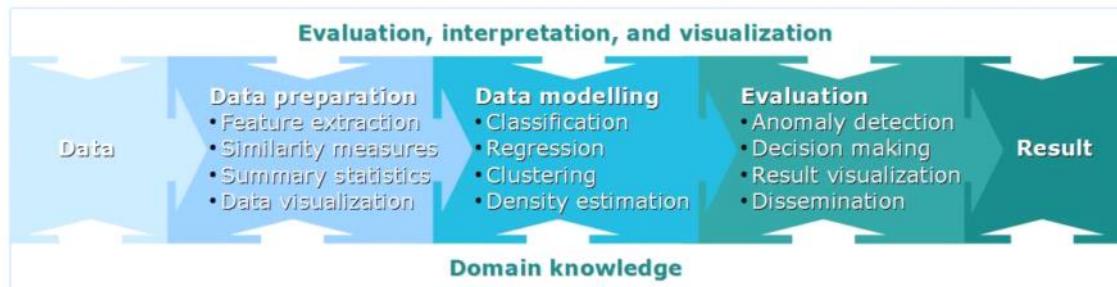
30 November: C1-C21

Online help: Forum on DTU Learn

Videos of lectures: <https://video.dtu.dk>

Streaming of lectures: Zoom (link on DTU Learn)

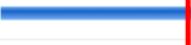
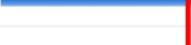
Lecture 8 26 October, 2021

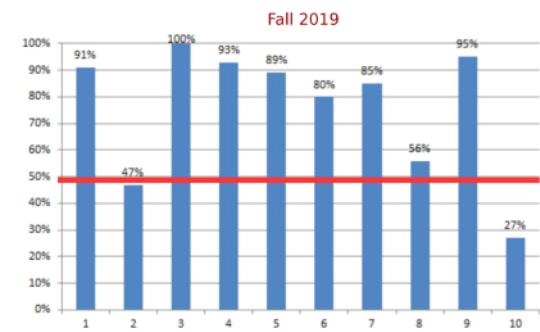
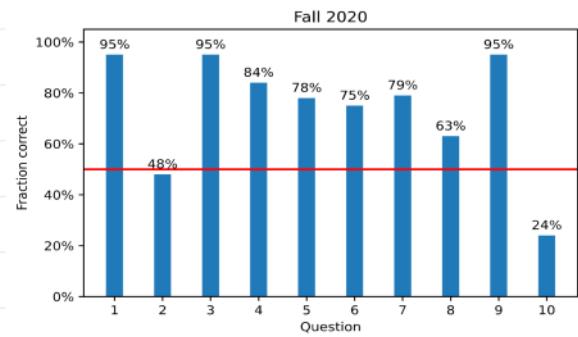


Learning Objectives

- Understand the Bias-Variance decomposition
- Understand and apply regularized least squares regression (i.e. ridge regression)
- Understand the principles behind artificial neural networks (ANNs) and how ANNs can be used for classification and regression
- Understand how logistic regression and ANNs can be extended to multi-class classification

Midterm practice test results

Question 1		94.87 %
Question 2		56.41 %
Question 3		94.87 %
Question 4		87.18 %
Question 5		87.18 %
Question 6		78.21 %
Question 7		78.21 %
Question 8		57.69 %
Question 9		94.87 %
Question 10		24.36 %



Solutions are at the end of this presentation

Question 2:

Consider the classification problem given in figure 1 and the Decision Tree in figure 2 with two decisions denoted A and B. We will let x_n define the x_1 and x_2 coordinates of a given observation whereas $x_n - 0.5 \cdot \mathbf{1}$ denotes the subtraction of 0.5 from x_1 and x_2 .

Which one of the following classification rules would lead to a correct classification of the data?

- A: A: $\|x_n - 0.5 \cdot \mathbf{1}\|_1 \leq 0.25$, B: $\|x_n\|_\infty \leq 1$
- B: A: $\|x_n\|_1 \leq 1$, B: $\|x_n - 0.5 \cdot \mathbf{1}\|_2 \leq \infty$
- C: A: $\|x_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B: $\|x_n\|_\infty \leq 1$
- D: A: $\|x_n - 0.5 \cdot \mathbf{1}\|_\infty \leq 0.25$, B: $\|x_n\|_1 \leq 1$
- E: Don't know

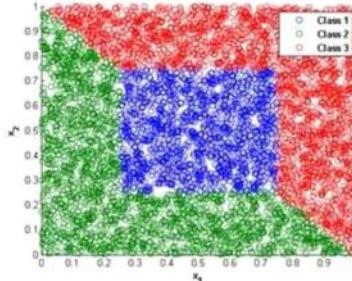
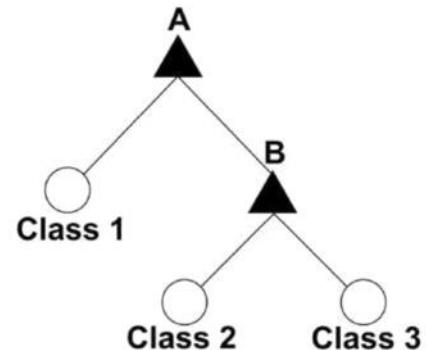


Figure 1

Figure 2
Lecture 8 26 October, 2021

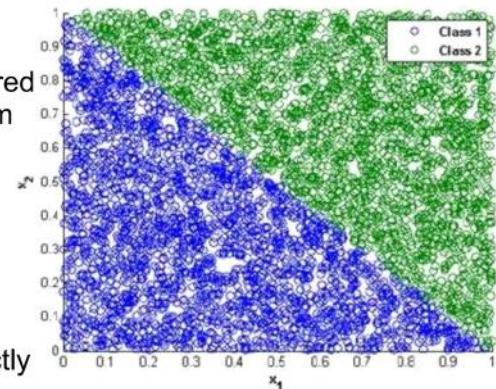
Question 8:

When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values $s_1=4$, $s_2=2$, $s_3=1$, and $s_4=0$. Which one of the following statements is wrong?

- A: The first principal component accounts for more than 60 % of the variation in the data.
- B: The third principal component accounts for less than 5 % of the variation in the data.
- C: The second principal component accounts for more than 20 % of the variation in the data.
- D: The data can be perfectly represented in a three dimensional sub-space.
- E: Don't know.

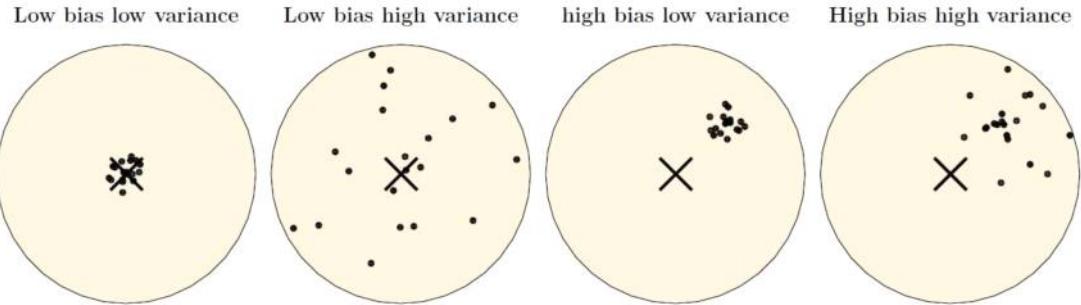
Question 10:

Consider the classification problem given in Figure 5 where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w_0 . Which one of the following statements is wrong?



- A: The two classes can be perfectly separated by a logistic regression model using x_1 and x_2 as features.
- B: A decision tree with less than five nodes can perfectly separate the classes using only x_1 and x_2 as features.
- C: A logistic regression model can perfectly separate the two classes using only the feature t given by $t = x_1 + x_2$.
- D: In logistic regression the probability that each observation belongs to the two classes can be derived from the logistic function.
- E: Don't know.

What is bias and what is variance?



The cross in the middle is the correct value.

- 1) On average all the samples match the target
- 2) The average is really close to the target -> low bias

Regularized least squares

- Recall cost function from linear regression

$$E(\mathbf{w}) = \|\mathbf{y} - \tilde{\mathbf{X}}\mathbf{w}\|^2$$

Reducing some weights towards zero \rightarrow simpler model because it's like removing some features.

- A parsimonious model can be obtained by **forcing** parameters towards zero.
- Problem: Columns of \mathbf{X} have very different scale (i.e. require large/small values of \mathbf{w}) \rightarrow we need to standardize \mathbf{X}
- Therefore, standardize \mathbf{X} :

$$\hat{X}_{ij} = \frac{X_{ij} - \mu_j}{\hat{s}_j}, \quad \mu_j = \frac{1}{N} \sum_{i=1}^N X_{kj}, \quad \hat{s}_j = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_{ij} - \mu_j)^2}$$

- Note $\hat{\mathbf{X}}$ contains no constant term.

Does not contain the offset on the y-axes

Initially, we want to find the weights that minimize the linear regression error, Now I also want to minimize the magnitude of the w vector. How? Take the derivative of the error E and put equal to 0. (where $\lambda \|w\|^2$ is a constant -> so the derivative is 0)

- Introduce regularization term $\lambda \|w\|^2$ to penalize large weights:

$$E_\lambda(w, w_0) = \sum_{i=1}^N (y_i - w_0 - \hat{x}_i^\top w)^2 + \lambda \|w\|^2 = \left\| \mathbf{y} - w_0 \mathbf{1} - \hat{\mathbf{X}} \mathbf{w} \right\|^2 + \lambda \|w\|^2$$

- We can solve for w_0 and w :

$$\begin{aligned} \frac{dE_\lambda}{dw_0} &= \sum_{i=1}^N -2(y_i - w_0 - \hat{x}_i^\top w) = -2N\mathbb{E}[y] - 2Nw_0 - N \left(\frac{1}{N} \sum_{i=1}^N \hat{x}_i^\top \right) w \\ &\Rightarrow w_0 = \mathbb{E}[y] \end{aligned}$$

Because it's standardize this term is 0:
Is the mean of all he columns

- With $\hat{y}_i = y_i - \mathbb{E}[y]$

$$E_\lambda = \left\| \hat{\mathbf{y}} - \hat{\mathbf{X}} \mathbf{w} \right\|^2 + \lambda \|w\|^2$$

- Setting the derivative wrt. w equal to zero and solving for w yields

$$\mathbf{w}^* = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}} + \lambda \mathbf{I}) \setminus (\hat{\mathbf{X}}^\top \hat{\mathbf{y}})$$

I do select the balance between minimize the magnitude of w and minimize the error? Changing lambda.

Selecting λ

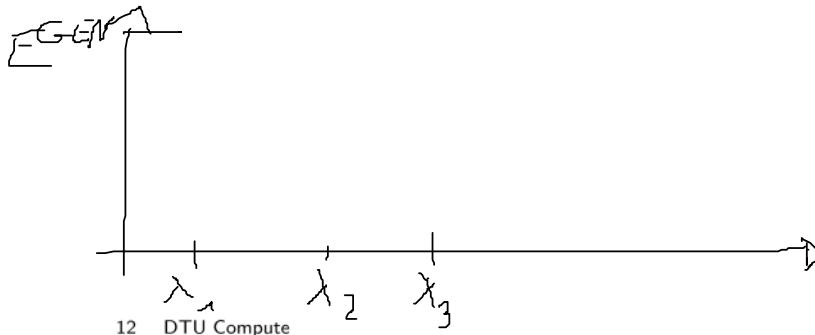
In 1 dimension: if lambda = 0, no effect, only minimize the error
 if lambda $\rightarrow \infty$, it means forget about the least squares error, just minimize the magnitude to 0.



- Suppose

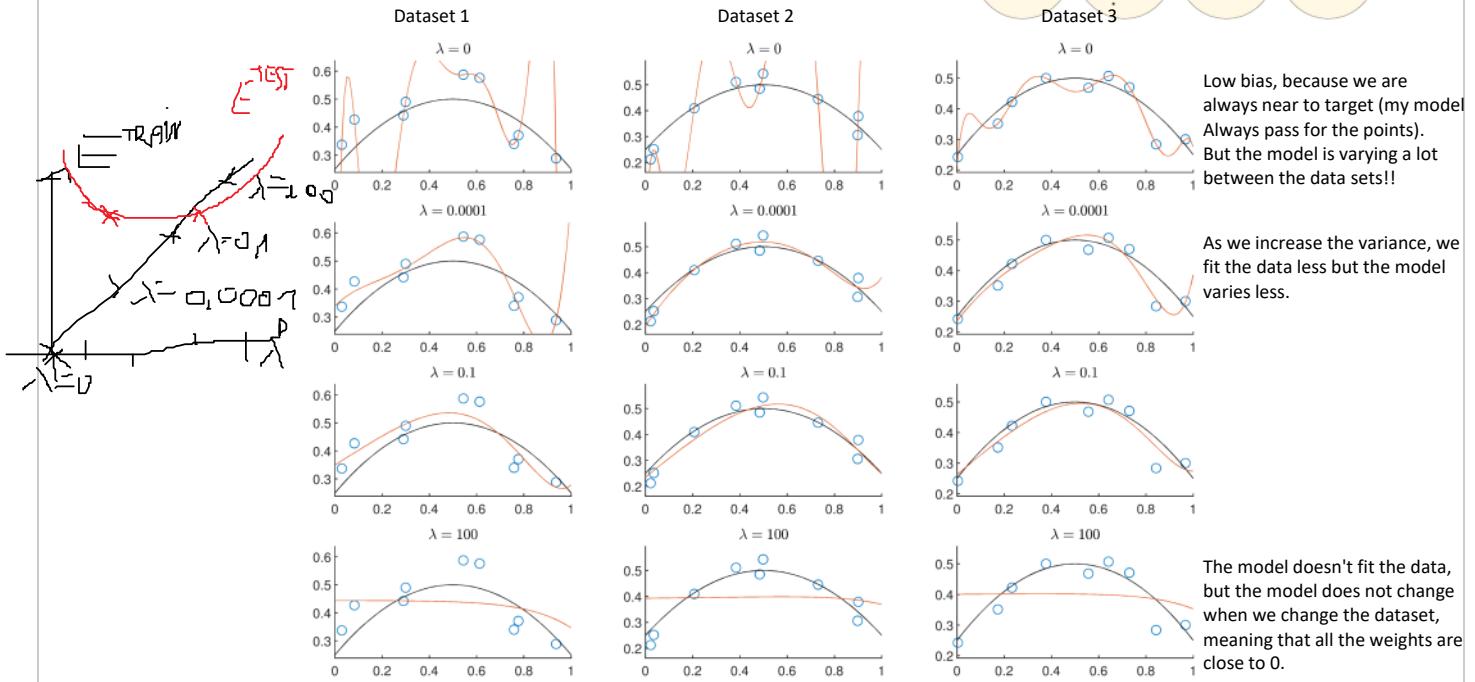
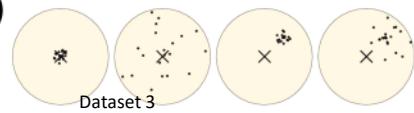
$$\mathbf{w}^* = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}} + \lambda \mathbf{I}) \setminus (\hat{\mathbf{X}}^\top \hat{\mathbf{y}}) \propto \frac{\mathbf{X}\mathbf{y}}{\mathbf{X}^2 + \lambda}$$

- So if $\lambda = 0$ then no effect, else if $\lambda \rightarrow \infty$ then $\mathbf{w}^* \rightarrow 0$
- λ controls complexity of model. Select λ using cross-validation



Lecture 8 26 October, 2021

How does different values of λ (vertical) affect the bias/variance of learned function (red lines)



13 DTU Compute

Lecture 8 26 October, 2021

The Bias-Variance decomposition

We want to decompose the Egen, to get a term which is about the bias and one about the variance

$$p(x,y) = p(x) p(y|x)$$

$$\mathbb{E}_{D,p(x,y)}[E] = E$$

How can we rewrite the $E(e_{\text{gen}}) ??$

$$\mathbb{E}_{\mathcal{D}} [E^{\text{gen}}] = \mathbb{E}_{\mathcal{D},(\mathbf{x},y)} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right]$$

Expectation over
the train data
Expectation over the
Test data ($D, (\mathbf{x}, y)$)

We first consider \mathbf{x} fixed

$$\begin{aligned} & \mathbb{E}_{\mathcal{D},y|\mathbf{x}} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{\mathcal{D},y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}) + \bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] + 2\mathbb{E}_{\mathcal{D},y|\mathbf{x}} [(y - \bar{y}(\mathbf{x})) (\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))] \end{aligned}$$

$$P(x,y) = P(x) P(y|x)$$

$$\mathbb{E}_{\dots} = \dots$$



$$\begin{aligned}
 P(x,y) &= P(x) P(y|x) \\
 \mathbb{E}_{D, P(x,y)}[E] &= \mathbb{E}_D[\mathbb{E}_{P(x,y)}[E]] \\
 &= \mathbb{E}_D[\mathbb{E}_{P_x}[\mathbb{E}_{P_{y|x}}[E]]]
 \end{aligned}$$



/commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_(cropped).jpg

p

The Bias-Variance decomposition



$$\mathbb{E}_D [E^{\text{gen}}] = \mathbb{E}_{D,(\mathbf{x},y)} \left[(y - f_D(\mathbf{x}))^2 \right]$$

We first consider \mathbf{x} fixed

$$\begin{aligned}
 &\mathbb{E}_{D,y|\mathbf{x}} \left[(y - f_D(\mathbf{x}))^2 \right] && \bar{y}(\mathbf{x}) = \mathbb{E}_{y|\mathbf{x}} [y] \\
 &= \mathbb{E}_{D,y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}) + \bar{y}(\mathbf{x}) - f_D(\mathbf{x}))^2 \right] \\
 &= \mathbb{E}_{y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}))^2 \right] + \mathbb{E}_D \left[(\bar{y}(\mathbf{x}) - f_D(\mathbf{x}))^2 \right] + \cancel{2\mathbb{E}_{D,y|\mathbf{x}} [(y - \bar{y}(\mathbf{x}))(\bar{y}(\mathbf{x}) - f_D(\mathbf{x}))]}
 \end{aligned}$$

Does't depend on the test data

Does't depend on the test data



[https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_\(cropped\).jpg](https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_(cropped).jpg)

The Bias-Variance decomposition



$$\mathbb{E}_{\mathcal{D},y|\mathbf{x}} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right] = \mathbb{E}_{y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right]$$

$$\begin{aligned}\bar{f}(\mathbf{x}) &= \mathbb{E}_{\mathcal{D}} [f_{\mathcal{D}}(\mathbf{x})] \\ \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] &= \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}) + \bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] + 2\mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x})) (\bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x})) \right]\end{aligned}$$



16 DTU Compute

[https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_\(cropped\).jpg](https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_(cropped).jpg) Lecture 8 26-October, 2021

The Bias-Variance decomposition



$$\mathbb{E}_{\mathcal{D},y|\mathbf{x}} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right] = \mathbb{E}_{y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right]$$

$$\bar{f}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}} [f_{\mathcal{D}}(\mathbf{x})]$$

It's the mean value of the expectations.

$$\mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right]$$

$$= \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}) + \bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right]$$

$$= \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] + \cancel{2\mathbb{E}_{\mathcal{D}} [(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))(\bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))]}$$

$$\mathbb{E}_{\mathcal{D},y|\mathbf{x}} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right]$$

$$= \mathbb{E}_{y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}))^2 \right] + (\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 + \mathbb{E}_{\mathcal{D}} \left[(\bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right]$$

$$= \text{Var}_{y|\mathbf{x}} [y] + (\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 + \text{Var}_{\mathcal{D}} [f_{\mathcal{D}}(\mathbf{x})]$$

The Bias-Variance decomposition



$$\mathbb{E}_{\mathcal{D}} [E^{\text{gen}}] = \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{\mathcal{D}, y|\mathbf{x}} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \right]$$
$$\mathbb{E}_{\mathcal{D}} [E^{\text{gen}}] = \mathbb{E}_{\mathbf{x}} \left[\text{Var}_{y|\mathbf{x}} [y] + (\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 + \text{Var}_{\mathcal{D}} [f_{\mathcal{D}}(\mathbf{x})] \right]$$



18 DTU Compute

[https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_\(cropped\).jpg](https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_(cropped).jpg) Lecture 8 – 26 October, 2021

The Bias-Variance decomposition

That's how we can rewrite the expectation $E(e_{\text{gen}})$.

$$\mathbb{E}_{\mathcal{D}} [E^{\text{gen}}] = \mathbb{E}_{\mathbf{x}} \left[\text{Var}_{y|\mathbf{x}} [y] + (\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 + \text{Var}_{\mathcal{D}} [f_{\mathcal{D}}(\mathbf{x})] \right]$$

- 1) We can do anything about it, it is not related to the model -> there is going to be a lower bound on our generalization error (if there is a high variance in our data)
- 2) Refers to the bias, if it close to 0 it means that my model is un-biased, and viceversa.

The first term does not depend at all upon our choice of model but simply represents the intrinsic difficulty of the problem. We cannot make this term any larger or smaller by selecting one model over another.

The second term is the **bias** term. It tells us how much the average values of models trained on different training datasets differ compared to the true mean of the data.

The third term is the **variance** term. It tells us how much the model wiggles when trained on different sets of training data. That is, when you train the models on N different (random) sets of training data and the models (the prediction curves) are nearly the same this term is small.

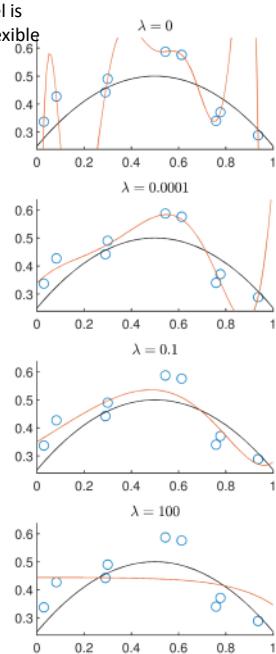
We can understand the model by decomposing the generalization error.



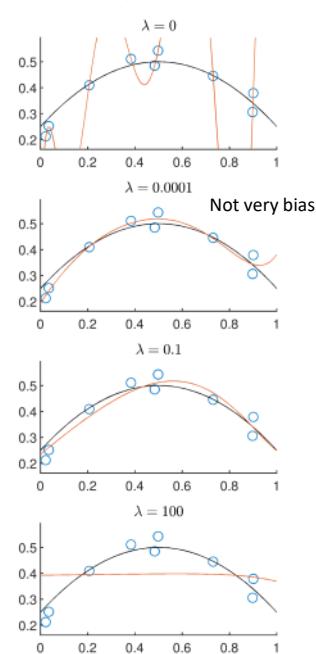
The bias variance decomposition

When lambda = 0, the model is more flexible -> the more flexible the more variance.

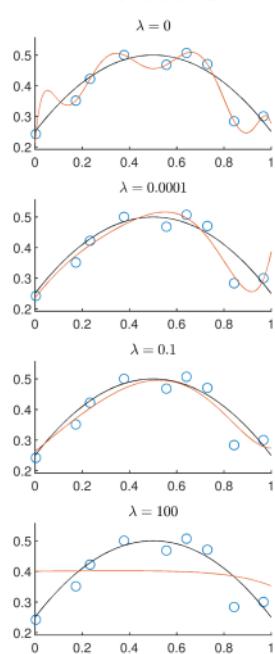
Dataset 1



Dataset 2



Dataset 3



The black line is the mean value of our dataset.

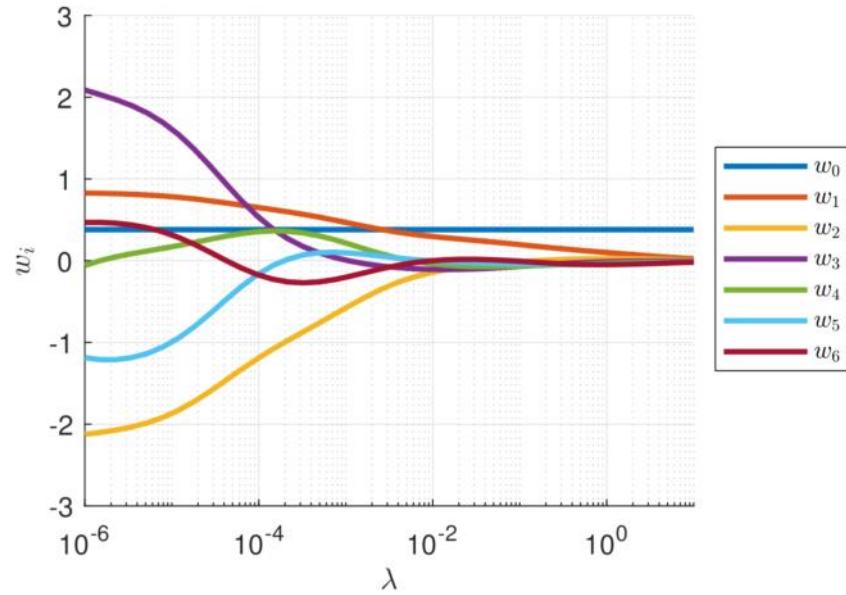
By regularization we can tradeoff bias and variance, in particular, we can hope to substantially reduce variance without introducing too much bias!

A model with 6 features. As we increase the lambda, we squeeze all the weights towards zero, but we don't affect w0 because we do not regularize w0.

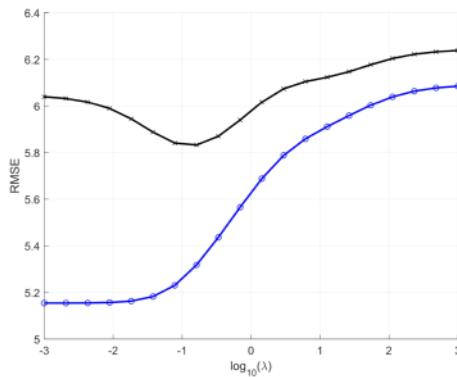


Parameters w^* as function of λ

$$E_\lambda(\mathbf{w}) = \sum_{i=1}^N (\hat{y}_i - w_0 - \hat{\mathbf{x}}_i^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|^2$$



Quiz 1, Bias-variance (Fall 2017)



Using 54 observations of a dataset about Basketball, we would like to predict the average points scored per game (y) based on the four features. For this purpose we consider regularized least squares regression which minimizes with respect to \mathbf{w} the following cost function:

$$E(\mathbf{w}) = \sum_n (y_n - [1 \ x_{n1} \ x_{n2} \ x_{n3} \ x_{n4}] \mathbf{w})^2 + \lambda \mathbf{w}^\top \mathbf{w},$$

We consider 20 different values of λ and use leave-

one-out cross-validation to estimate the performance (measured by mean-squared error) of each of these different values of λ and plot the result in the figure. For the value of $\lambda = 0.6952$ the following model is identified:

$$f(\mathbf{x}) = 2.76 - 0.37x_1 + 0.01x_2 + 7.67x_3 + 7.67x_4.$$

Which one of the following statements is correct?

- A. In the figure the blue curve with circles corresponds to the training error whereas the black curve with crosses corresponds to the test error.
- B. According to the model defined for $\lambda = 0.6952$ increasing a players height x_1 will increase his average points scored per game.
- C. There is no optimal way of choosing λ since increasing λ reduces the variance but increases the bias.
It's not true because the optimal way is using cross-validation.
- D. As we increase λ the 2-norm of the weight vector \mathbf{w} will also increase.
- E. Don't know.

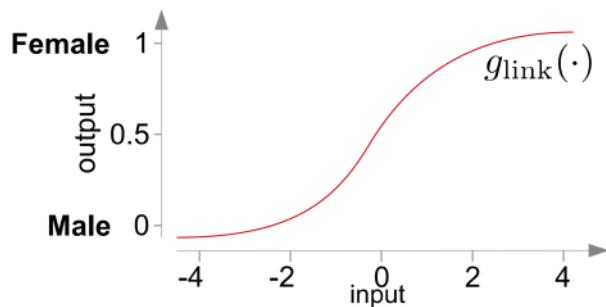
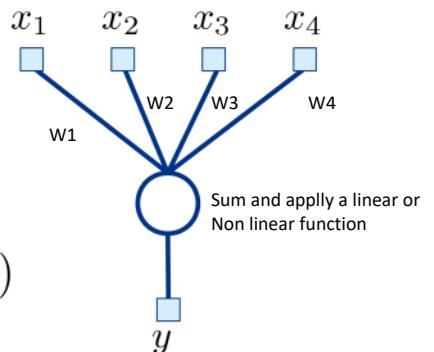
22 DTU Compute We take a linear model and we train it with regularization.
As I increase lambda, we overfit less and less but then I introduce a bias (black line)
The blue is the test error 26 October, 2021

General linear model

- Remember the generalized linear model?

- Data $\{\mathbf{x}_n, y_n\}_{n=1}^N$
- Model $f(\mathbf{x}) = g_{\text{link}}(\mathbf{x}^\top \mathbf{w})$ Sigmoid
- Cost function $d(y, f(\mathbf{x}))$ Square norm
- Parameters $\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{n=1}^N d(y_n, f(\mathbf{x}_n))$

We multiply x by the weights

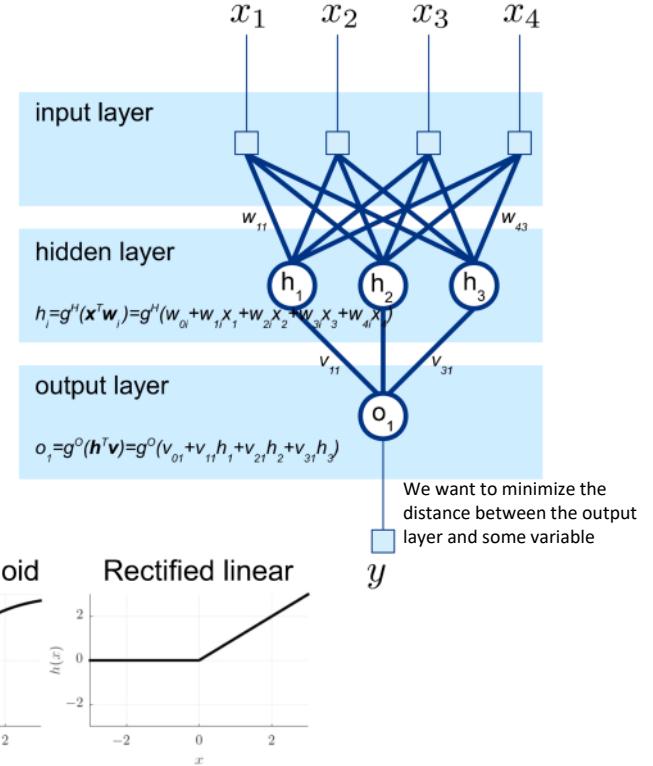


Each node/neuron is a linear model

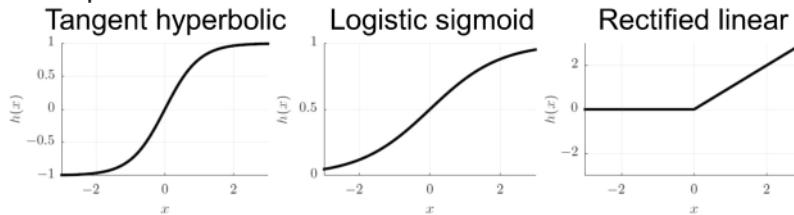
Artificial neural networks

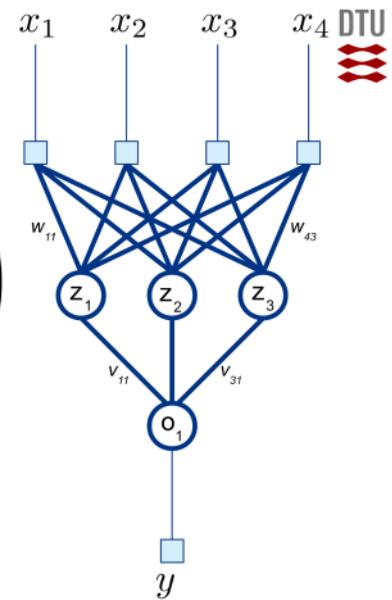
Feed forward network

- Each "neuron"
 - Computes a non-linear function of the sum of its inputs
 - Is just like a generalized linear model
 - Has its own set of parameters
- Modeling choices
 - Cost function
 - Non-linearities
 - Number of neurons and hidden layers
 - Selection of inputs
- Parameter estimation using numerical optimization methods
- Very flexible model: Can easily overfit



Example of non-linearities:





Data: $\{\mathbf{x}_i, y_i\}$

Even having different layer, we end up with only one model.

$$\text{Model: } f(\mathbf{x}) = h^{(2)} \left(v_{10} + \sum_{j=1}^H v_{1j} h^{(1)} (\tilde{\mathbf{x}}^\top \mathbf{w}_j) \right)$$

Distance: $d(y, f(\mathbf{x}))$

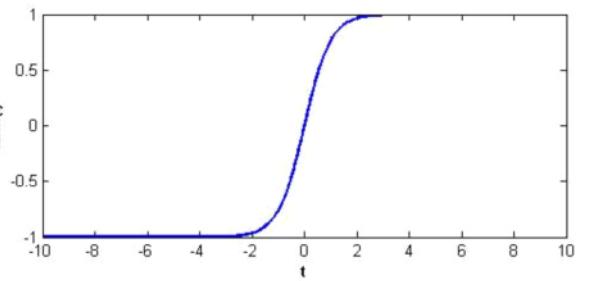
$$\text{Cost: } E = \sum_{i=1}^N d(y_i, f(\mathbf{x}_i))^2$$

Common choices

$$h^{(1)}(x) = \tanh(x)$$

$$h^{(2)}(x) = x$$

$$d(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$$



Neurons and layers

Recall:

$$f(\mathbf{x}) = h^{(2)} \left(v_{10} + \sum_{j=1}^H v_{1j} h^{(1)} (\tilde{\mathbf{x}}^\top \mathbf{w}_j) \right)$$

h is the non linear function (sigmoid)

- Let $z_j^{(1)}$ be output of j 'th hidden unit

$$z_j^{(1)} = h^{(1)} \left(\mathbf{w}_j^{(1) \top} \tilde{\mathbf{x}} \right)$$

Abbreviated $\mathbf{z}^{(1)} = h^{(1)} (\mathbf{W}^{(1)} \tilde{\mathbf{x}})$

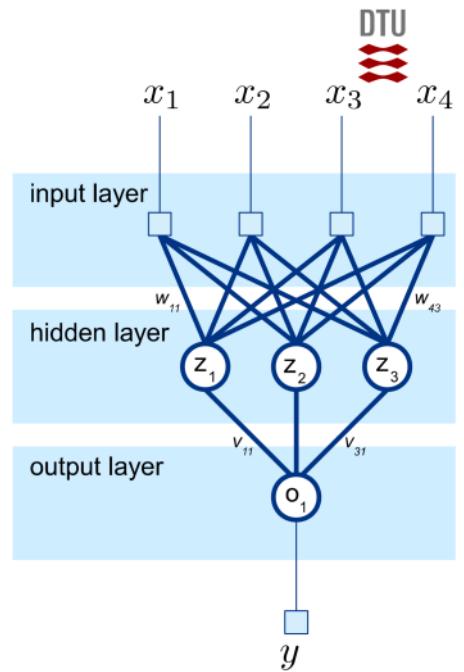
- Output

$$f(\mathbf{x}) = h^{(2)} \left(v_{10} + \sum_{j=1}^H v_{1j} z_j^{(1)} \right) = h^{(2)} \left(\mathbf{W}^{(2)} \mathbf{z}^{(1)} \right)$$

We consider each $z_j^{(1)}$ a neuron and $\mathbf{z}^{(1)}$ a (hidden) layer

26 DTU Compute

Lecture 8 26 October, 2021



$$\mathbf{z}^{(1)} = \begin{bmatrix} z_1^{(1)} \\ z_2^{(1)} \\ z_3^{(1)} \end{bmatrix} = h^{(1)} \left[\begin{bmatrix} \mathbf{w}_1^{(1)} & \mathbf{w}_2^{(1)} & \mathbf{w}_3^{(1)} \end{bmatrix} \mathbf{x} \right]$$

Quiz 2, Artificial Neural Network (Fall 2017)



We will consider an artificial neural network (ANN) trained to predict the average score of a player (i.e., y). The ANN is based on the model:

$$f(\mathbf{x}, \mathbf{w}) = w_0^{(2)} + \sum_{j=1}^2 w_j^{(2)} h^{(1)}([1 \ \mathbf{x}] \mathbf{w}_j^{(1)}).$$

where $h^{(1)}(x) = \max(x, 0)$ is the rectified linear function used as activation function in the hidden layer (i.e., positive values are returned and negative values are set to zero). We will consider an ANN with two hidden units in the hidden layer defined by:

$$\mathbf{w}_1^{(1)} = \begin{bmatrix} 21.78 \\ -1.65 \\ 0 \\ -13.26 \\ -8.46 \end{bmatrix}, \quad \mathbf{w}_2^{(1)} = \begin{bmatrix} -9.60 \\ -0.44 \\ 0.01 \\ 14.54 \\ 9.50 \end{bmatrix},$$

and $w_0^{(2)} = 2.84$, $w_1^{(2)} = 3.25$, and $w_2^{(2)} = 3.46$.

What is the predicted average score of a basketball player with observation vector $\mathbf{x}^* = [6.8 \ 225 \ 0.44 \ 0.68]$?

- A. 1.00
- B. 3.74
- C. 8.21
- D. 11.54

$$f(\mathbf{x}, \mathbf{w}) = w_0^{(0)} + w_1^{(1)} \max(0, [\mathbf{1} \ \tilde{\mathbf{x}}]^T \mathbf{w}_1^{(1)}) + w_2^{(2)} \max(0, [\mathbf{1} \ \tilde{\mathbf{x}}]^T \mathbf{w}_2^{(2)}) + \varepsilon$$

$$= 2.84 + 3.25 \max(0, \begin{bmatrix} 1 \\ 68 \\ 0.25 \\ 0.44 \\ 0.63 \end{bmatrix}^T \begin{bmatrix} 2128 \\ -165 \end{bmatrix})$$

Generalization 1: Multiple outputs

- As before define: $\mathbf{z}^{(1)} = h^{(1)}(\mathbf{W}^{(1)} \tilde{\mathbf{x}})$

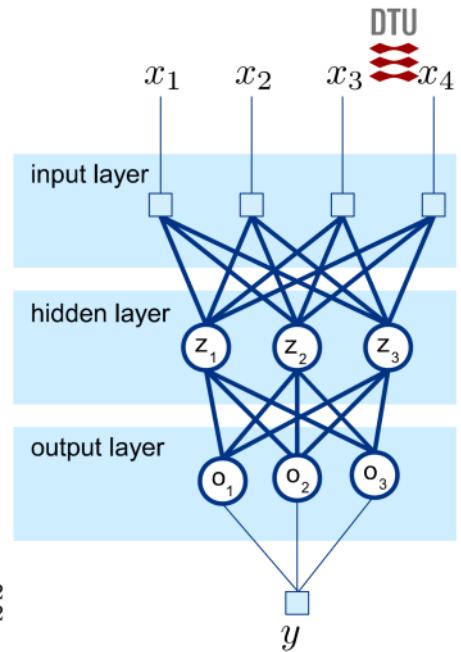
- Now let $\mathbf{W}^{(2)}$ be a $C \times H$ matrix then:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) = h^{(2)}(\mathbf{W}^{(2)} \mathbf{z}^{(1)})$$

will be C -dimensional

- Re-define error function

$$E = \sum_{i=1}^N \|\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i)\|_2^2$$



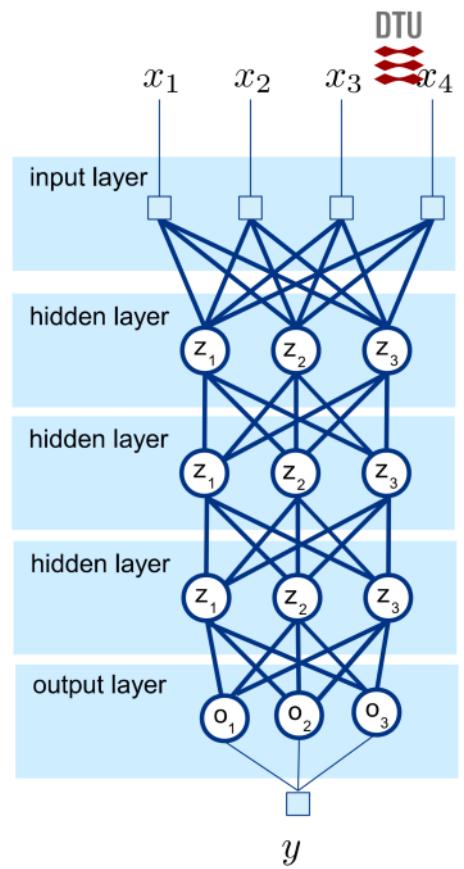
Generalization 2: Multiple layers

- Define $\mathbf{z}^{(0)} = \mathbf{x}$
- For each layer $l = 1, \dots, L$ compute

$$z_j^{(l)} = h^{(l)} (\mathbf{W}^{(l)} \mathbf{z}^{(l-1)})$$

- Output is simply

$$\mathbf{f}(\mathbf{x}) = \mathbf{z}^{(L)}$$



For learning the weights.

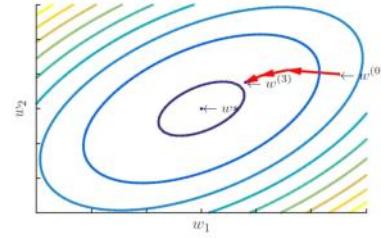
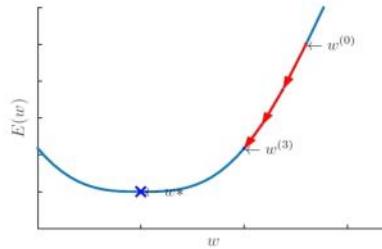
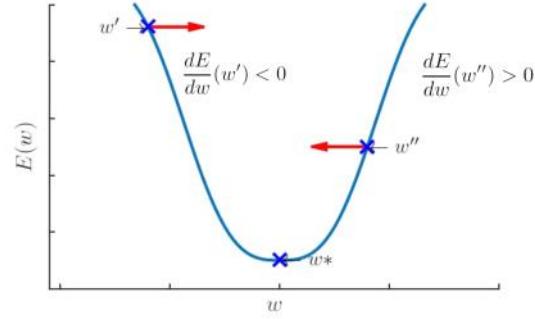


Gradient descent

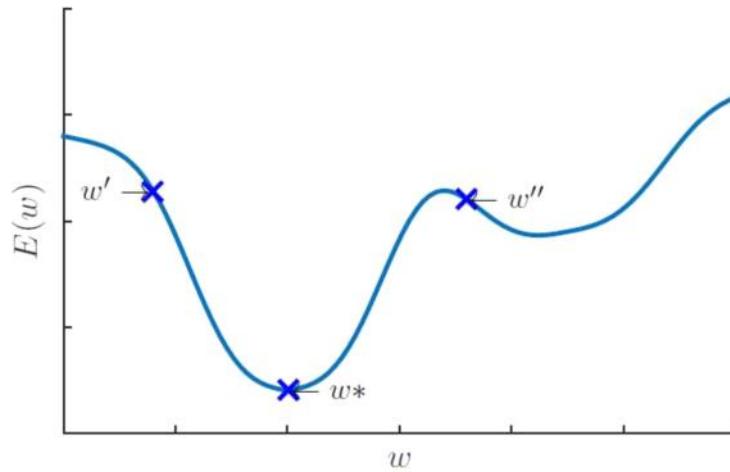
- Start from an initial guess at \mathbf{w}^* , $\mathbf{w}^{(0)}$
- At step t of the algorithm, modify $\mathbf{w}^{(t-1)}$ to produce a better guess $\mathbf{w}^{(t)}$:

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \epsilon \frac{dE}{dw}(\mathbf{w}^{(t-1)})$$

ϵ tells how big the step along the minimum gradient direction should be: the direction tell us which directions we should move to to minimize the error



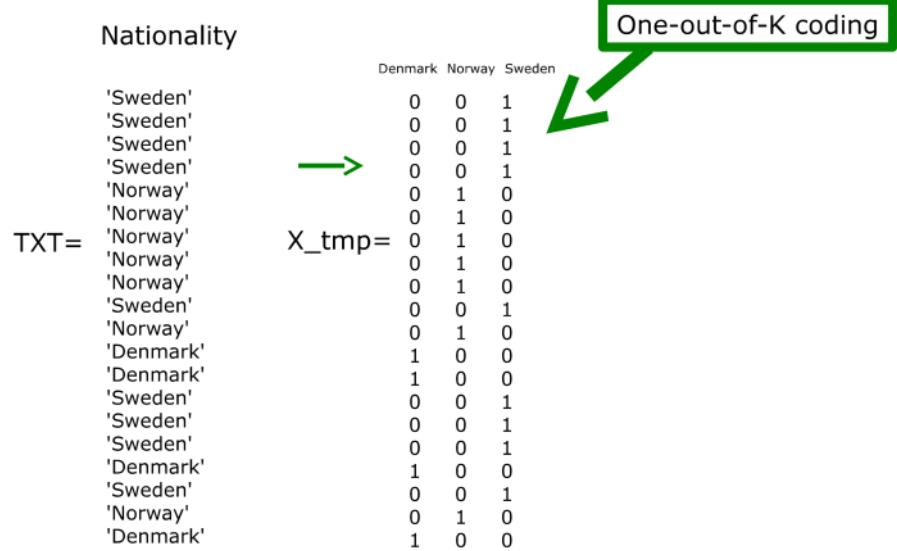
Contrary to least-squares linear regression and logistic regression ANNs have issues of local minima



There are local min max, so the gradient descent really depends on the initialization because if we start in a local minima then the gradient descent tell to don't move.

Single and multi-class: One out of K coding

Nationality



		Denmark	Norway	Sweden
'Sweden'	X_tmp=	0	0	1
'Sweden'		0	0	1
'Sweden'		0	0	1
'Sweden'		0	0	1
'Norway'		0	1	0
'Norway'		0	1	0
'Norway'		0	1	0
'Norway'		0	1	0
'Sweden'		0	0	1
'Norway'		0	1	0
'Denmark'		1	0	0
'Denmark'		1	0	0
'Sweden'		0	0	1
'Sweden'		0	0	1
'Denmark'		1	0	0
'Sweden'		0	0	1
'Norway'		0	1	0
'Denmark'		1	0	0

Multi-class classification

- Logistic regression, $y = 0, 1$:

$$p(y|\theta) = \theta^y (1 - \theta)^{1-y}$$

$$\theta = \sigma(\mathbf{x}^\top \mathbf{w})$$

- Multinomial regression, $y = 1, 2, \dots, K$

z_k : one-of- K encoding of y ,

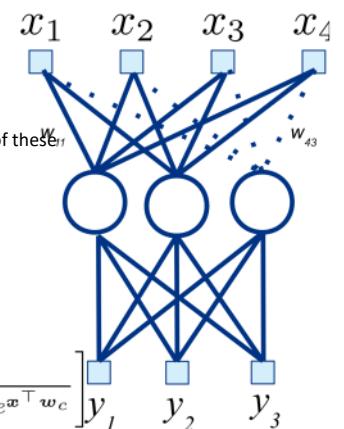
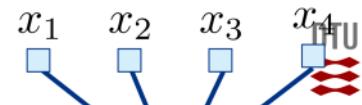
Theta vector: $p(y|\theta) = \prod_{i=1}^K \theta_k^{z_k}$ Categorical distribution: if we take a one-hot-encoding then only one of these z_k is going to be equal to one, the other will be 0s.

Class 1

Theta2 is the prob of being in class 2 $\theta = \text{softmax}([\mathbf{x}^\top \mathbf{w}_1 \dots \mathbf{x}^\top \mathbf{w}_K])$

Softmax: everything sum to 1. It's a sort of regularization of the sigmoid function

$$\text{or: } \theta = \left[\frac{e^{\mathbf{x}^\top \mathbf{w}_1}}{1 + \sum_{c=1}^{K-1} e^{\mathbf{x}^\top \mathbf{w}_c}} \dots \frac{e^{\mathbf{x}^\top \mathbf{w}_{K-1}}}{1 + \sum_{c=1}^{K-1} e^{\mathbf{x}^\top \mathbf{w}_c}} \frac{1}{1 + \sum_{c=1}^{K-1} e^{\mathbf{x}^\top \mathbf{w}_c}} \right] y_1 \quad y_2 \quad y_3$$



We can both apply the softmax to linear regression or to neural network.



Connection to neural networks

Multinomial regression:

- Define:

$$\boldsymbol{\theta} = \begin{bmatrix} \frac{e^{\mathbf{x}^\top \mathbf{w}_1}}{1 + \sum_{c=1}^{K-1} e^{\mathbf{x}^\top \mathbf{w}_c}} & \dots & \frac{1}{1 + \sum_{c=1}^{K-1} e^{\mathbf{x}^\top \mathbf{w}_c}} \end{bmatrix}$$

- Cost function is (z_{ic} is one-of- K encoding of y_i)

$$E = -\sum_{i=1}^N \log p(y_i | \mathbf{x}_i) = -\sum_{i=1}^N \sum_{c=1}^K z_{ic} \log \theta_{ic}$$

Multi-class neural network:

- Suppose $\tilde{y}_1, \dots, \tilde{y}_K$ are outputs of a neural network

- Define

$$\boldsymbol{\theta} = \begin{bmatrix} \frac{e^{\tilde{\mathbf{y}}^\top \mathbf{w}_1}}{\sum_{c=1}^K e^{\tilde{\mathbf{y}}^\top \mathbf{w}_c}} & \dots & \frac{e^{\tilde{\mathbf{y}}^\top \mathbf{w}_K}}{\sum_{c=1}^K e^{\tilde{\mathbf{y}}^\top \mathbf{w}_c}} \end{bmatrix}$$

- Cost function is:

$$E = -\sum_{i=1}^N \log p(y_i | \tilde{\mathbf{y}}_i) = -\sum_{i=1}^N \sum_{c=1}^K z_{ic} \log \theta_{ic}$$

Take an example point: $x = [0 \ 1]$ then it should be classified as class 1



Quiz 3, Multinomial Regression (Spring 2016)

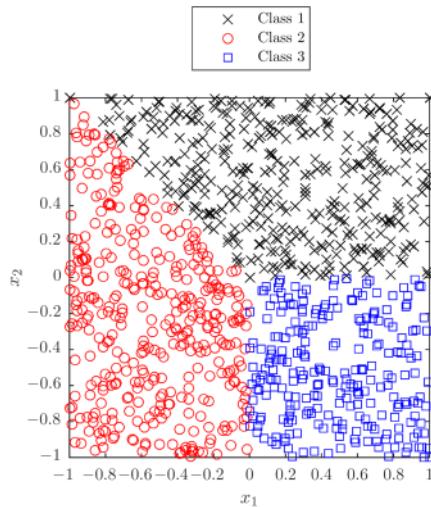


Figure 1: Observations labelled with the most probable class

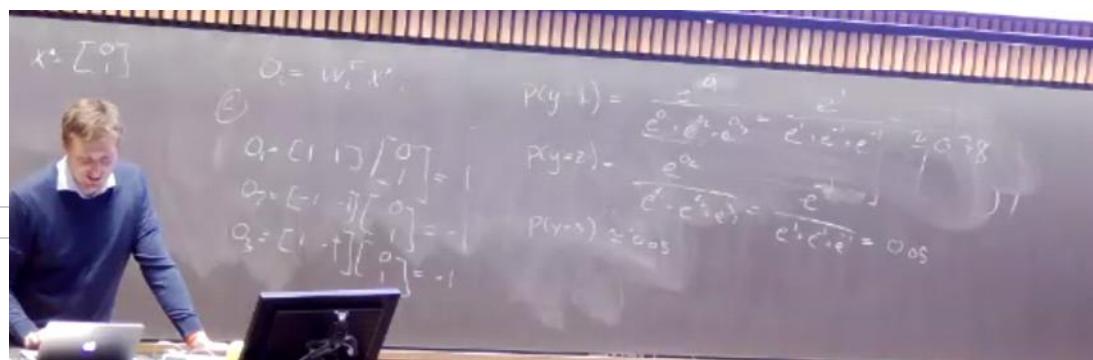
Consider a multinomial regression classifier for

a three-class problem where for each point $\mathbf{x} = [x_1 \ x_2]^\top$ we compute the class-probability using the softmax function

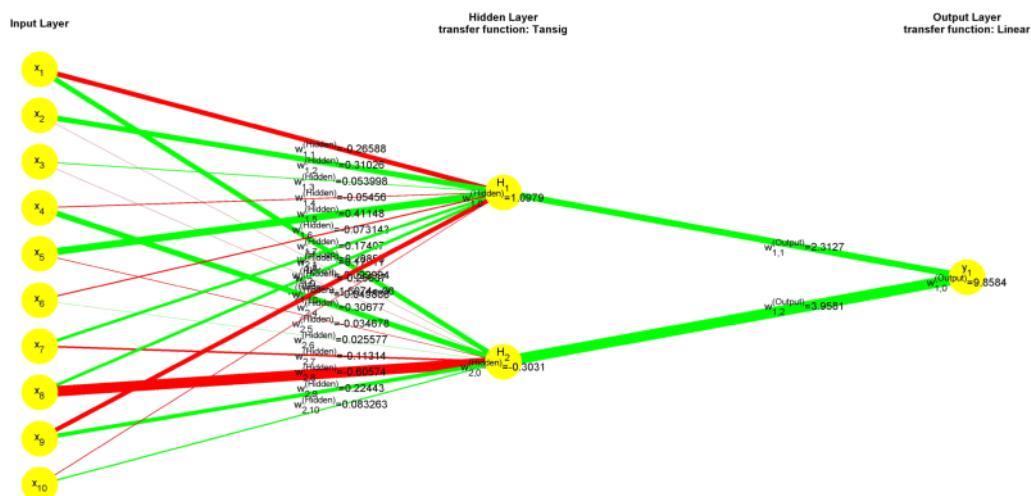
$$P(\hat{y} = k) = \frac{e^{\mathbf{w}_k^\top \mathbf{x}}}{e^{\mathbf{w}_1^\top \mathbf{x}} + e^{\mathbf{w}_2^\top \mathbf{x}} + e^{\mathbf{w}_3^\top \mathbf{x}}}.$$

A dataset of $N = 1000$ points where each point is labeled according to the maximum class-probability is shown in Figure 1. Which setting of the weights was used?

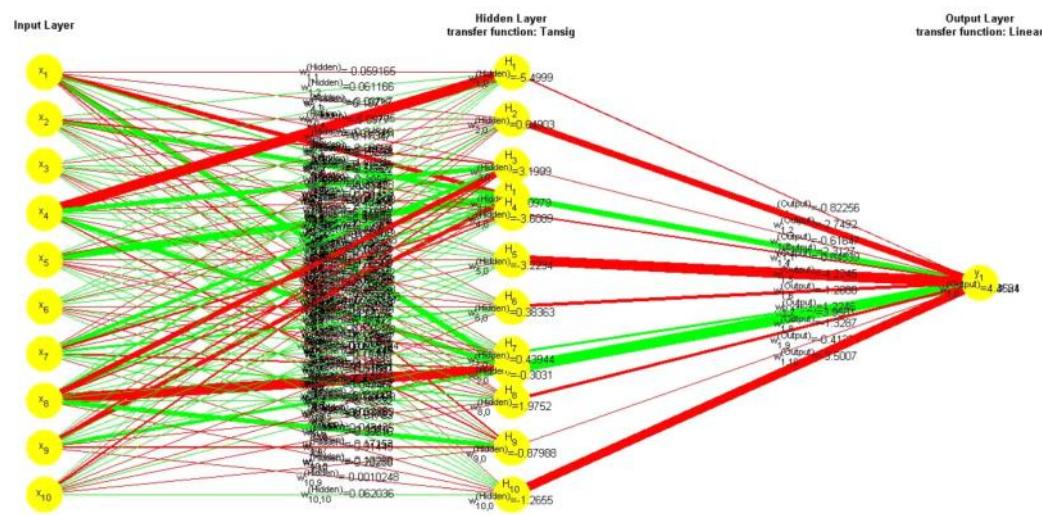
- A. $w_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
- B. $w_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
- C. $w_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
- D. $w_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$
- E. Don't know.



Interpreting neural networks can be difficult



Interpreting neural networks can be difficult



Resources

<https://www.youtube.com> Excellent video resource explaining the concepts behind neural networks

(https://www.youtube.com/watch?v=aircAruvnKk&list=PLZHQ0b0WTQDNU6R1_67000Dx_ZCJB-3pi)

<http://playground.tensorflow.org> Sleek interactive neural network example where you can examine the effect of different number of hidden neurons, activation functions, and many other things on training (<http://playground.tensorflow.org/>)

<https://www.tensorflow.org> Most popular and well-documented deep learning framework. While well documented, notice it requires some python knowledge (<https://www.tensorflow.org/>)

<https://pytorch.org> Upcoming (and in some ways slightly simpler) framework for deep learning; alternative to tensorflow

(<https://pytorch.org/>)

Mid-term quiz 1

In the analysis of house prices the following attributes were collected for a house: The year the house was built (denoted YEAR), the size of the house given in square meters (denoted SIZE) the county in which the house is located (denoted LOCATION). Which statement about the three attributes is correct?

- A. YEAR is ratio, SIZE is interval and LOCATION is nominal
- B. YEAR is interval, SIZE is ratio and LOCATION is nominal
- C. YEAR is interval, SIZE is ratio and LOCATION is ordinal
- D. YEAR is interval, SIZE is ratio and LOCATION is interval
- E. Don't know.

Mid-term quiz 2

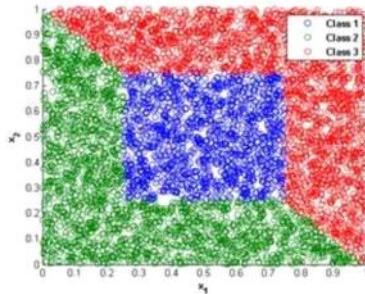
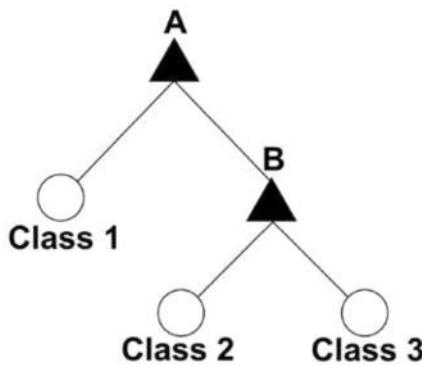


Figure 1

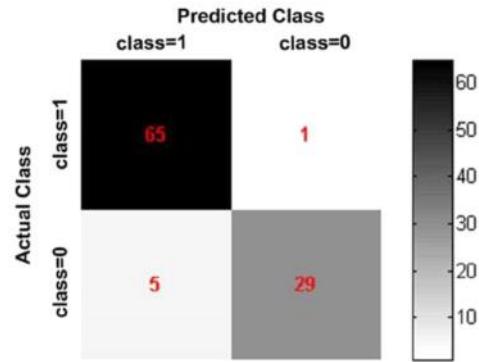


Consider the classification problem given in figure 1 and the Decision Tree shown below it with two decision nodes denoted A and B . We will let $\mathbf{x}_n = (x, y)$ denote a 2-dimensional observation such that $\mathbf{x}_n - 0.5 \cdot \mathbf{1}$ denotes the subtraction of 0.5 from each of the two coordinates of \mathbf{x}_n . Which one of the following classification rules would lead to a correct classification of the data?

- A. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_1 \leq 0.25$, B: $\|\mathbf{x}_n\|_\infty \leq 1$
- B. A: $\|\mathbf{x}_n\|_1 \leq 1$, B: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq \infty$
- C. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B: $\|\mathbf{x}_n\|_\infty \leq 1$
- D. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_\infty \leq 0.25$, B: $\|\mathbf{x}_n\|_1 \leq 1$
- E. Don't know.

Mid-term quiz 3

A classifier has the confusion matrix given in the figure below. Which statement about the classifier is correct?



- A. The Accuracy is 94% and the Error rate is 6%
- B. The Accuracy is 6% and the Error rate is 94%
- C. The Accuracy is 65% and the Error rate is 35%
- D. There is insufficient information in the confusion matrix to determine the Accuracy and Error rate.
- E. Don't know.

Mid-term quiz 4

Which statement about crossvalidation is wrong?

- A. Cross-validation can be used to estimate the generalization error.
- B. Leave one out cross-validation is more computationally expensive than 10 fold crossvalidation.
- C. Holding out one third of the data for validation is faster but less accurate than performing 10 fold cross-validation.
- D. The same test set can be used for model selection as well as evaluation of the generalization performance of the model.
- E. Don't know.

Mid-term quiz 5

Consider a data set of four features: A , B , C , and D that are applied in a classification algorithm. The table below shows the cross-validated Error rate when using different combinations of the features.

Feature(s)	Error rate
A	0.40
B	0.45
C	0.33
D	0.42
A and B	0.20
A and C	0.25
A and D	0.34
B and C	0.29
B and D	0.42
C and D	0.40
A and B and C	0.13
A and B and D	0.17
B and C and D	0.10
A and C and D	0.15
A and B and C and D	0.28

We will apply a forward feature selection algorithm. Which feature set will the selection algorithm choose?

- A. C
- B. B and C and D
- C. A and B
- D. A and B and C
- E. Don't know.

Mid-term quiz 6

When training a decision tree we will use the classification error as impurity measure $I(t)$ given by $I(t) = 1 - \max_i[p(i|t)]$ where $p(i|t)$ denotes the fraction of data objects belonging to class i at a given node t . We will use Hunt's algorithm to grow the tree and recall that the purity gain is given by:

$$\Delta = I(\text{Parent}) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j)$$

where N is the total number of data objects at the parent node, k is the number of child nodes and $N(v_j)$ is the number of data objects associated with the child node, v_j . We will consider classification of Iris flowers into Iris-Setosa, Iris-Virginica and Iris-Versicolor. At a potential split we have:

- Before the split: 5 Iris-Setosa, 10 Iris-Virginica and 10 Iris Versicolor.

After the split

- 0 Iris-Setosa, 8 Iris-Virginica and 2 Iris-Versicolor in the left node.
- 5 Iris-Setosa, 2 Iris-Virginica and 8 Iris-Versicolor in the right node.

Which statement is correct?

- A. The purity gain is $\Delta = \frac{3}{5}$
- B. The purity gain is $\Delta = \frac{3}{15}$
- C. The purity gain is $\Delta = \frac{6}{25}$
- D. The purity gain is $\Delta = \frac{7}{15}$
- E. Don't know.

Mid-term quiz 7

When people are well rested and take an exam their chance of passing the exam is 90%, however, when people are not well rested there chance of passing the exam is only 40%. On any given day 80% of people are well-rested. What is the chance that a person passing

the test is well rested?

- A. $\frac{4}{10}$
- B. $\frac{8}{10}$
- C. $\frac{9}{10}$
- D. $\frac{10}{11}$
- E. Don't know.

Mid-term quiz 8

When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values $\sigma_1 = 4$, $\sigma_2 = 2$, $\sigma_3 = 1$, and $\sigma_4 = 0$.

Which one of the following statements is wrong?

- A. The first principal component accounts for more than 60% of the variation in the data.
- B. The third principal component accounts for less than 5% of the variation in the data.
- C. The second principal component accounts for more than 20% of the variation in the data.
- D. The data can be perfectly represented in a three dimensional sub-space.
- E. Don't know.

Mid-term quiz 9

Consider the following sequence of numbers

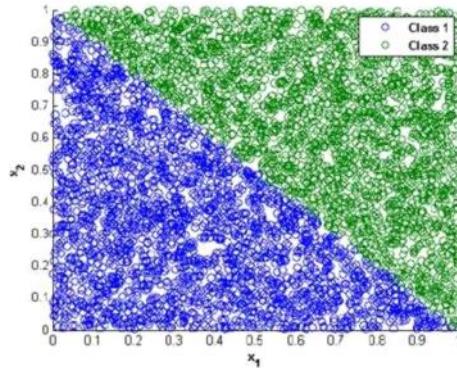
$$x = [0 \ 1 \ 1 \ 1 \ 2 \ 3 \ 4 \ 4 \ 5 \ 14].$$

What is the sum of the mean, the median and the mode of these numbers, i.e. what is the value: $y =$

$\text{mean}(x) + \text{median}(x) + \text{mode}(x)$?

- A. $y = 1$
- B. $y = 6$
- C. $y = 7$
- D. $y = 11$
- E. Don't know.

Mid-term quiz 10



Consider the classification problem given in the figure below where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w_0 . Which one of the following

statements is **wrong**?

- A. The two classes can be perfectly separated by a logistic regression model using x_1 and x_2 as features.
- B. A decision tree with less than five nodes, all of the usual axis-aligned form $x_1 > a$ or $x_2 > b$ for different values of a, b , can perfectly separate the classes using only x_1 and x_2 as features.
- C. A logistic regression model can perfectly separate the two classes using only the feature z given by $z = x_1 + x_2$.
- D. In logistic regression the probability that each observation belong to the two classes can be derived from the logistic function.
- E. Don't know.

02450: Introduction to Machine Learning and Data Mining

AUC and ensemble methods

Jes Frellsen

DTU Compute, Technical University of Denmark (DTU)

DTU Compute

Department of Applied Mathematics and Computer Science

Today

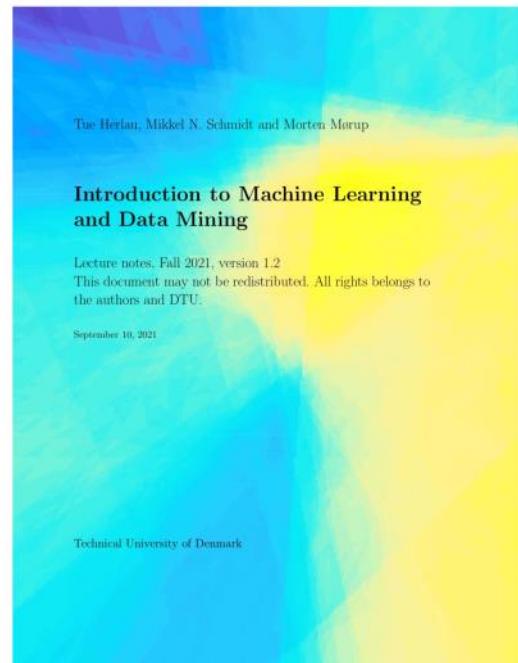
Feedback Groups of the day:

Adrian Roed Schøning, Alaina Ann Martinez,
Amanda Aagaard Uldal, Anna Venetsanou, Annette
Lien, Antoni Wojciech Skrobisz, Asbjørn Ebenezer
Magnussen, Asha Omar Abdirahman Haji, Asta Marie
Nielsen, Camilla Lind Ommen, Caroline Ulstrup
Larsen, Christian Adam Deding Nielsen, Christine
Ibæk Topp Lindenhoff, Claes Jens Sjælborg
Lindhardt, Daniel Pedrosa Martin, Danyu Shen,
Dmitrij Mordasov, Elea Seidlmann, Emilie Østerdal
Nilsson, Frey Emil Vestergaard, Grzegorz Wojciech
Zaba, Hans Christian Godballe Lundberg, Harshit
Shrivastava, Helene Scheel Wegener, Huanjun Liu,
Jakob Malte Skou Lindstad, Jens Peter Sparresø, Jiří
Tykva, Johan Kristian Petersen, Johannes Nørskov
Toke, Josep Marín Llaó, Kenneth Paulsen, Kevin
Thomas McCabe, Line Egerod Lund, Mads
Bundgaard Nørløv, Mads Cort Nielsen, Mads Dudzik
Møller, Maria Maniati, Mario Cesar Rodriguez,
Marion Motard, Mathilde Albrechtsen Mortensen,
Mikkel Ditlev Sjøgren Olsen, Mikkel Thestrup, Niels
Peter Lindegaard, Nils Rehtanz, Ole Martin Sørensen,
PANAGIOTIS PAPADAMOS, Peter Gustav Rilenter
Leunbach, Philip Tinggaard Thomsen, Rafael Parrado
Gorga, Saud Abdulaziz A Shaheen, Sebastian
Christian Harhoff Pieters, Signe Wulff-Andersen,
Tanja Sølvsten, Thomas Spyrou, Torben Truong
Nguyen, Txomin Perthuisot, Varun Shankar, Viktor
Wilhelm Johannes Nordström, William Frederik
Foldøy Steffens, Íris Björk Snorradóttir

², DTU Compute

Reading material:

Chapter 16, Chapter 17



Lecture 9 2 November, 2021

Lecture Schedule

1 Introduction

31 August: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

7 September: C2, C3

3 Measures of similarity, summary statistics and probabilities

14 September: C4, C5

4 Probability densities and data visualization

21 September: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

28 September: C8, C9

6 Overfitting, cross-validation and Nearest Neighbor

5 October: C10, C12 (Project 1 due before 13:00)

7 Performance evaluation, Bayes, and Naive Bayes

12 October: C11, C13

8 Artificial Neural Networks and Bias/Variance

26 October: C14, C15

9 AUC and ensemble methods

2 November: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 November: C18

11 Mixture models and density estimation

16 November: C19, C20 (Project 2 due before 13:00)

12 Association mining

23 November: C21

Recap

13 Recap and discussion of the exam

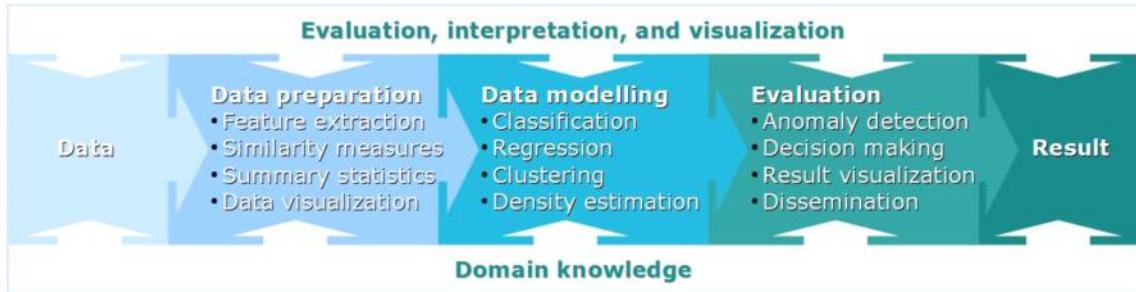
30 November: C1-C21

Online help: Forum on DTU Learn

Videos of lectures: <https://video.dtu.dk>

Streaming of lectures: Zoom (link on DTU Learn)

Lecture 9 2 November, 2021



Learning Objectives

- Explain the principle behind boosting and bagging and apply it to improve classifiers
- Be able to address issues of class-imbalance and resampling
- Understand the definition of Precision, Recall, ROC, and AUC

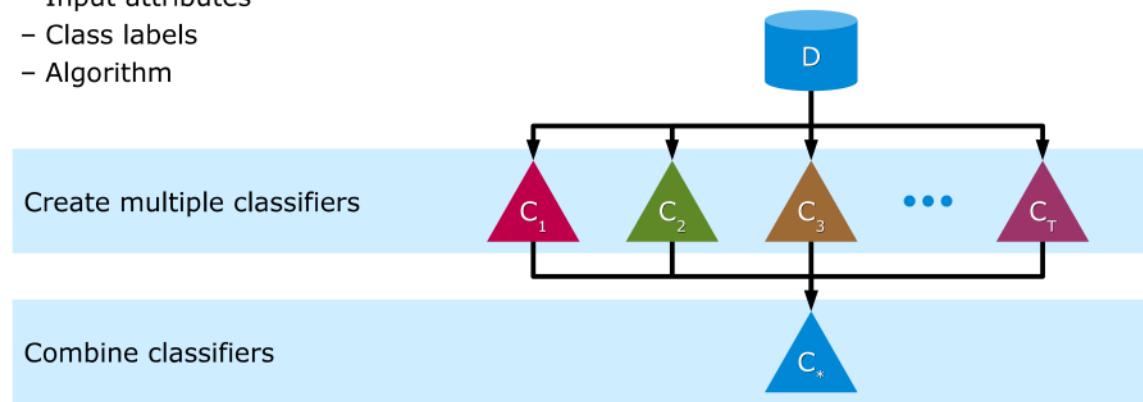
Combine multiple weak classifier (LR, decision tree) with a strong one

With ensemble you start from a simple model and make it more complex, is the opposite of regularization where tuning lambda we can obtain simpler classifier

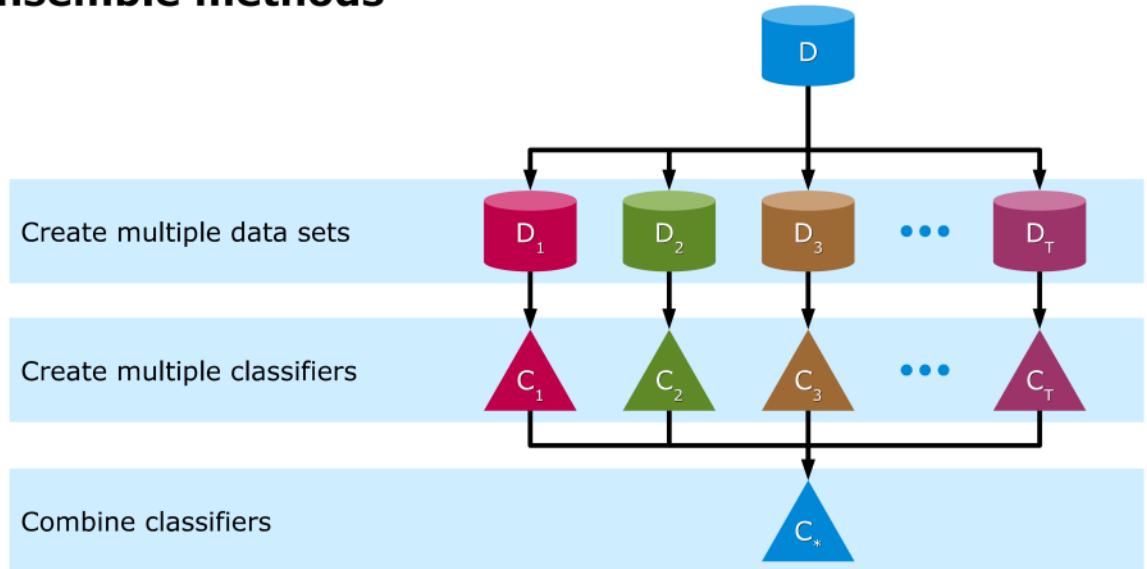


Ensemble methods

- Combine multiple (weak) classifiers into one (strong) classifier
- Each classifier trained using different variations of
 - Data set
 - Input attributes
 - Class labels
 - Algorithm



Ensemble methods



Why ensemble methods?

- Can improve classification algorithms in terms of
 - Better classification accuracy
 - Increased stability
 - Reduced variance
 - Less overfitting
- Consider T independent classifiers for binary classification, each with accuracy p .
- The probability a classifier which use majority voting is correct is then given by:
Majority voting is correct if at least half of the classification are correct

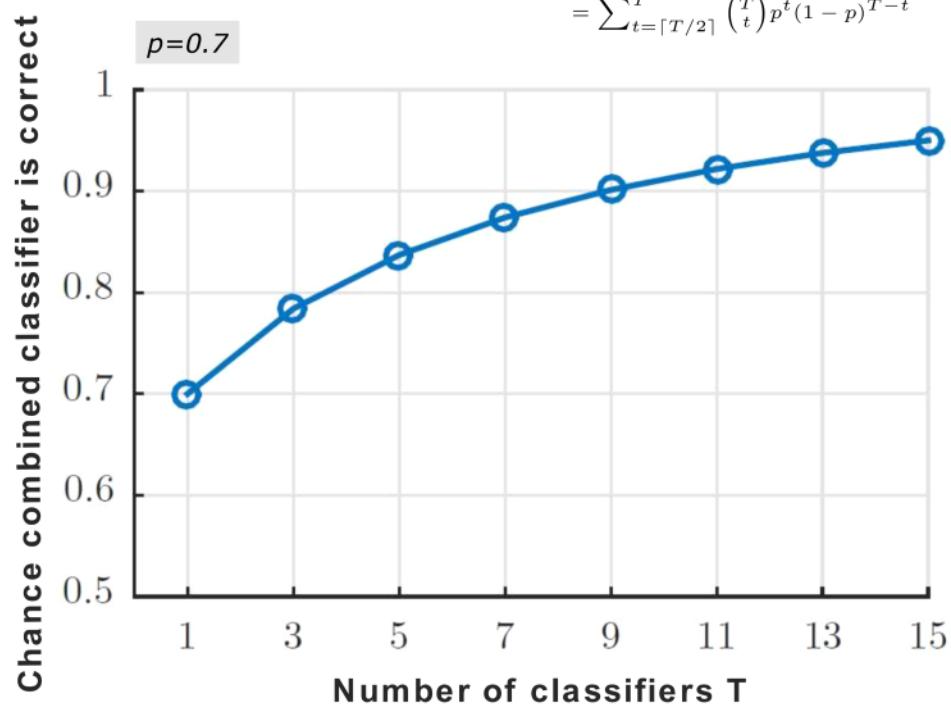
Look at why we get better classification accuracy

$$\begin{aligned} P(\text{Majority voting is correct}) &= \sum_{t=\lceil T/2 \rceil}^T P(\{t \text{ classifiers are correct}\}) \\ &= \sum_{t=\lceil T/2 \rceil}^T \binom{T}{t} p^t (1-p)^{T-t} \end{aligned}$$

Even if we have bad classifier, as long they are just above 50% of accuracy, we can get a better classifier combining the weak ones: ideally if we combine infinite number of classifiers, we get close to 1 of accuracy

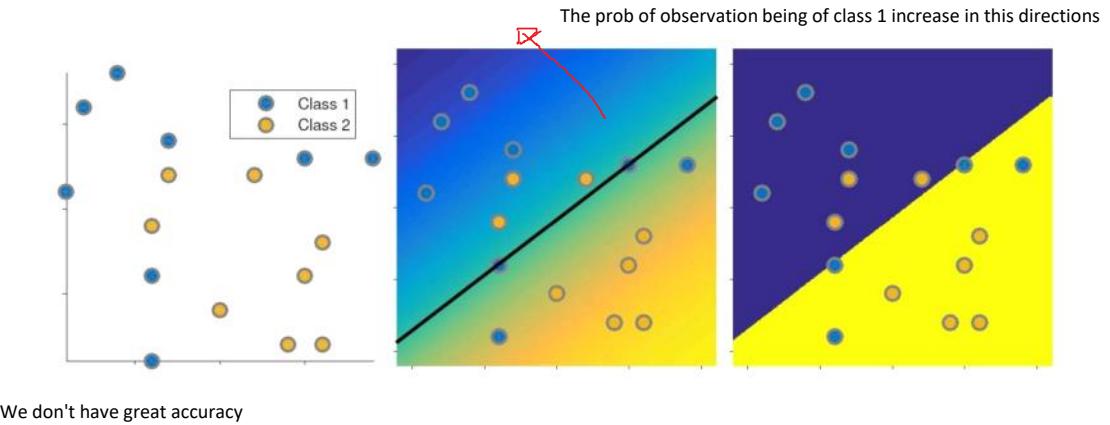


$$\begin{aligned} P(\text{Majority voting is correct}) &= \sum_{t=\lceil T/2 \rceil}^T P(\{t \text{ classifiers are correct}\}) \\ &= \sum_{t=\lceil T/2 \rceil}^T \binom{T}{t} p^t (1-p)^{T-t} \end{aligned}$$



Data example

- Classification using logistic regression



One way to create a new dataset is : randomly resample the original dataset to create random subsampling



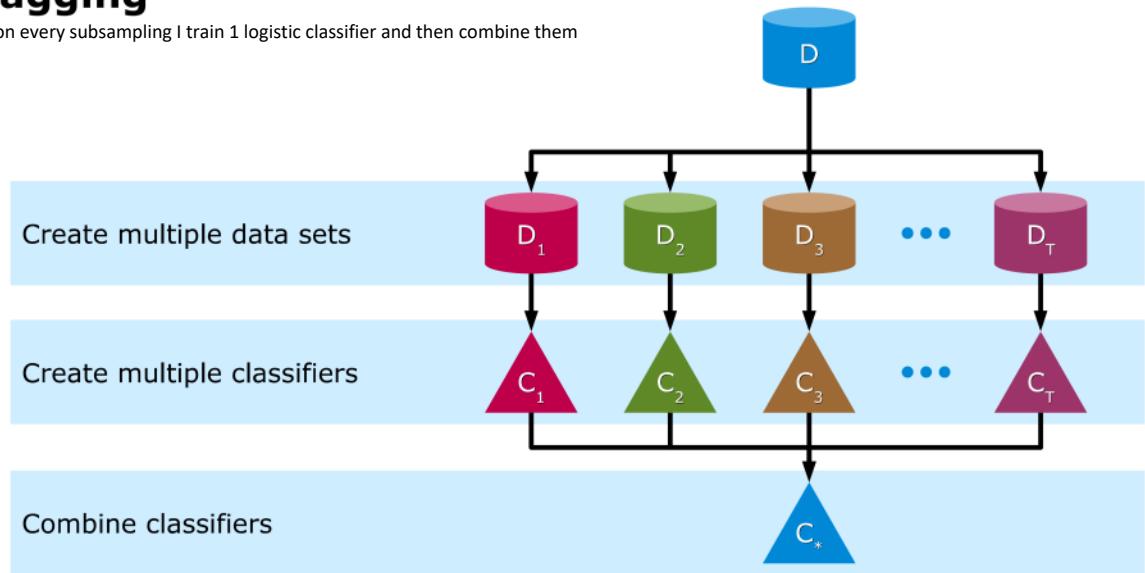
Bagging

- New training data sets drawn randomly from pool with replacement

Pool of training data	1	2	3	4	5	6	7	8	9	10
	3	5	4	3	9	7	9	5	1	1
	5	8	2	6	2	3	8	3	5	1
New training data sets	1	7	4	1	10	6	10	8	8	7
	4	3	8	5	2	4	7	10	10	8

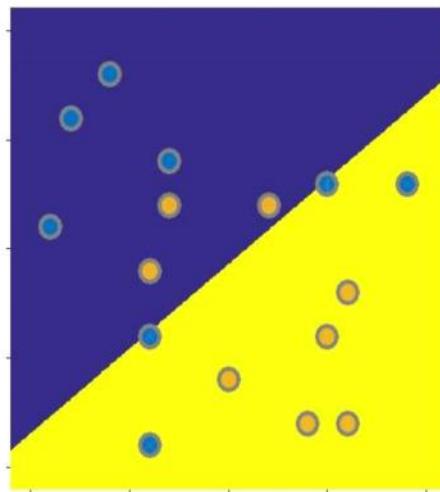
Bagging

So on every subsampling I train 1 logistic classifier and then combine them

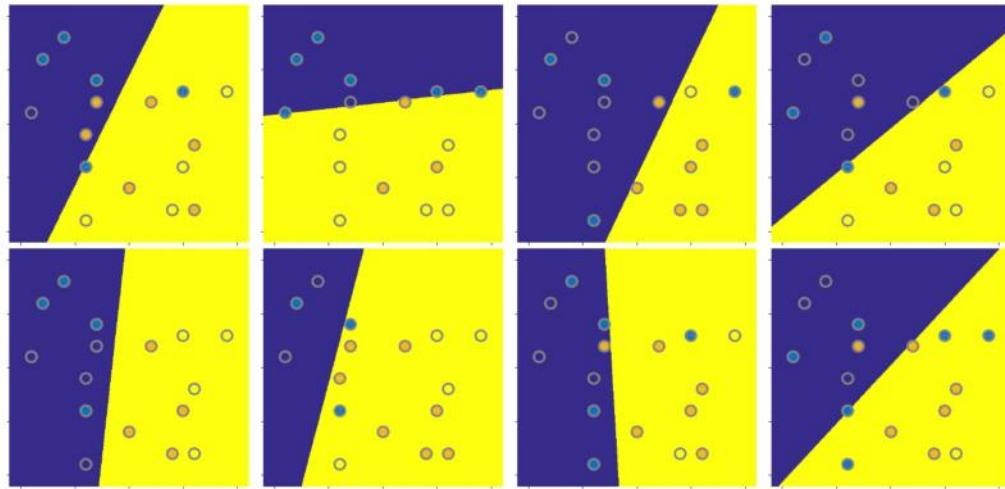


Bagging

- **Single classifier**
 - Logistic regression
 - Two features, (x, y)



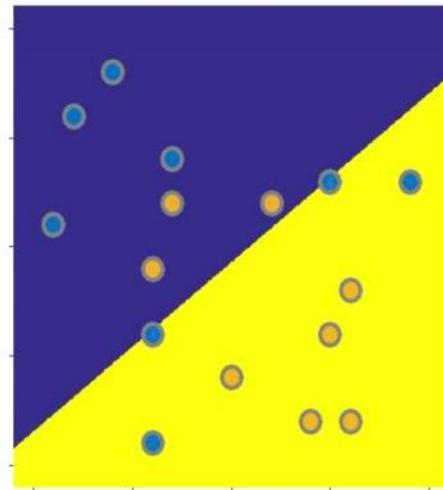
Bagging



Notice, hollow dots are observations not included in bagging round

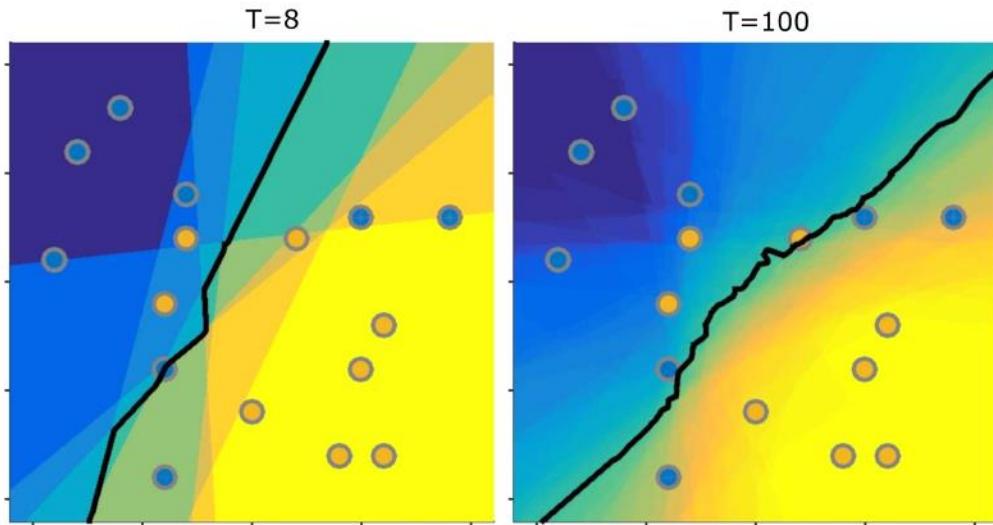
Bagging

- Single classifier



Bagging

Combine classifier: the black line is the new decision boundaries. Color correspond to the prob of an observation to be classified and disagreement



When we do BAGGING, when we combine a lot of linear model,
We get something that is not linear anymore: how to improve?? --> BOOSTING: instead of the
subsampling, I use the probability

Boosting

Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
New training data set	3	5	4	3	9	7	9	5	1	1
Train classifier									c_1	

Boosting

Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
New training data set	3	5	4	3	9	7	9	5	1	1
Train classifier	 C_1									
Classify all data objects	1✓	2✗	3✓	4✗	5✓	6✗	7✓	8✓	9✓	10✓
	wrong									

Boosting

Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
New training data set	3	5	4	3	9	7	9	5	1	1
Train classifier										
Classify all data objects	1✓	2✓	3✓	4✓	5✓	6✓	7✓	8✓	9✓	10✓
Update weights	.07	.17	.07	.17	.07	.17	.07	.07	.07	.07

When the classifier is correct, then decrease the probability (meaning that the point is probably an easy point to classify while we want to concentrate on the difficult one) and increases the miss-classified -> but the dataset should sum up to 1 (normalized) -> the idea is to look into more details at the difficult point for that specific classifier

Boosting

Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
New training data set	3	5	4	3	9	7	9	5	1	1
Train classifier									 C_1	
Classify all data objects	1✓	2✓	3✓	4✓	5✓	6✓	7✓	8✓	9✓	10✓
Update weights	.07	.17	.07	.17	.07	.17	.07	.07	.07	.07
New training data set	6	4	7	3	2	4	10	2	5	6
Train classifier									 C_2	
										⋮

1. We have some weights for the data point for the first classifier we train
2. We have T classifier
3. Create the first dataset, using regular sumsampling (equal weights)
4. Train
5. Calculate a weighted error: delta is going to be 1 when the prediction is correct (1 - delta) is just counting the miss-predicted. Divided by N i get the error rate
6. Alpha = it's a way to express the importance of the classifier: when the argument is close to 0, alpha goes to infinity: so if the error is low, we consider this classifier really important. In this way we don't take into account bad classifier. (alpha goes to -inf)



AdaBoost

Algorithm 6: AdaBoost algorithm

```

1: Initialize  $w_i(1) = \frac{1}{N}$  for  $i = 1, \dots, N$ 
2: for  $t = 1, \dots, T$  do
3:   Create  $\mathcal{D}_t$  by sampling (with replacement) from  $\mathcal{D}$  according to  $w(t)$ 
4:   Let  $f_t$  be the classifier trained on  $\mathcal{D}_t$ 
5:    $\epsilon_t = \sum_{i=1}^N w_i(t) (1 - \delta_{f_t(\mathbf{x}_i), y_i})$  (weighted error of  $f_t$  on all data)
6:    $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$ 
7:   For each  $i$  update weights using eq. (15.7):

```

$$w_i(t+1) = \frac{\tilde{w}_i(t+1)}{\sum_{j=1}^N \tilde{w}_j(t+1)}, \quad \tilde{w}_i(t+1) = \begin{cases} w_i(t)e^{-\alpha_t} & \text{if } f_t(\mathbf{x}_i) = y_i \\ w_i(t)e^{\alpha_t} & \text{if } f_t(\mathbf{x}_i) \neq y_i. \end{cases}$$

```

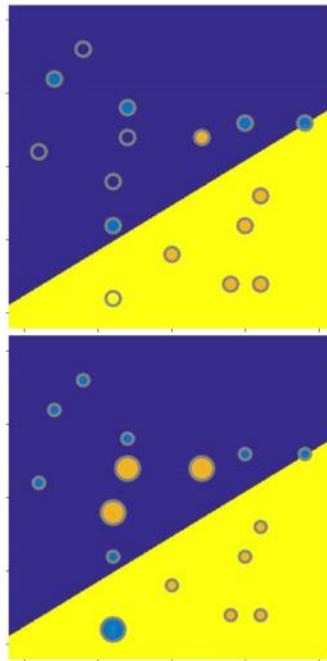
8: end for
9:  $f^*(\mathbf{x}) = \arg \max_{y=1,2} \sum_{t=1}^T \alpha_t \delta_{f_t(\mathbf{x}), y}$  (Majority voting classifier)

```

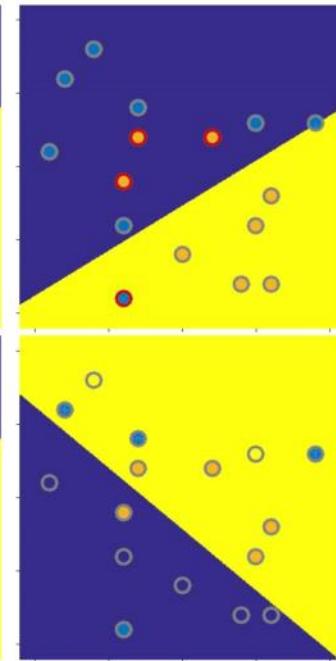
7. Update weight: if it a good classifier, we are going to put less weight on the easy point(good classified) and increase the weight in the miss-predicted ones -> give more importance to the wrong classification (difficult points)
9. We combine these classifier with Majority Voting: if we just assume that alpha = 1, then is just majority voting (how many times each classifier is 1. But with alpha is the importance, so if a classifier has low error rate, then alpha has big alpha so that 1 is more important. If a classifier that say opposite thing,(80% of the times is wrong and 20 % correct, than make sense to invert the classifier) and this is possible using alpha negative

Boosting

A:
A dataset is sampled with replacement and a classifier trained.



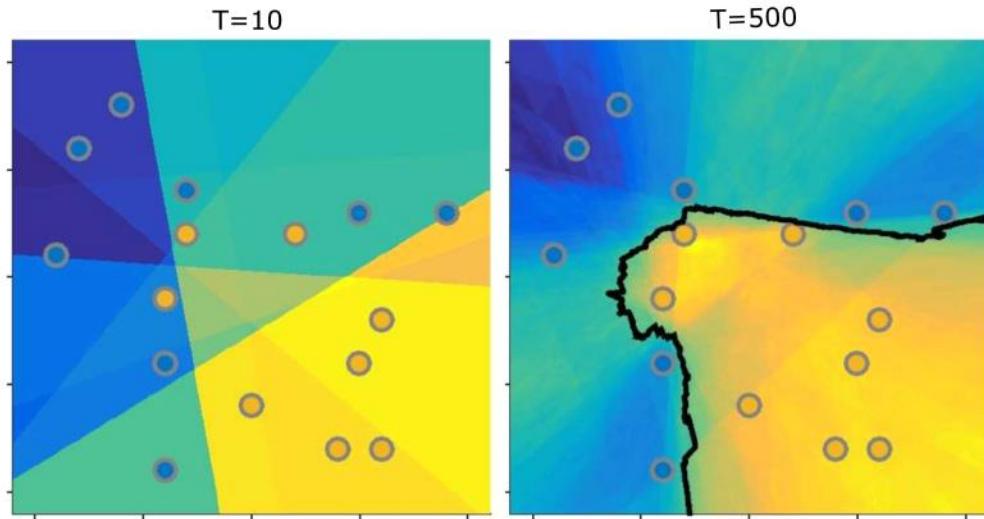
C:
Weights are updated such that more emphasis is given to these misclassified observations.



B:
Mis-classified observations are identified.

New round:
Based on the updated weights a new dataset is sampled and a classifier trained (shown), mis-classified observations identified and given more emphasis...

Boosting



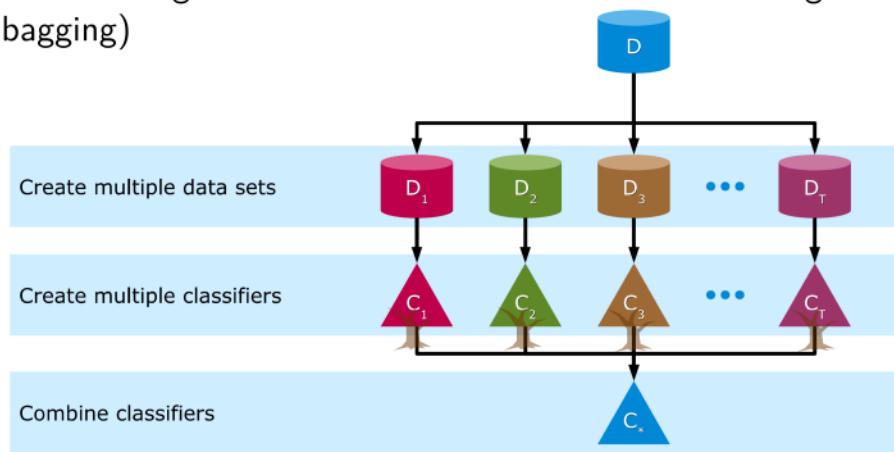
Bagging example: Random forest

If we do Bagging for Decision Tree we get Random forest.

Each tree is generated as follows:

- Sample dataset with replacement
- When generating each node in the tree, randomly select a subset of the features and only consider splits using these features

A large number of trees are generated and the trees are combined using majority voting (bagging)



Quiz 1 (please answer on Piazza): Adaboost (Spring 2016)

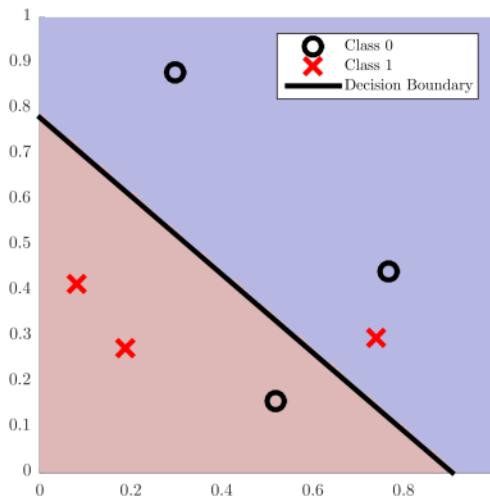


Figure 1: A binary classification problem and the decision boundary obtained by logistic regression. Observations left of the boundary are classified as belonging to the positive class 1 (red crosses) and observations right of the boundary to the negative class 0 (black circles)

We wish to apply a logistic regression model to the binary classification problem shown in Figure 1. We attempt to improve the performance by applying AdaBoost. AdaBoost works by first sampling a new dataset with replacement, then training a classifier on the dataset and then proceeding with the subsequent steps of the AdaBoost algorithm.

Suppose in the first iteration of the AdaBoost algorithm the classification boundary of the trained classifier is as indicated by the black line (i.e. observations left of the black line are classified as in the positive class). What is the resulting value of the weights \mathbf{w} ?

- A. $\mathbf{w} = [0.125 \ 0.250 \ 0.125 \ 0.125 \ 0.125 \ 0.250]$
- B. $\mathbf{w} = [0.026 \ 0.447 \ 0.026 \ 0.026 \ 0.026 \ 0.447]$
- C. $\mathbf{w} = [0.235 \ 0.029 \ 0.235 \ 0.235 \ 0.235 \ 0.029]$
- D. $\mathbf{w} = [0.1 \ 0.3 \ 0.1 \ 0.1 \ 0.1 \ 0.3]$
- E. Don't know.

(Hint: First compute ε_1 , then α_1 , then the weights)

24 DTU Compute

$W1 = [1/6 \ 1/6 \ 1/6 \ 1/6 \ 1/6 \ 1/6]$

There are two point that are mis-classified -> $(1-\delta) = 1$ because $\delta = 0$

Eps = 1/3 (error rate) in the first loop.

Alpha = $1/2 * \log(1-eps / eps) = 0,15$

Lecture 9 2 November, 2021

Class imbalance problem

- Many data sets have **imbalanced class distributions**
 - Example: Detection of defects that only occur rarely (e.g. 1/1,000,000)
 - Danger: Algorithm that says nothing is defect will be 99.999% correct
- **Solution approaches**
 - Resample to balance data sets
 - Modify existing classification algorithms
 - Measure performance in a way that takes balance into account

Resampling balanced data

- New sample has equal number of data objects from each class
- **Approaches**
 - **Undersampling** majority class: Throws out potentially useful data
 - **Oversampling** minority class: Increase data size and computational burden
 - **Somewhere in between...**

Imbalanced training data	1	2	3	4	5	6	7	8	9	10
Oversampling	1	2	3	4	5	7	9	10	6	6
	6	6	8	8	8	8				
Undersampling	3	5	6	8						
Somewhere in between	3	5	4	3	9	6	6	8	8	8

We usually do something in between, using a dataset of the same size of the original one but throwing some of the observations for balancing

Confusion matrix

		<i>Predicted</i>	
		<i>positive</i>	<i>negative</i>
<i>Actual</i>	<i>positive</i>	TP True Positive	FN False Negative
	<i>negative</i>	FP False Positive	TN True Negative

Precision and recall

High precision means that we can trust the positive predictions, so something that has been predicted

- **Precision** as positive is highly probable that is really positive

- Fraction of true positive among objects predicted to be positive

$$p = \frac{TP}{TP + FP}$$

- **Recall**

- Fraction of objects predicted to be positive among all positive objects

$$r = \frac{TP}{TP + FN}$$

		<i>Predicted</i>	
		<i>positive</i>	<i>negative</i>
<i>Actual</i>	<i>positive</i>	TP	FN
	<i>negative</i>	True Positive	False Negative
	<i>negative</i>	FP	TN
	<i>positive</i>	False Positive	True Negative



Quiz 2 (please answer on Piazza): Precision/Recall

Consider two different classifiers, and suppose on a test set with 20 positive observations:

- Classifier 1 detects 54 positive of which 18 are actually positive
- Classifier 2 detects 16 positive of which 14 are actually positive

		Predicted	
		positive	negative
Actual	positive	TP True Positive	FN False Negative
	negative	FP False Positive	TN True Negative

What is the precision and recall of the two classifiers?

- A. Classifier 1: $p_1 = \frac{2}{3}, r_1 = \frac{7}{10}$
Classifier 2: $p_2 = \frac{1}{3}, r_2 = \frac{1}{3}$
- B. Classifier 1: $p_1 = \frac{1}{3}, r_1 = \frac{9}{10}$
Classifier 2: $p_2 = \frac{2}{3}, r_2 = \frac{9}{10}$
- C. Classifier 1: $p_1 = \frac{2}{3}, r_1 = \frac{7}{10}$
Classifier 2: $p_2 = \frac{1}{3}, r_2 = \frac{9}{10}$
- D. Classifier 1: $p_1 = \frac{1}{3}, r_1 = \frac{9}{10}$
Classifier 2: $p_2 = \frac{7}{8}, r_2 = \frac{7}{10}$



Which classifier would you use if the objective was to detect credit-card fraud (the positive class corresponds to fraud)

Look at good is the recall: but high recall will produce high False Positive
Then low precision

• Precision

- Fraction of true positive among objects predicted to be positive

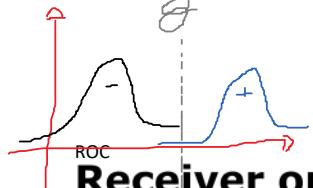
$$p = \frac{TP}{TP + FP}$$

• Recall

- Fraction of objects predicted to be positive among all positive objects

$$r = \frac{TP}{TP + FN}$$

Assume we have a classifier : $y_{\text{hat}} = 1$ with $p(y=1|x) > \theta$ and $y_{\text{hat}} = 0$ with $p(y=0|x) < \theta$

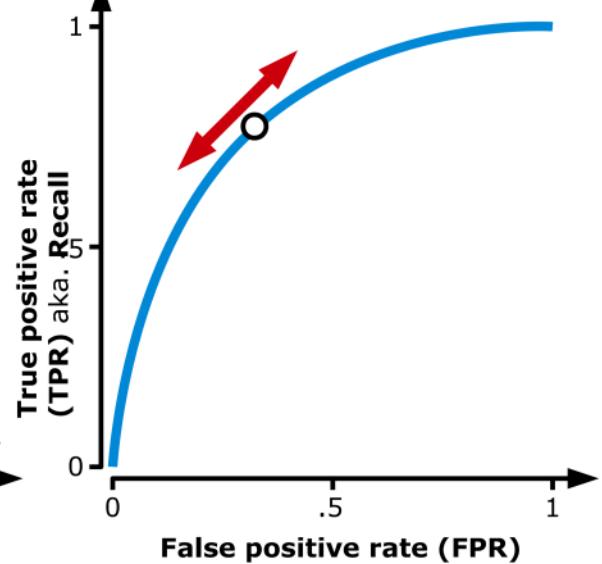
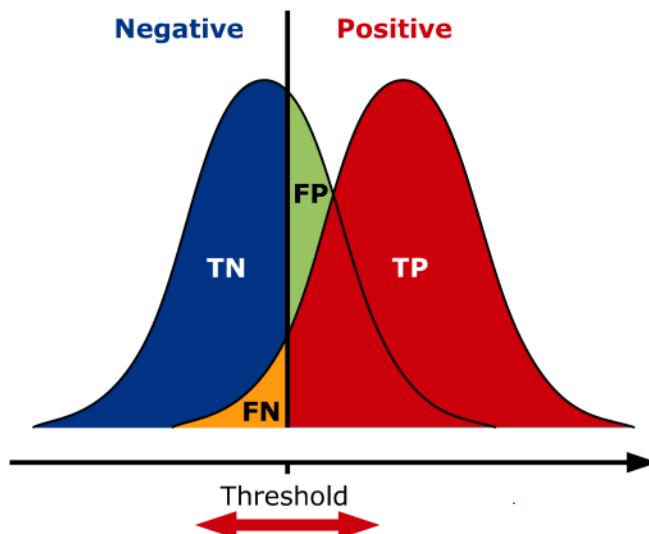


But in reality there is always overlap between the two plots -> everything left to the line will be predicted as negative and anything to the right as positive, so i create some FN and FP respectively.
I can tune the threshold to play between recall and precision -> ROC : curve TPR/FPR



Receiver operating characteristic

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

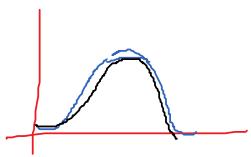


False positive rate (FPR)

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

Lecture 9 November, 2021

The worstcase is when I have two overlapping curve



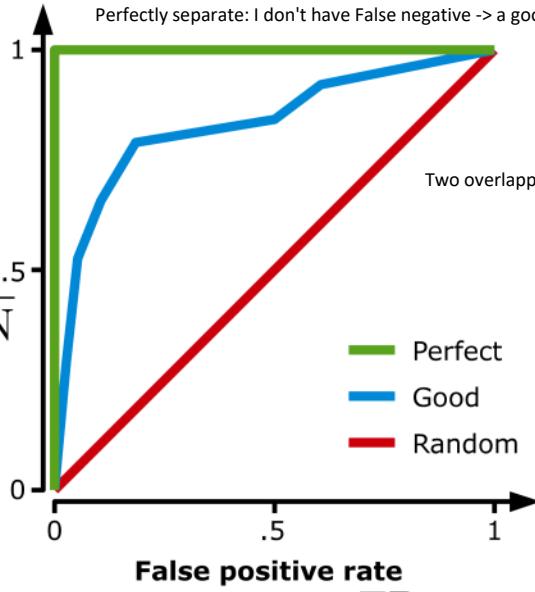
Receiver operating characteristic

True positive rate
aka. **Recall**

$$TPR = \frac{TP}{TP + FN}$$

Perfectly separate: I don't have False negative \rightarrow a good classifier will have an area below the curve
Between 1 and 0 (area below perfect = 1
And area below the random = 1/2).

Two overlapping curve \rightarrow random classifier



$y_{\text{hat}} = 1$ if $f(i) > \text{threshold}$
 $y_{\text{hat}} = 0$ otherwise

Let's take a threshold theta = 4,5 -> only 5 gear belongs to positive class

$$\begin{aligned} \text{FPR} &= 2 (\text{FP}) / 15 (\text{TN}) + 2 (\text{FP}) = 2/17 \\ \text{TPR} &= 3 / 3 + 13 = 8/15 \end{aligned}$$

Take another thresh :

Quiz 3 (please answer on Piazza): AUC (Spring 2017)



	3 gears ($x_5 = 3$)	4 gears ($x_5 = 4$)	5 gears ($x_5 = 5$)
Low mpg ($y = 0$)	13	2	2
High mpg ($y = 1$)	2	10	3

Table 1: Number of low mpg and high mpg cars (i.e. $y = 0$ and $y = 1$) according to the number of gears, i.e. $x_5 = 3$, $x_5 = 4$, or $x_5 = 5$.

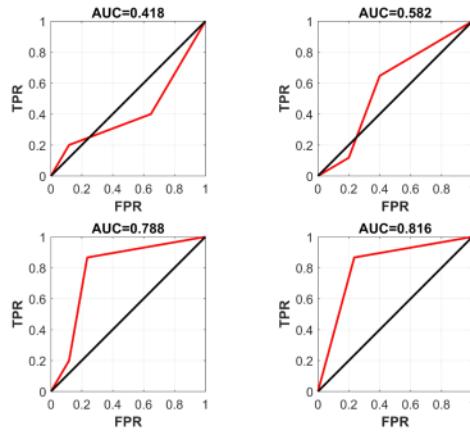


Figure 1: Four different receiver operator characteristic (ROC) curves and their area under curve (AUC) value.

A dataset representing cars contain an attribute x_5 corresponding to the number of gears. We wish to evaluate how well the number of gears predict low mpg, ($y = 0$, considered the negative class) from high mpg, ($y = 1$, considered the positive class) based on the data given in Table 1. For this purpose, we will evaluate the area under curve (AUC) of the receiver operator characteristic (ROC) using the feature x_5 . Which one of the ROC curves given in Figure 1 corresponds to using x_5 to discriminate between low mpg ($y = 0$) and high mpg ($y = 1$)?

- A. The curve having AUC=0.418
- B. The curve having AUC=0.582
- C. The curve having AUC=0.788
- D. The curve having AUC=0.816
- E. Don't know.

(Hint: Select a value e.g. $x_5 = 4$. We then predict cars with 4 or more gears as being in the positive class and otherwise negative. Compute the FPR and TPR using this prediction and use the (FPR, TPR) values to discriminate between the curves)

Resources

<https://www.youtube.com> Video tutorial on ROC curve and AUC

(<https://www.youtube.com/watch?v=0A16eAyP-yo>)

<https://towardsdatascience.com> More in-depth discussion of the Random Forrest algorithm and parameter choices

(<https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>)

<https://www.datacamp.com> Practical use of the random forest algorithm in python

(<https://www.datacamp.com/community/tutorials/random-forests-classifier-python>)

<https://citeseerx.ist.psu.edu> Justification for the AdaBoost algorithm (technical) (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.9525>)

Lecture 10

9. november 2021 12:36



02450: Introduction to Machine Learning and Data Mining

K-means and hierarchical clustering

Jes Frellsen

DTU Compute, Technical University of Denmark (DTU)



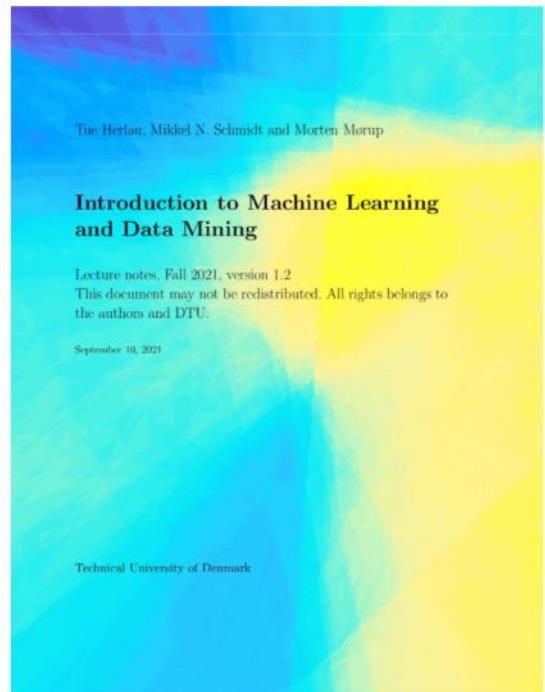
Today

Feedback Groups of the day:

Amaya Zaratiegui Pedrosa, Andreas Østerby Holst Rasmussen, Ane Carina Reiter, Anna Oliver Almirall, Bence Santha, Burak Özdemir, Camilla Becker Madsen, Carl Anton Schmidt, Chongchong Li, Christian Raasteen, Clara Mejlhede Lorenzen, Dagh Mikael Einer Nielsen, Daniel Schober, Daniel Scholz, David Bro Ludvigsen, Dominika Dora Melczer, Dung My Thi Trinh, Emil Sejer Pedersen, Emmanuel Mbecho Techago, Frithjof Prochnow Sletten, Hailin Liu, Hans Magnus Utne, Iñigo Velasco Ubillos, Jacob Østergaard, Jakob Tore Kammeyer Nielsen, Johan Nino Verrecchia, Jose Turegano Perez, Josephine Tanyous Rosenqvist, Julius Holbech Radzikowski, Kathe Hedegaard Schmidt, Lasse Bach Schrøder-Sevald, Liv olsen, Magnus Andreas Nielsen, Maksym Kryshchalov, Marah Osama marak, Marike Ikjær Weisbjerg, Martin Nguyen, Mathias Bonde Sørensen, Mathias Dal Møller, Mathias Krarup Lauridsen, Michel Qu, Mikkel Niklas Rasmussen, Morten Holmark Vandborg, Niels Georg Vendelø, Pablo Táboas Rivas, Peter Dalsgaard Nicolaisen, Rasmus Kjær Mortensen, Rasmus Thorsøe Molsted, Ruth Olerud, Simon Michael Widmer, Stine Lund Madsen, Syuan Yu Wang, Tamas Rudokasz, Troels Qvistgaard Ludwig, Victor Tadeusz Ulstrup Olszowski, Viktor Ryle Tamstorf, Vimal Velusamy Bharathi, Xinyu Liu, Yongzhi Song, Zhidie Wu

2 DTU Compute

Reading material: Chapter 18



Lecture 10 9 November, 2021

Lecture Schedule

1 Introduction

31 August: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

7 September: C2, C3

3 Measures of similarity, summary statistics and probabilities

14 September: C4, C5

4 Probability densities and data visualization

21 September: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

28 September: C8, C9

6 Overfitting, cross-validation and Nearest Neighbor

5 October: C10, C12 (Project 1 due before 13:00)

7 Performance evaluation, Bayes, and Naive Bayes

12 October: C11, C13

8 Artificial Neural Networks and Bias/Variance

26 October: C14, C15

9 AUC and ensemble methods

2 November: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 November: C18

11 Mixture models and density estimation

16 November: C19, C20 (Project 2 due before 13:00)

12 Association mining

23 November: C21

Recap

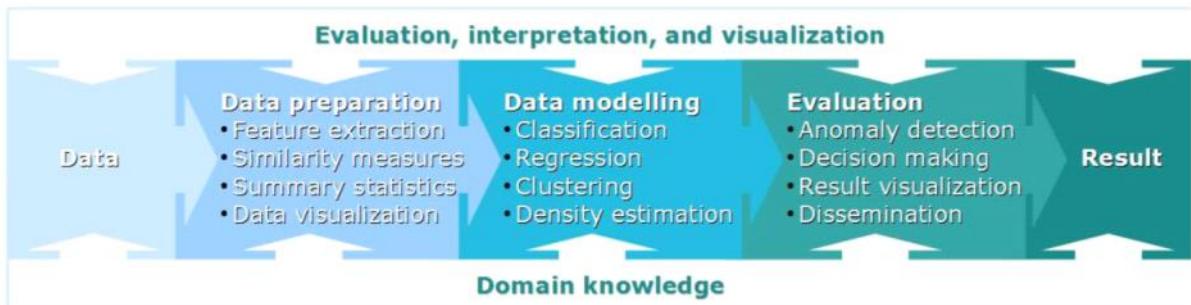
13 Recap and discussion of the exam

30 November: C1-C21

Online help: Forum on DTU Learn
 Videos of lectures: <https://video.dtu.dk>
 Streaming of lectures: Zoom (link on DTU Learn)

3 DTU Compute

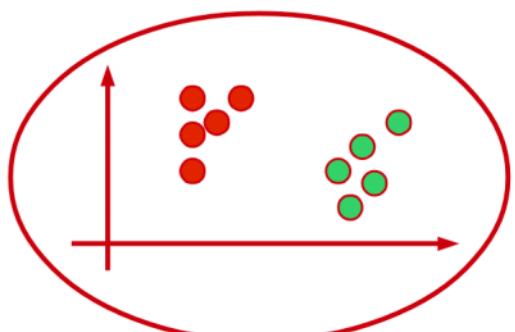
Lecture 10 9 November, 2021



Learning Objectives

- Understand the principles behind K-means and hierarchical clustering
- Understand how different linkage functions affects clustering types
- Evaluate clustering quality using class label information

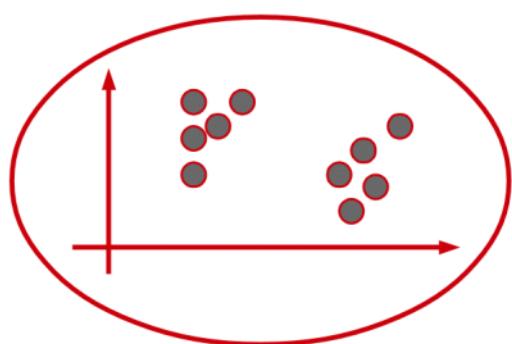
Supervised and Unsupervised learning



Supervised Learning

Input data x_n and output y_n

(Classification and Regression)



Unsupervised Learning

Input data x_n alone

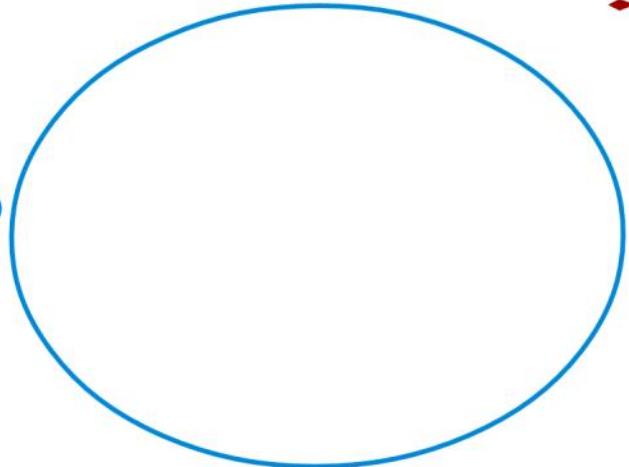
(Exploratory analysis)



Imagine you observe the world for the first time!



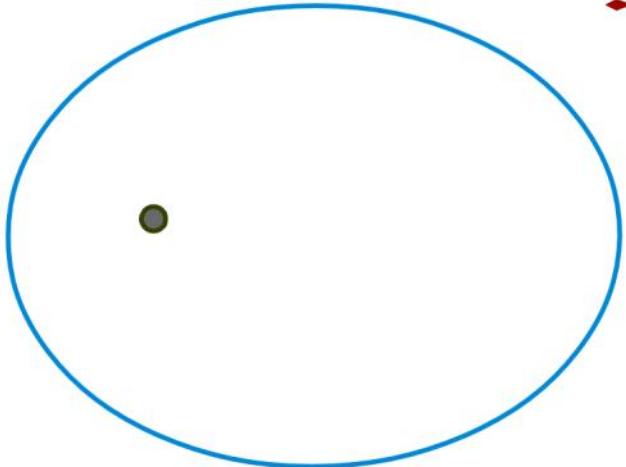
<http://www.clipartlord.com/category/baby-clip-art/>



Imagine you observe the world for the first time!



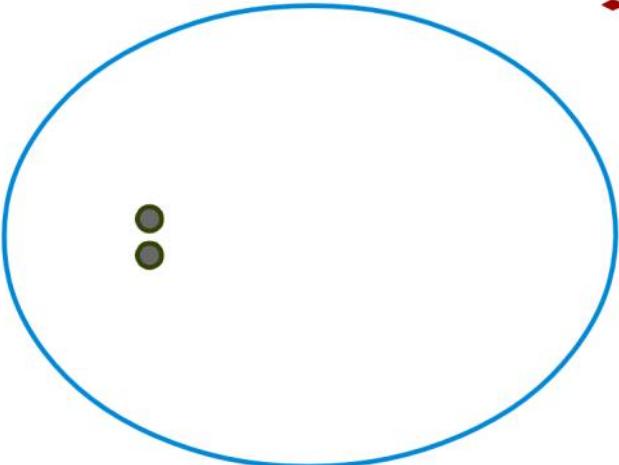
<http://www.clipartlord.com/category/baby-clip-art/>



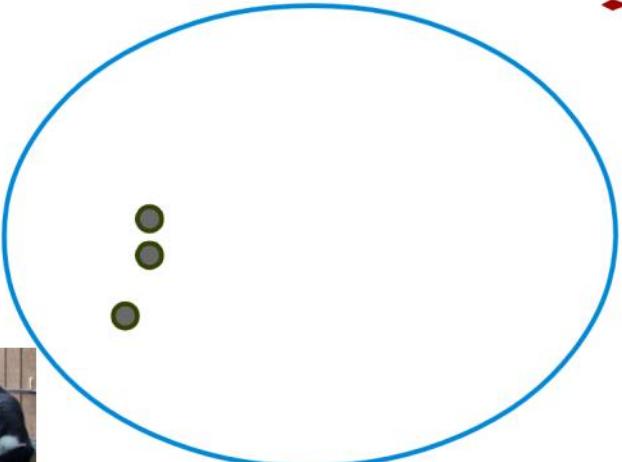
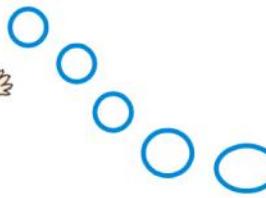
Imagine you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>



Imagine you observe the world for the first time!



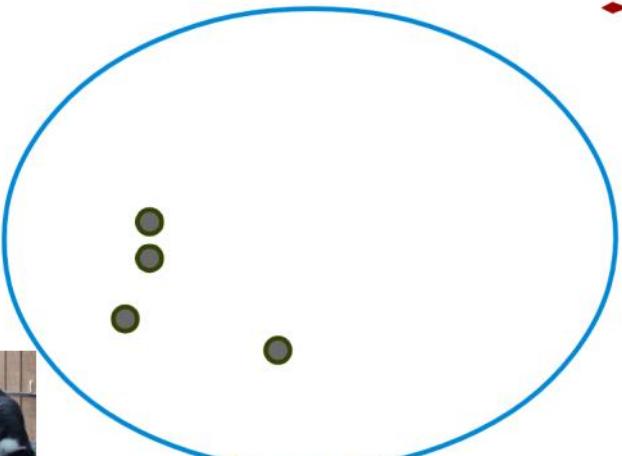
<http://www.clipartlord.com/category/baby-clip-art/>



Imagine you observe the world for the first time!



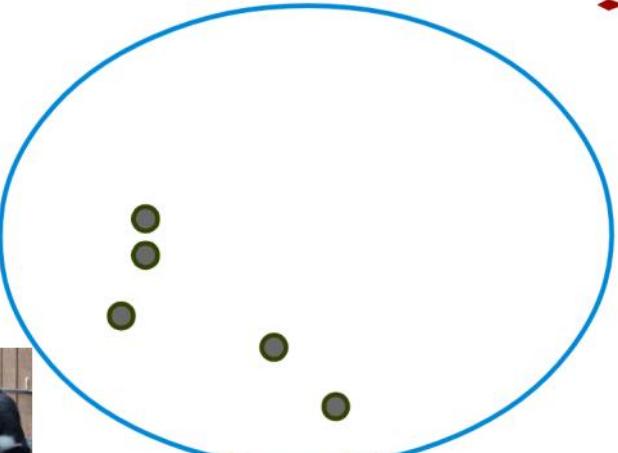
<http://www.clipartlord.com/category/baby-clip-art/>



Imagine you observe the world for the first time!



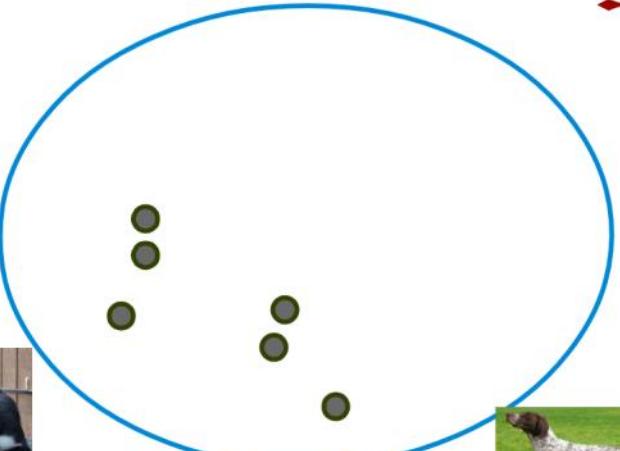
<http://www.clipartlord.com/category/baby-clip-art/>



Imagine you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>



12 DTU Compute

Lecture 10 9 November, 2021

Imagine you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>



Imagine you observe the world for the first time!



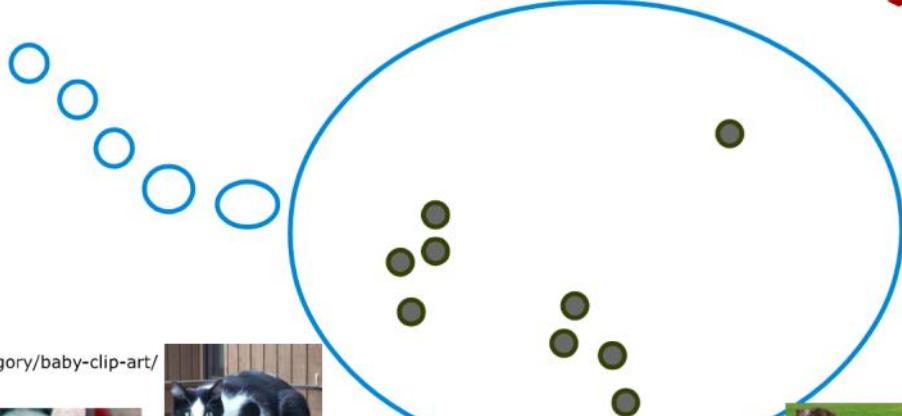
<http://www.clipartlord.com/category/baby-clip-art/>



Imagine you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>



Imagine you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>



We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?

16 DTU Compute

http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:MasPiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<https://commons.wikimedia.org/w/index.php?title=File:BluefacedBeechwoodnd.jpg&oldid=59100077>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>

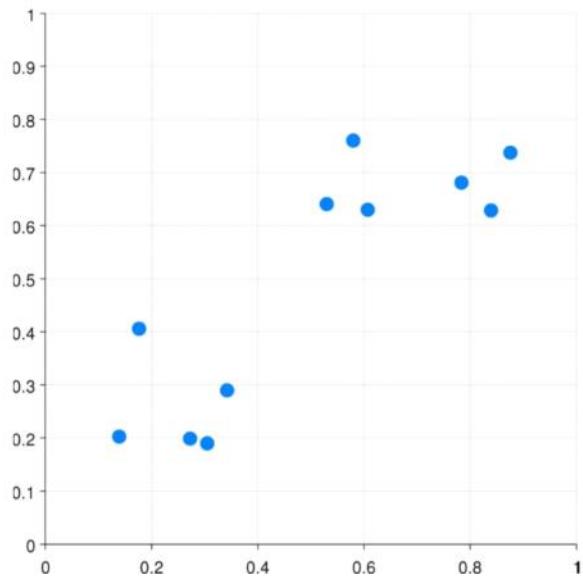
Unsupervised learning

- **Supervised learning**
 - Use the data to learn the output values
- **Unsupervised learning**
 - No output variables available
 - Sometimes called exploratory analysis
 - What to learn from the data?
 - Structure
 - Regularities
 - Hidden information
 - Etc.

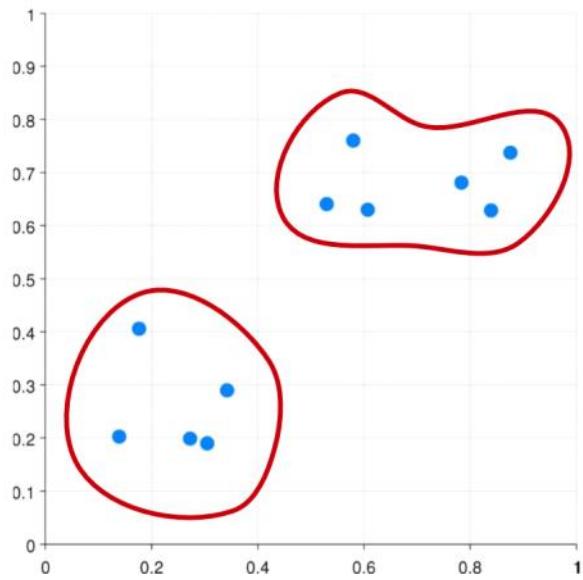
Clustering

- Divide data into groups (subsets/clusters) that are
 - **Meaningful:** Capture the natural structure of the data
 - **Useful:** Depends on purpose
- Observations in the same cluster are **similar in some sense**
- Unsupervised classification

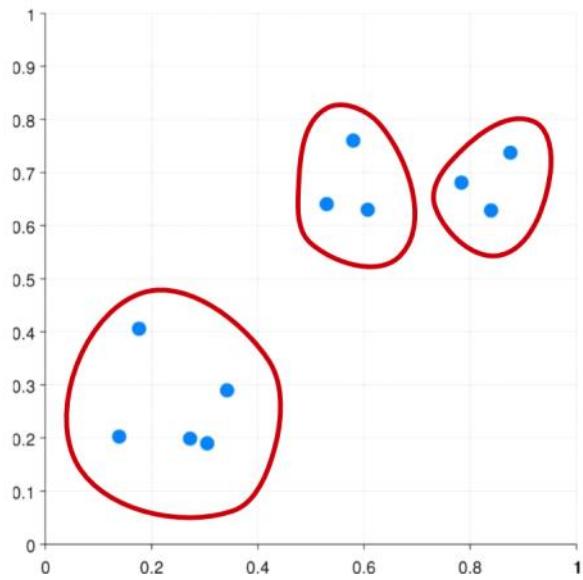
Clustering



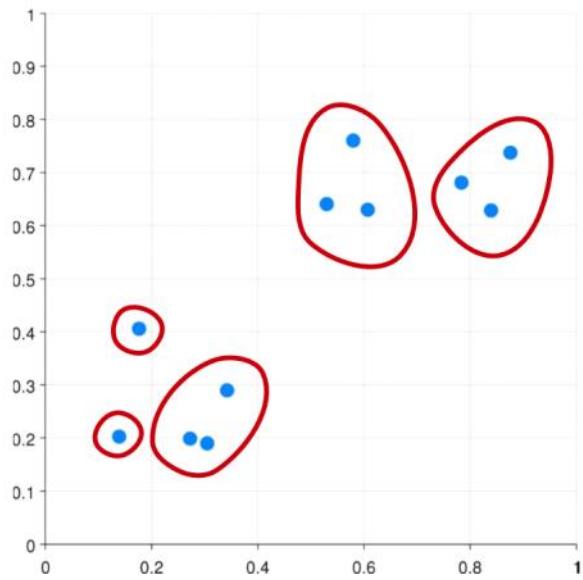
Clustering



Clustering

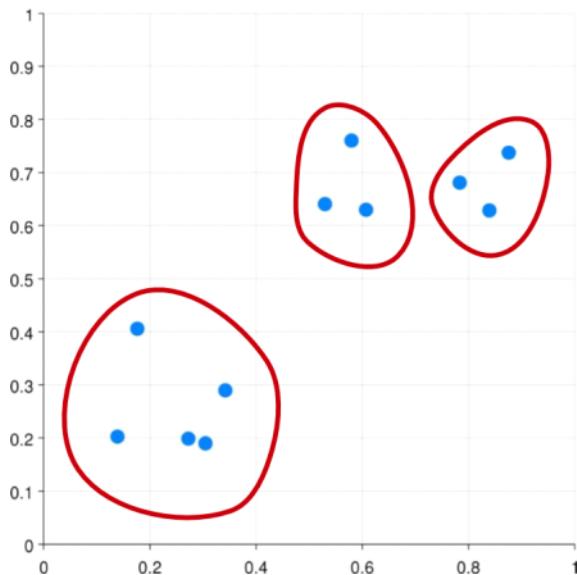


Clustering



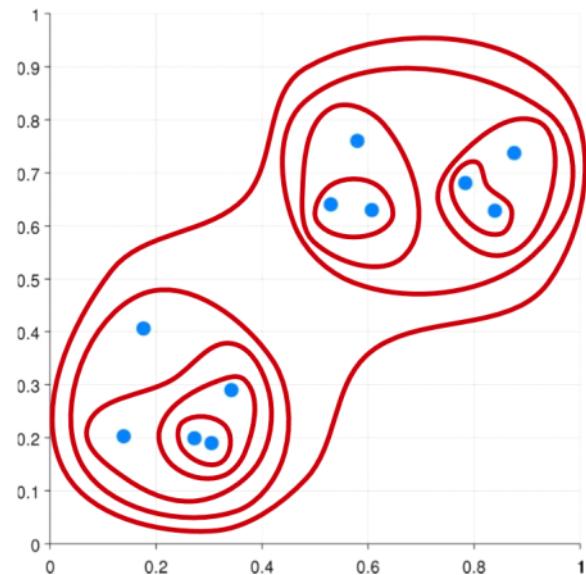
Partitional / hierarchical clustering

Partitional

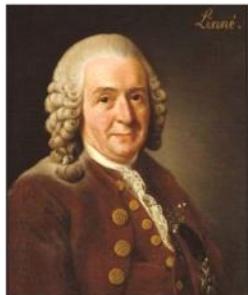


Hierarchical

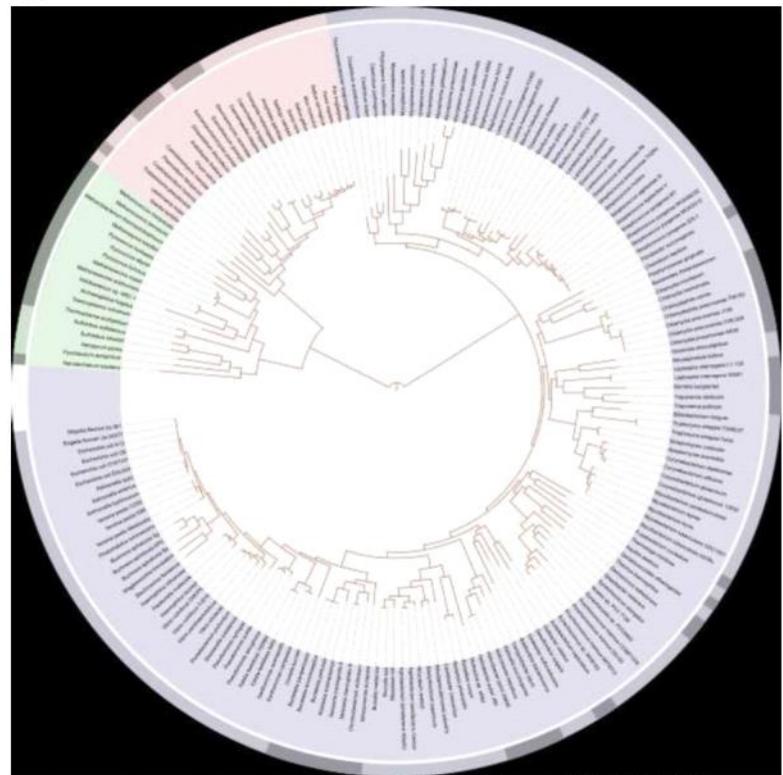
Nested partitioning of the data



Phylogenetic trees may be considered a type of hierarchical clustering



Carl Linnaeus
(1707 – 1778)
http://en.wikipedia.org/wiki/Carl_Linnaeus

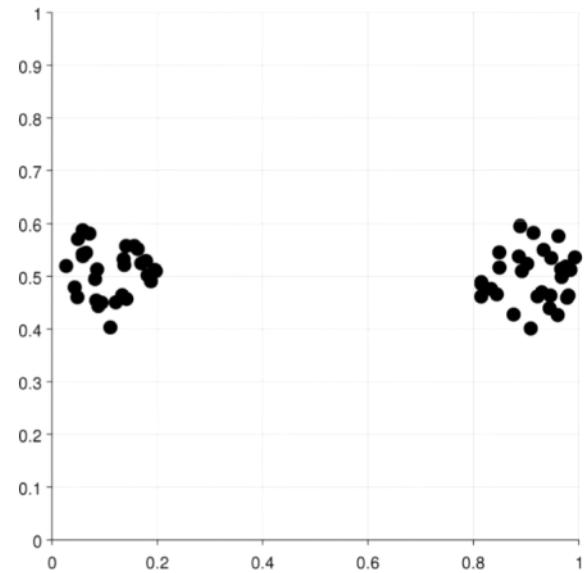


lecture 10 - 9 November 2021
http://en.wikipedia.org/wiki/File:Tree_of_life_SVG.svg

Types of clustering

Well-separated

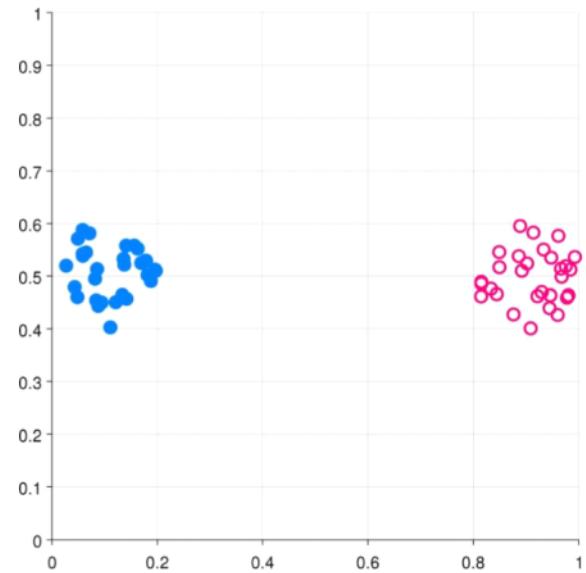
- Each point is closer to all points in its cluster than any point in another cluster



Types of clustering

Well-separated

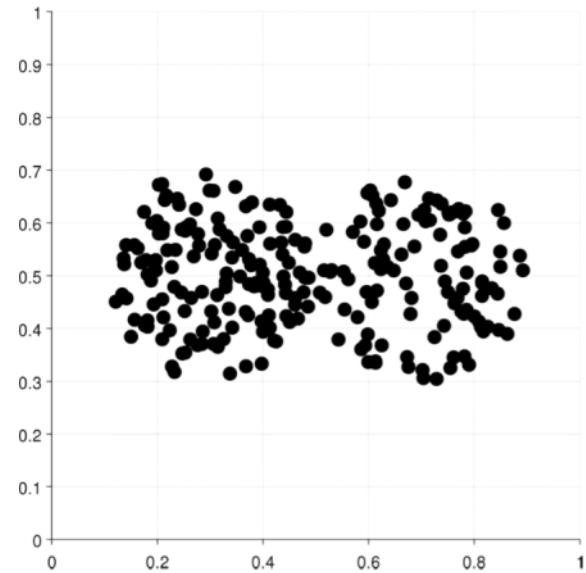
- Each point is closer to all points in its cluster than any point in another cluster



Types of clustering

Center-based

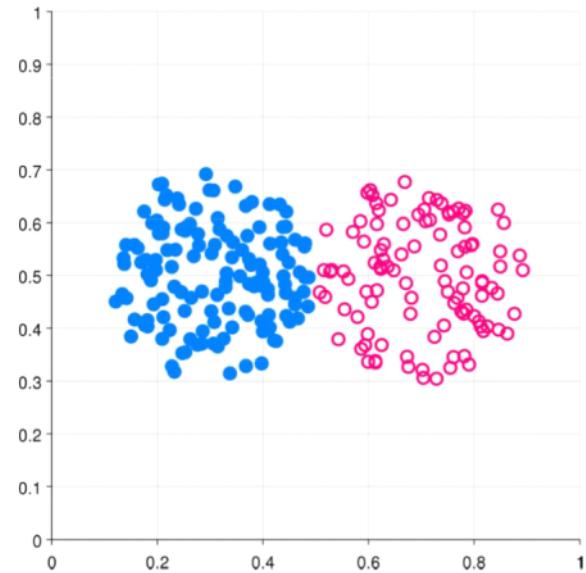
- Each point is closer to the center of its cluster than to the center of any other cluster



Types of clustering

Center-based

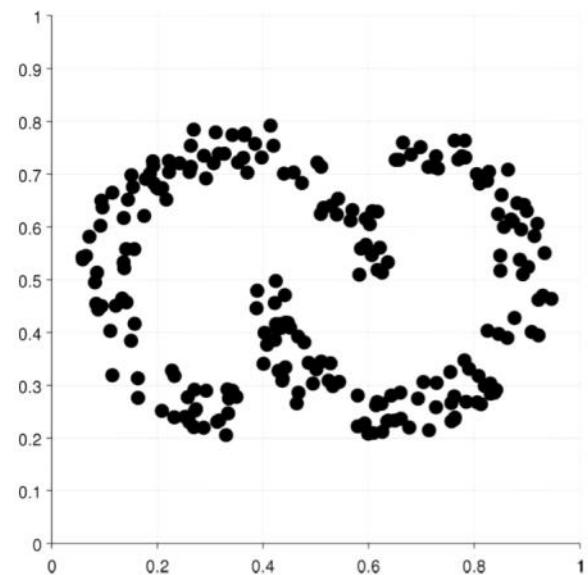
- Each point is closer to the center of its cluster than to the center of any other cluster



Types of clustering

Contiguity-based

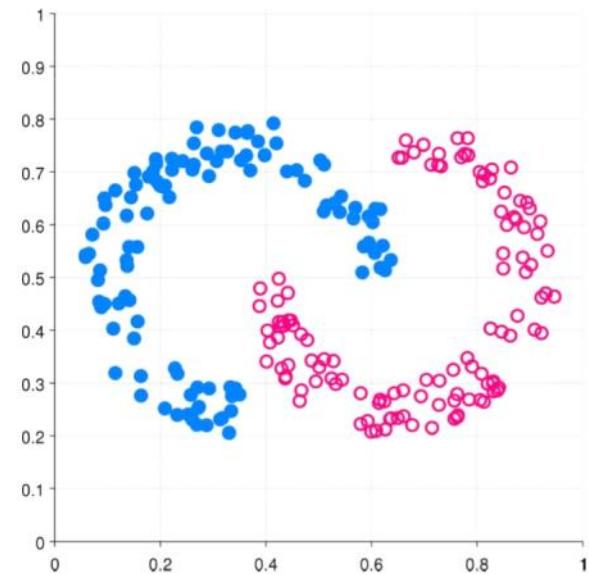
- Each point is closer to at least one point in its cluster than to any point in another cluster



Types of clustering

Contiguity-based

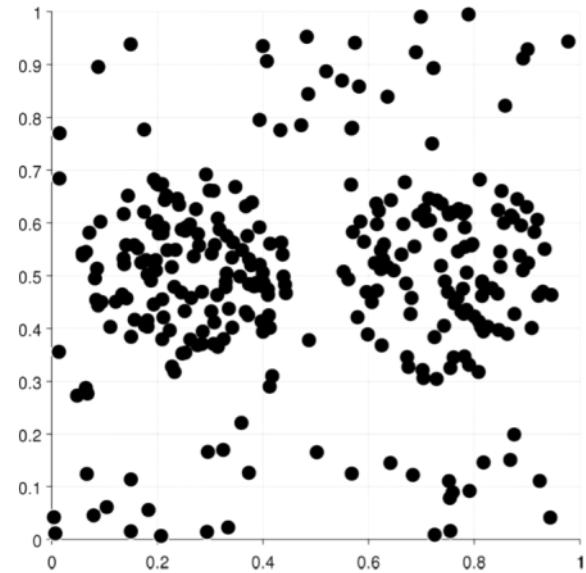
- Each point is closer to at least one point in its cluster than to any point in another cluster



Types of clustering

Density-based

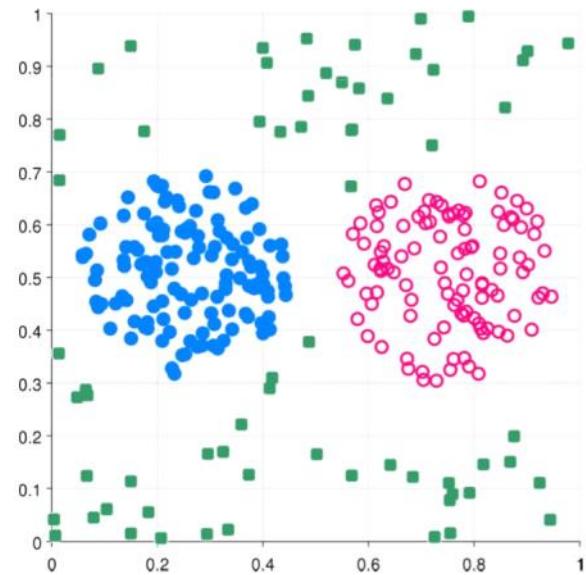
- Clusters are regions of high density separated by regions of low density



Types of clustering

Density-based

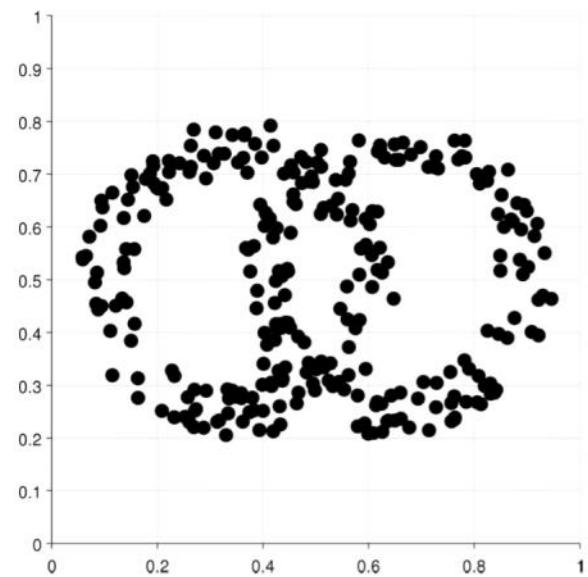
- Clusters are regions of high density separated by regions of low density



Types of clustering

Conceptual clusters

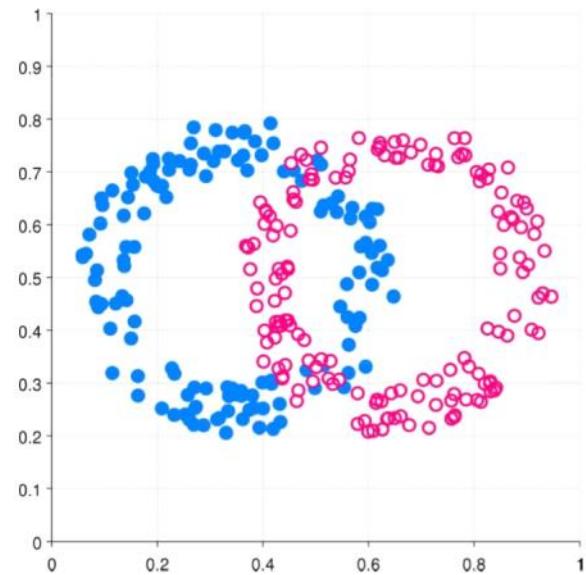
- Points in a cluster share some general property that derives from the entire set of points



Types of clustering

Conceptual clusters

- Points in a cluster share some general property that derives from the entire set of points



Quiz 01 (please answer on Piazza): Clustering types

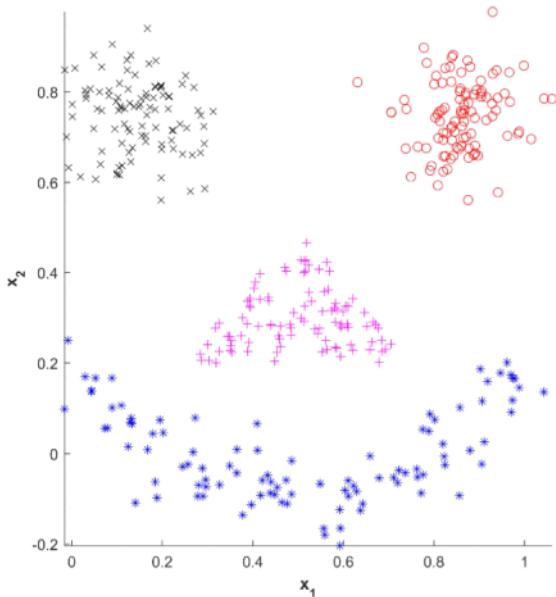


Figure 1: A clustering problem containing four clusters indicated by black crosses, red circles, magenta plusses and blue stars.

Consider the clustering problem given in Figure 1. Which clustering approach is *most* suited for correctly separating the data into the four groups indicated by black crosses, red circles, magenta plusses, and blue asterisks?

- A. A well-separated clustering approach.
- B. A contiguity-based clustering approach.
- C. A center-based clustering approach.
- D. A conceptual clustering approach.
- E. Don't know.

Clearly not weel-separated

Contiguity based: yes

Probably also center based clustering approach could be true

K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change

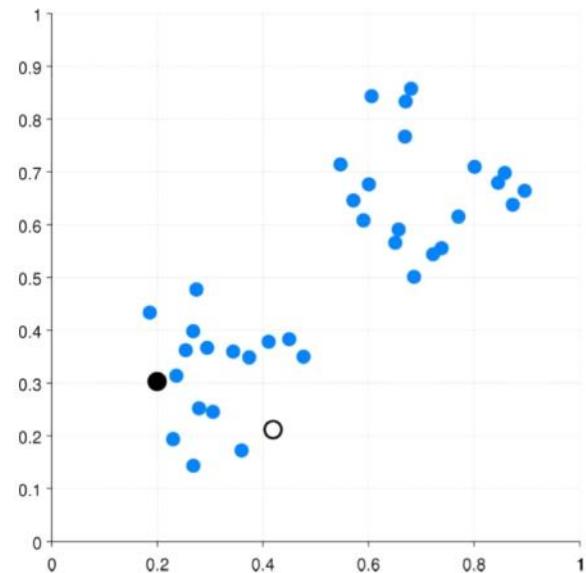
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



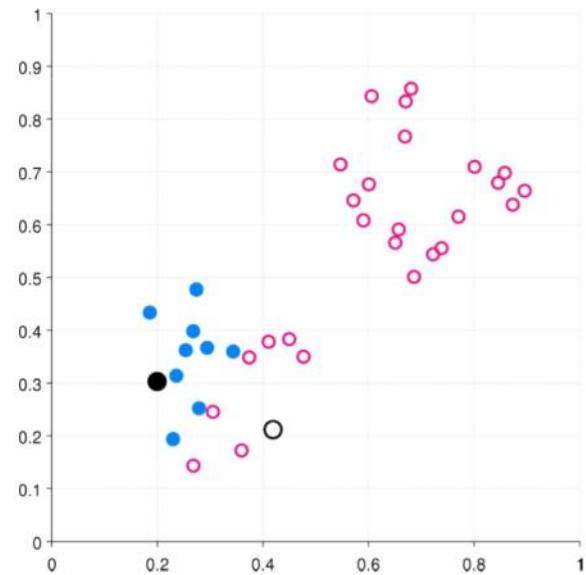
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



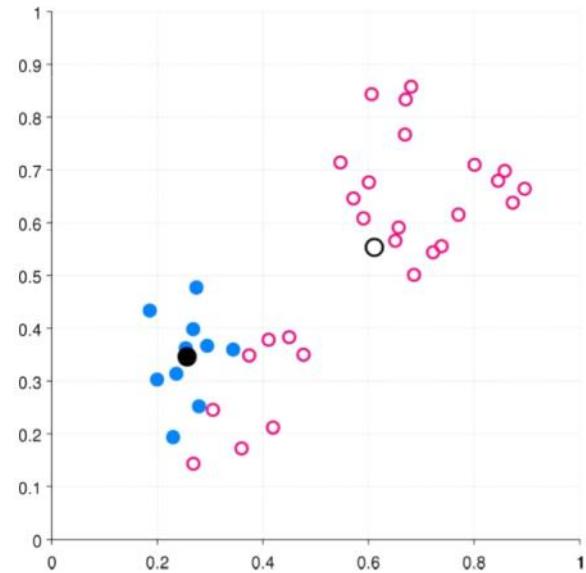
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



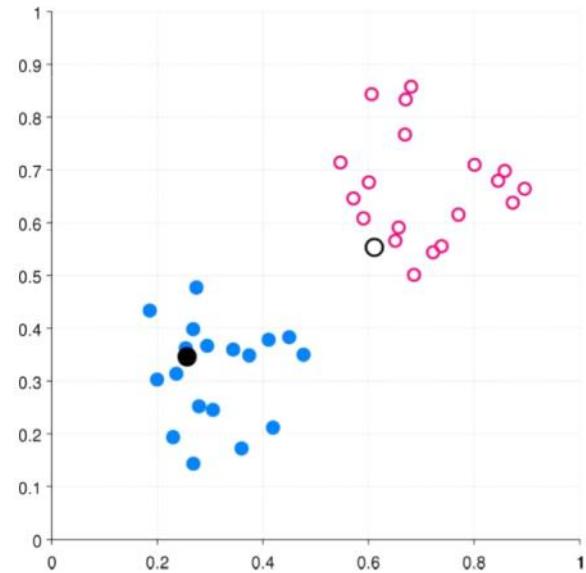
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



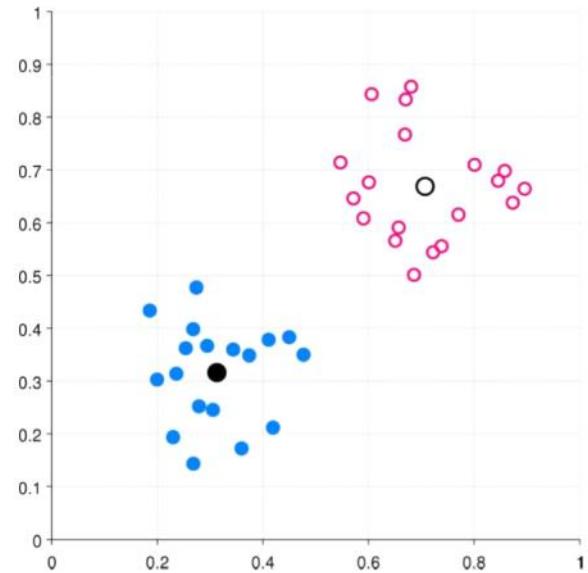
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



K-means clustering

How do I

- Find the closest centroid?
 - Use a suitable **dissimilarity/similarity measure**
- Compute the cluster centroids
 - Depends on dissimilarity/similarity measure
 - For example, for Euclidean distance the mean is optimal

Quiz 02 (please answer on Piazza): K-means

Consider the following dataset

$$\mathbf{X} = \{42, 60, 17, 48, 12\}$$

Select K points as initial centroids

Repeat

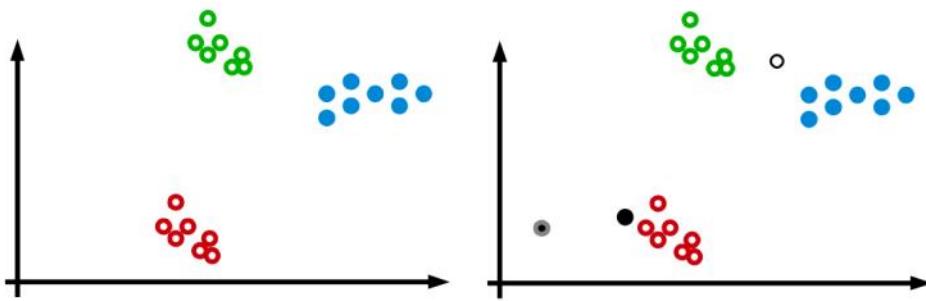
- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change

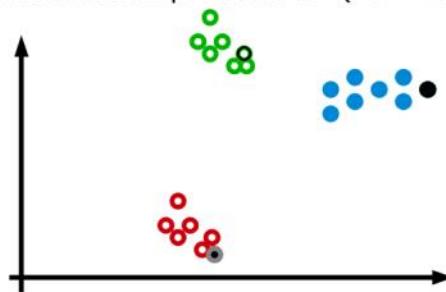
We wish to apply the K -means algorithm with $K = 2$ clusters to this dataset and we initialize with cluster centroids at $\mu_1 = 17$ and $\mu_2 = 12$. Carefully, using pen and paper, go through each step of the K -means algorithm until it converge. What is the final clustering?

- A. $\{60, 48\}, \{12, 17, 42\}$
- B. $\{42, 60, 48, 17\}, \{12\}$
- C. $\{60\}, \{12, 17, 42, 48\}$
- D. $\{42, 60, 48\}, \{12, 17\}$
- E. Don't know.

How will the data (top-left diagram) be clustered given the initialization of the three centroids shown at the right and at the bottom?



- What could we do if we have an empty cluster?
- What could be a good initialization procedure? (Farthest First)



Agglomerative hierarchical clustering

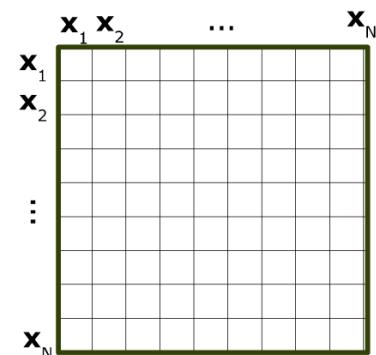
Initialize the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains

$$D_{ij} = \text{distance}(x_i, x_j)$$



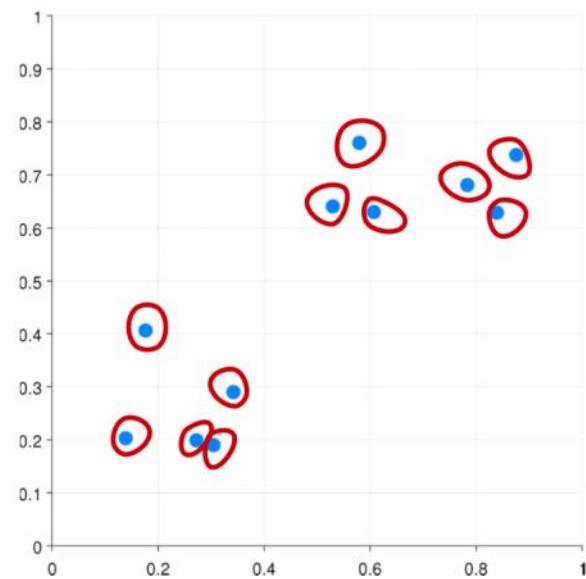
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



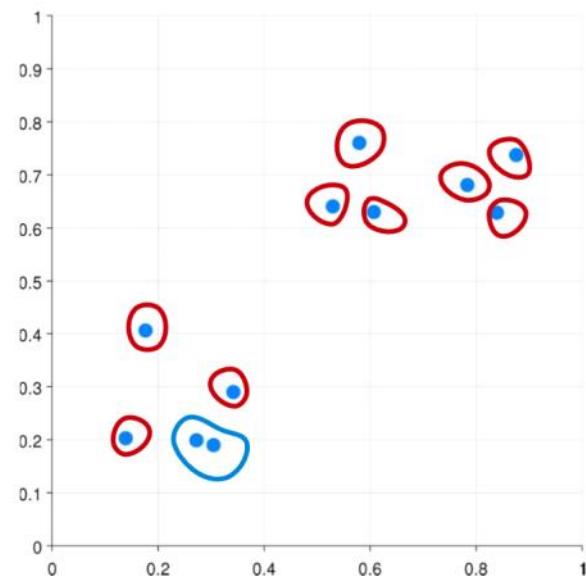
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



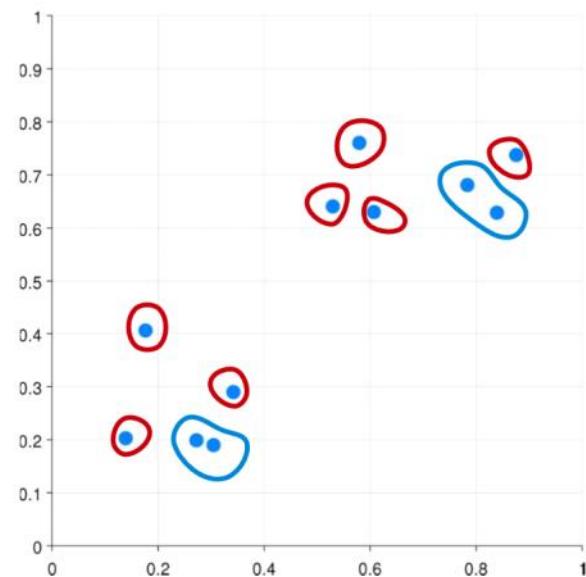
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



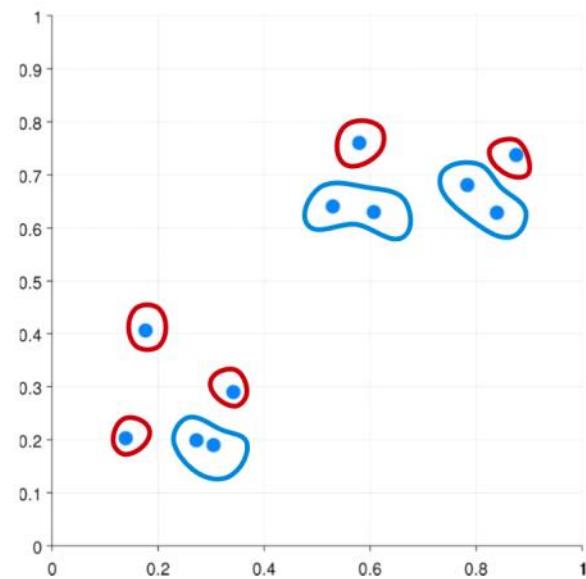
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



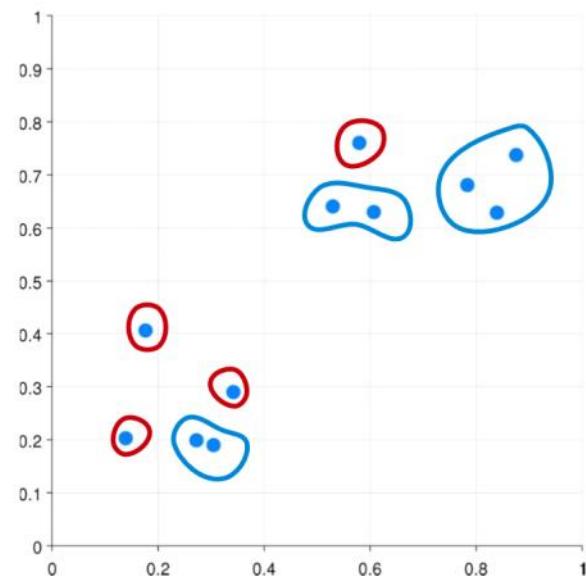
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



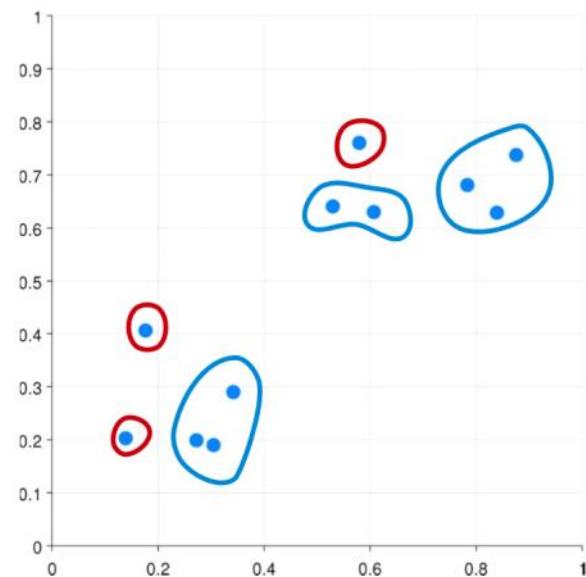
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



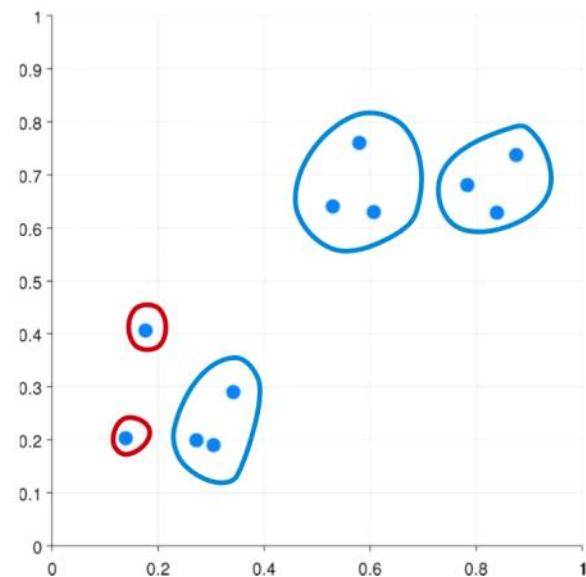
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



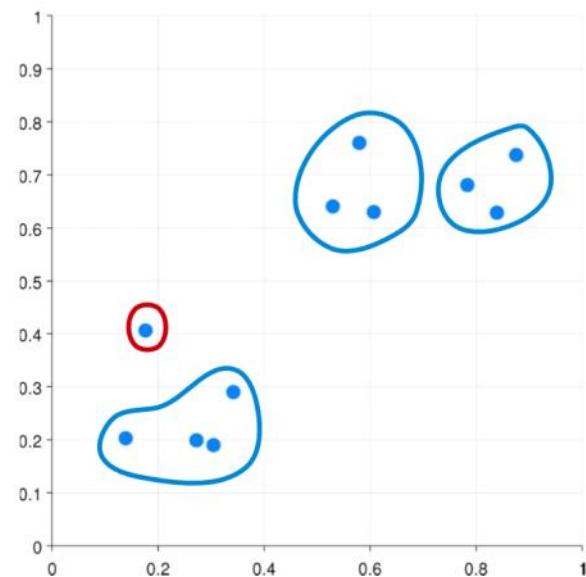
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



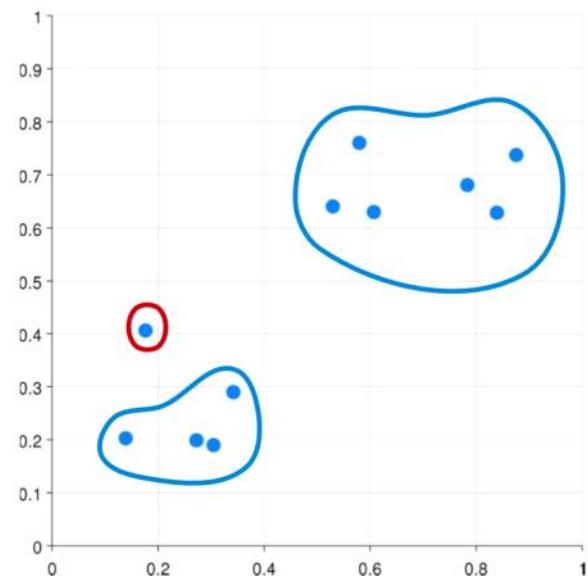
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



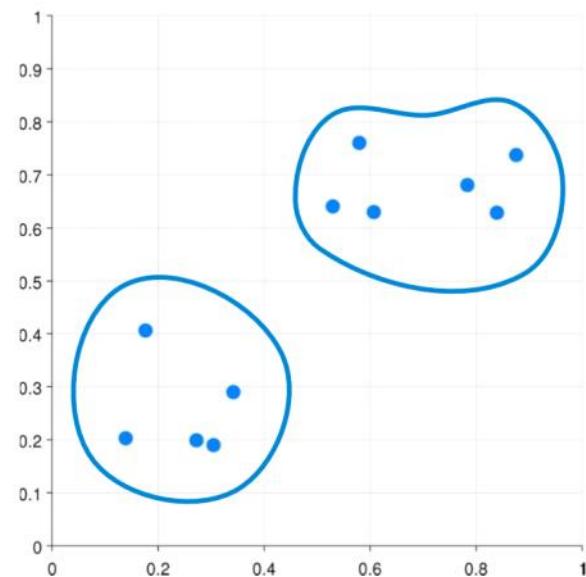
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



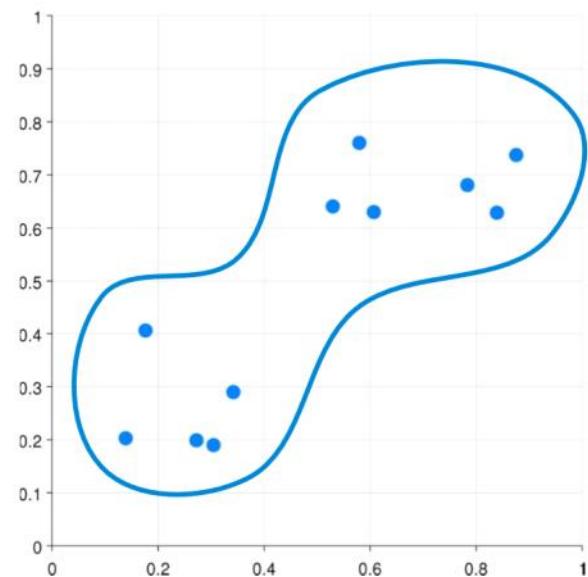
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

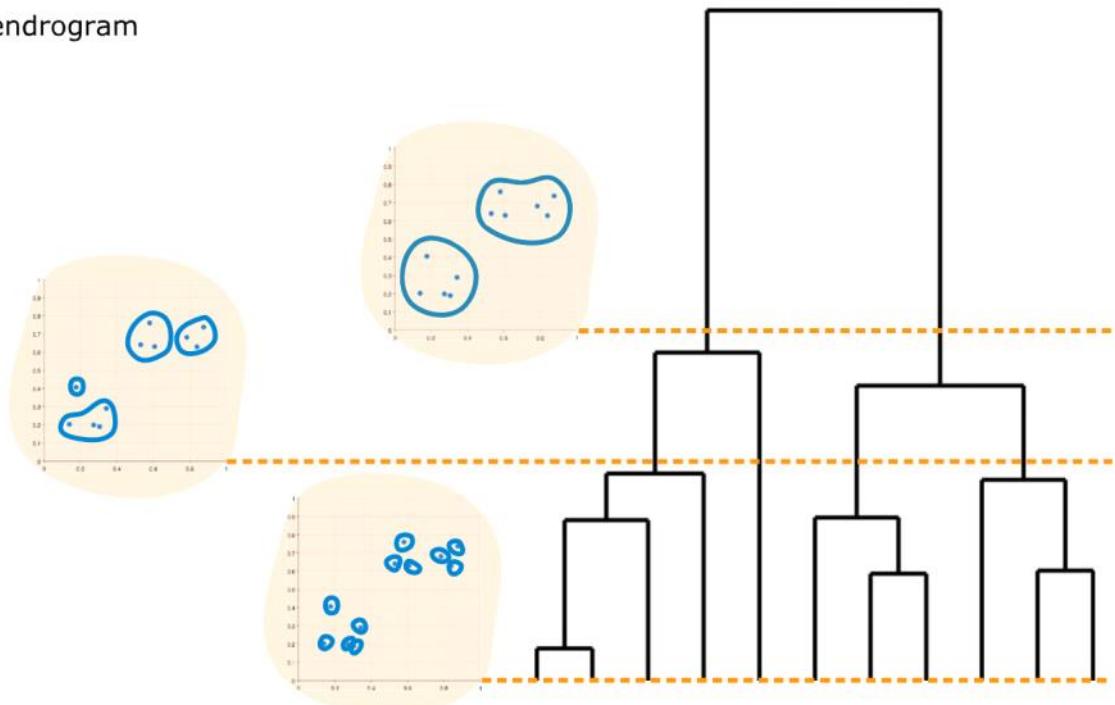
- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



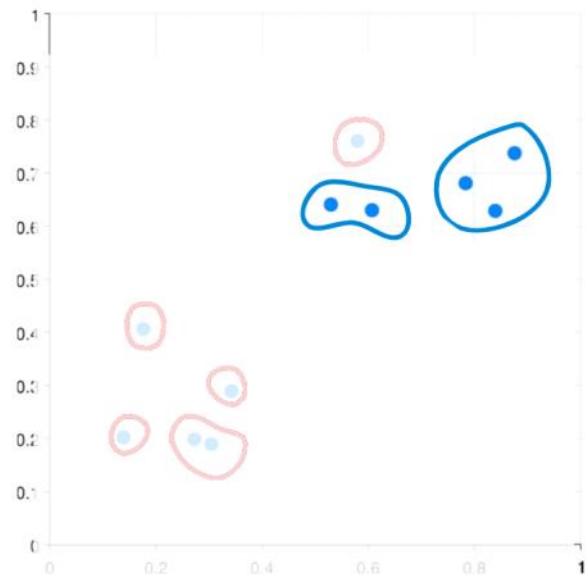
Agglomerative hierarchical clustering

- Dendrogram



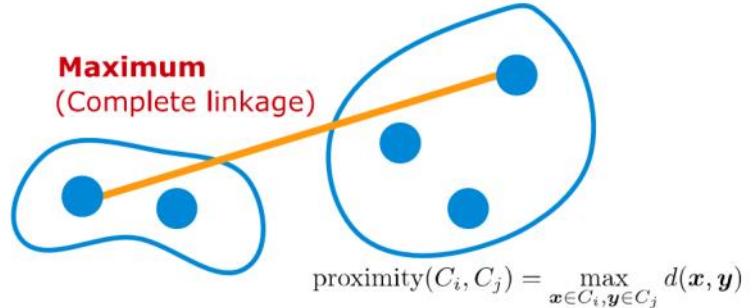
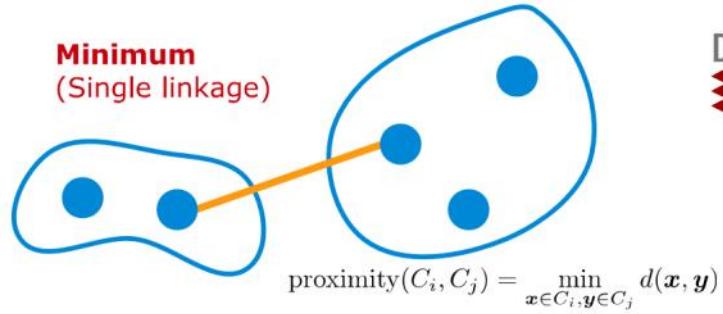
Similarity between clusters

- The **key operation** in agglomerative hierarchical clustering is measuring **distance (dissimilarity) between clusters**



Proximity between clusters

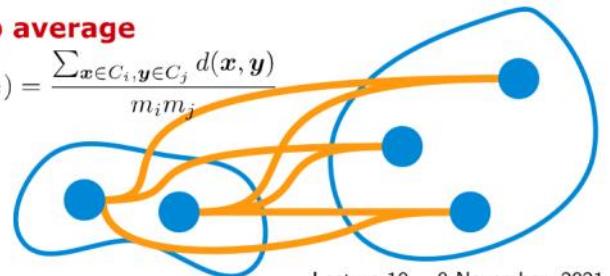
- Can be computed using **proximity between objects**
- In our example before we used Euclidian distance as proximity measure



C_i : Observations in cluster i
 C_j : Observations in cluster j
 m_i : Number of observations in cluster i
 m_j : Number of observations in cluster j

Group average

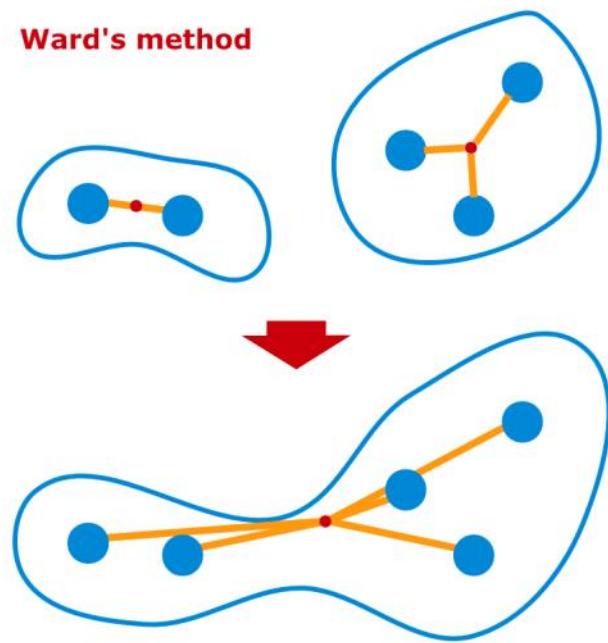
$$\text{proximity}(C_i, C_j) = \frac{\sum_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y})}{m_i m_j}$$



Similarity between clusters

- Increase in sum of squared error after merging the two clusters should be as small as possible

Ward's method

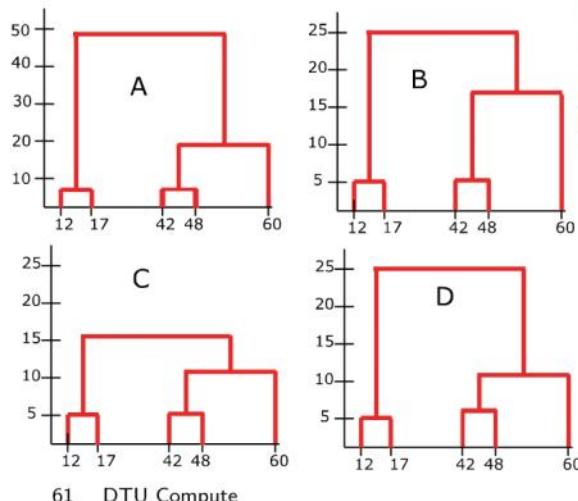


Quiz 03 (please answer on Piazza): Dendograms

Consider once more the dataset:

$$X = \{42, 60, 17, 48, 12\}$$

Using pen-and-paper, carefully build a dendrogram from X one step at a time using Euclidean distance and *minimum* (single) linkage. What will the dendrogram look like?



61 DTU Compute



Compute the proximity matrix

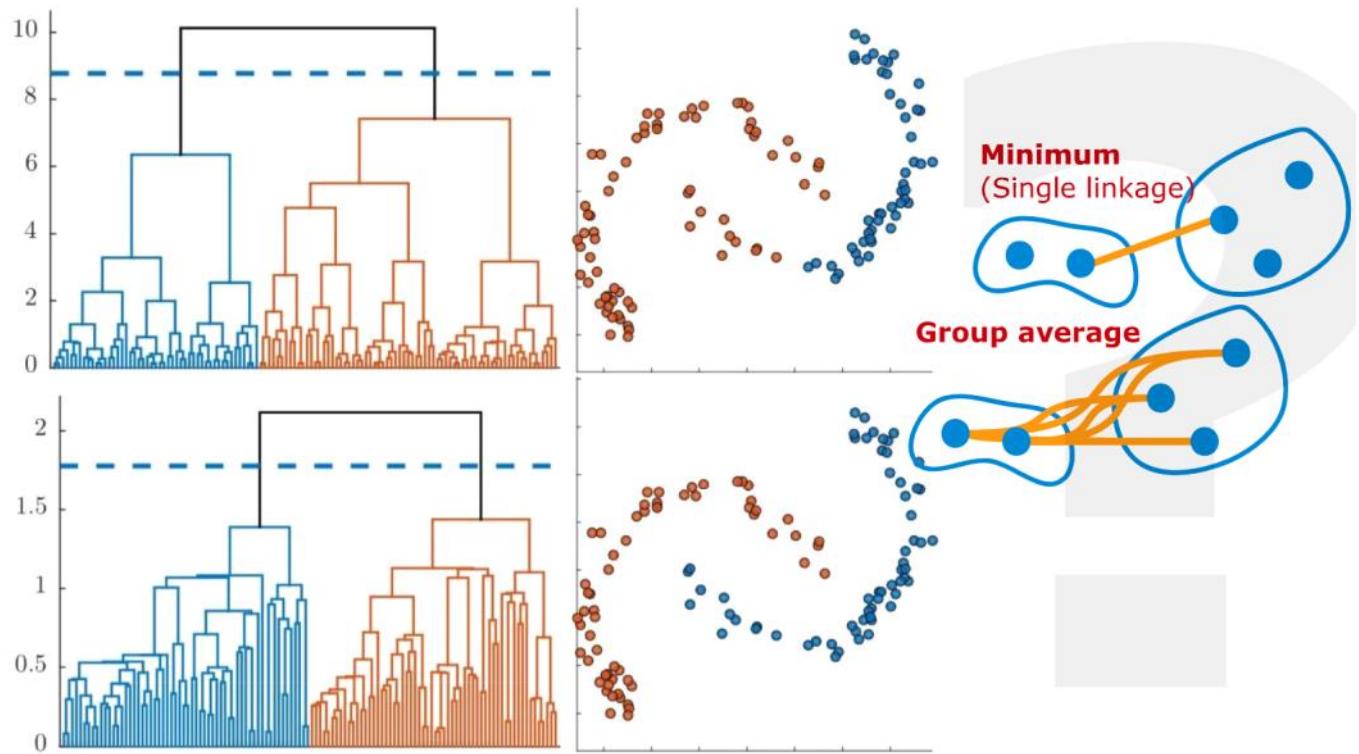
Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

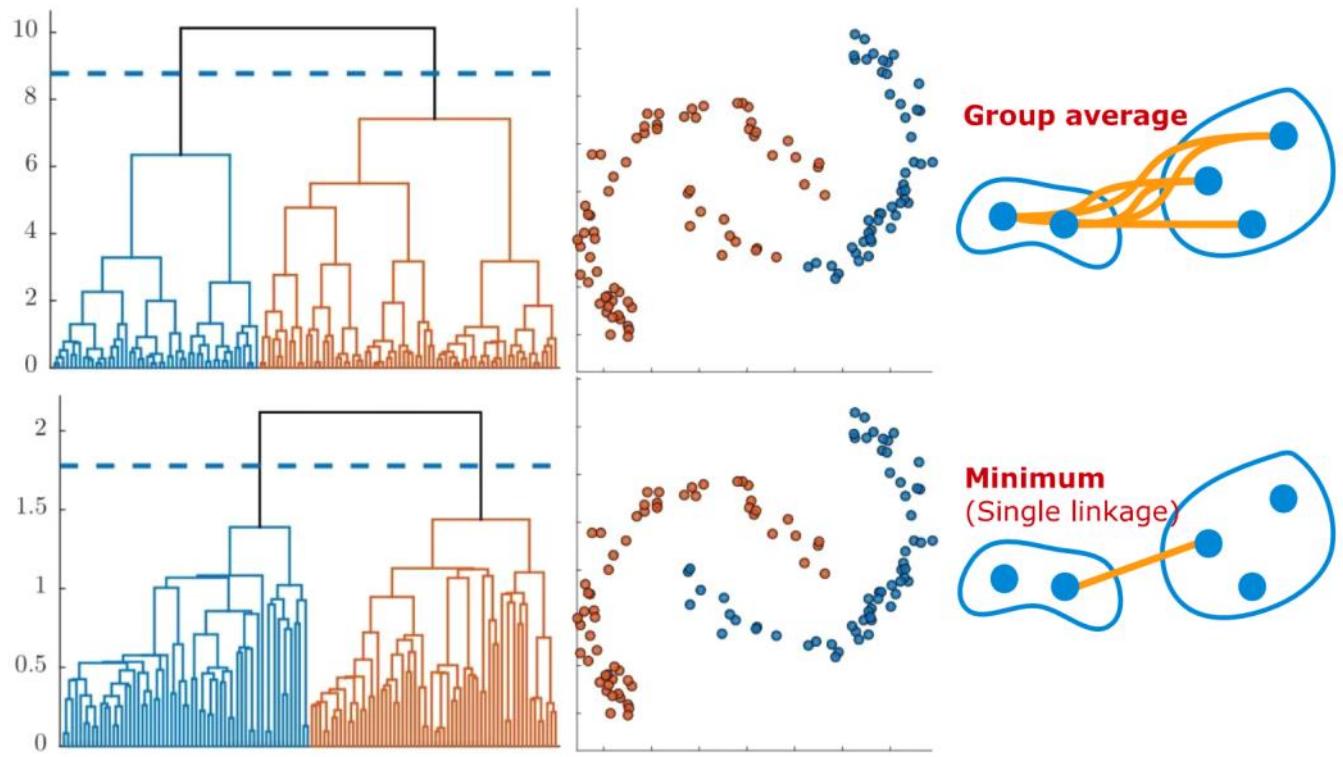
Until only one cluster remains

Lecture 10 9 November, 2021

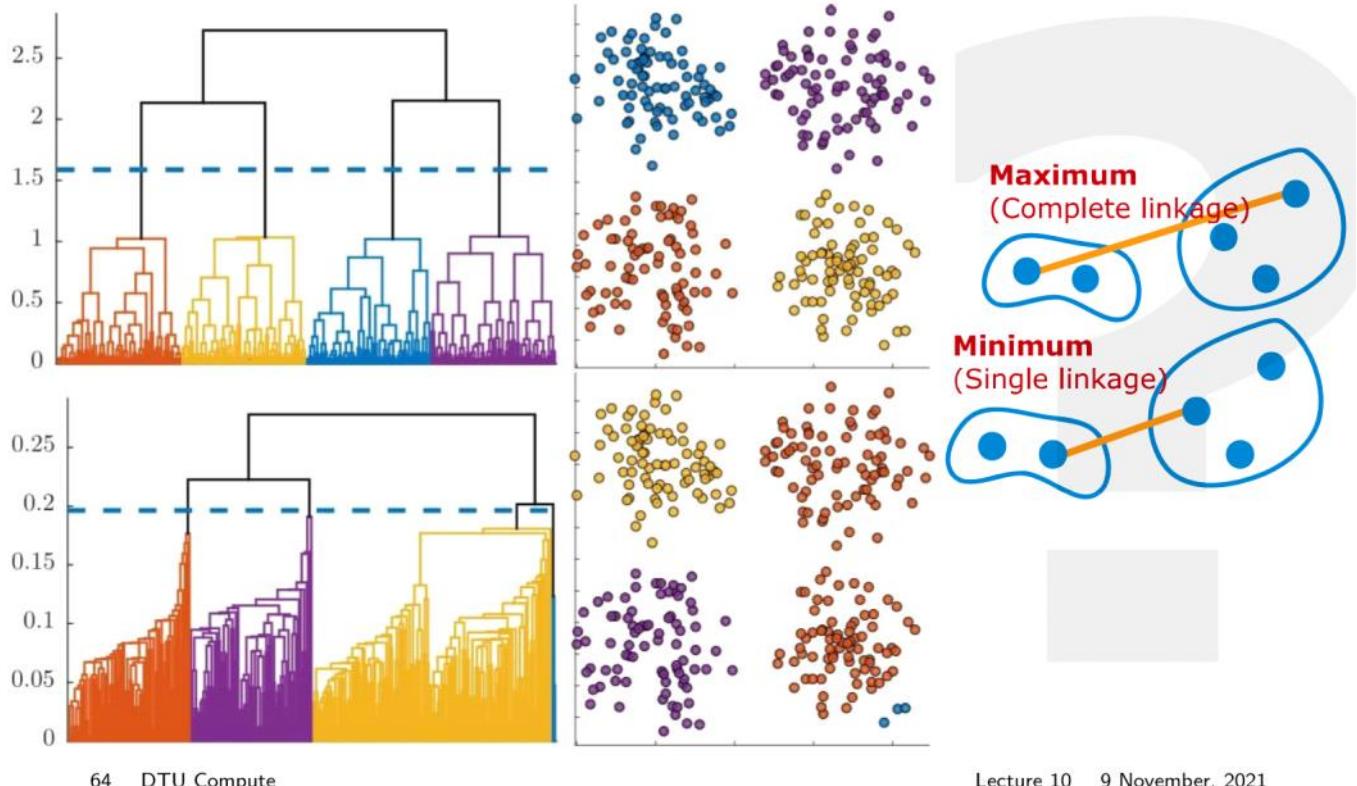
Clusterings and linkage function



Clusterings and linkage function



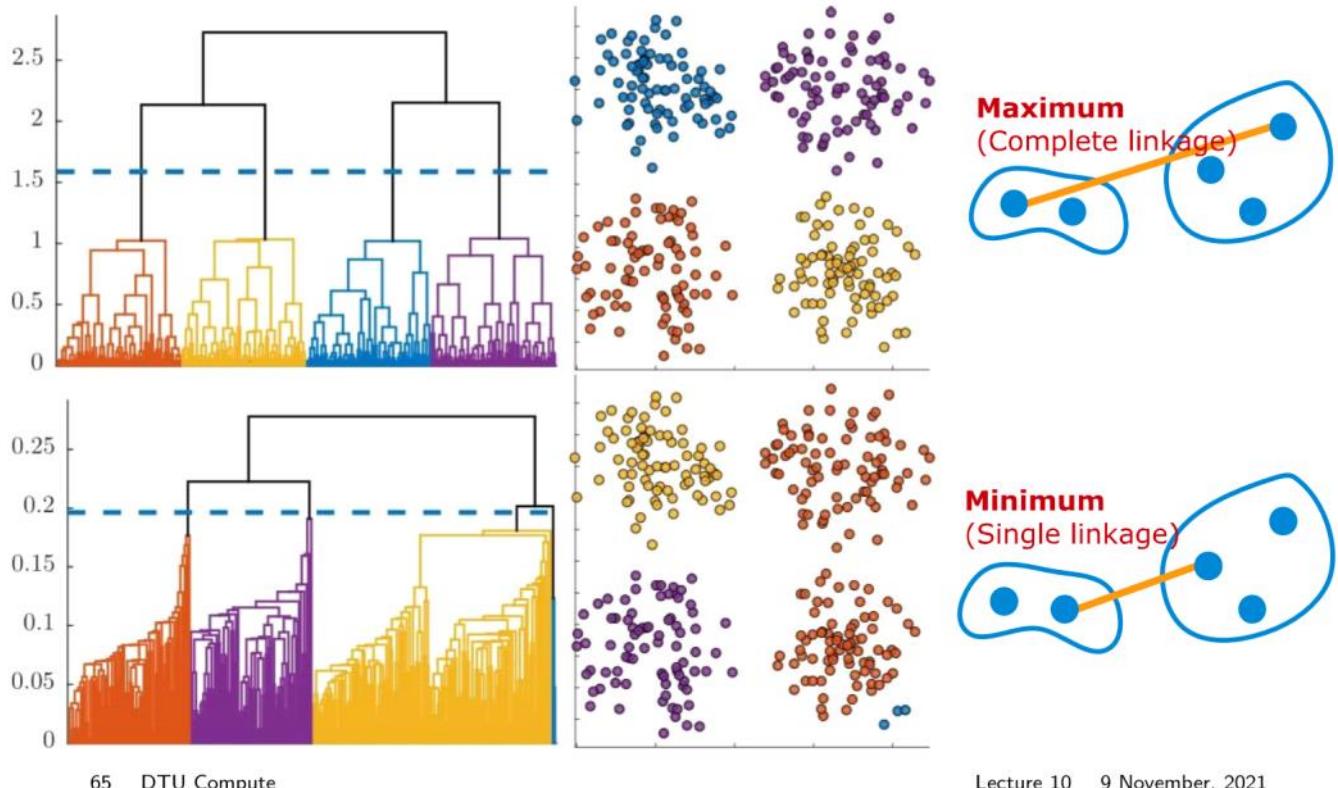
Clusterings and linkage function



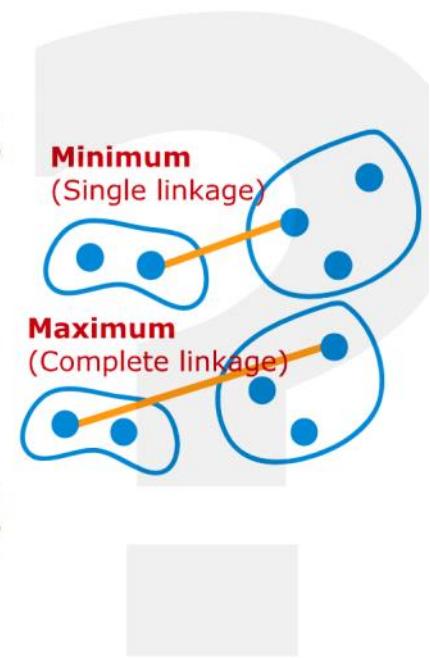
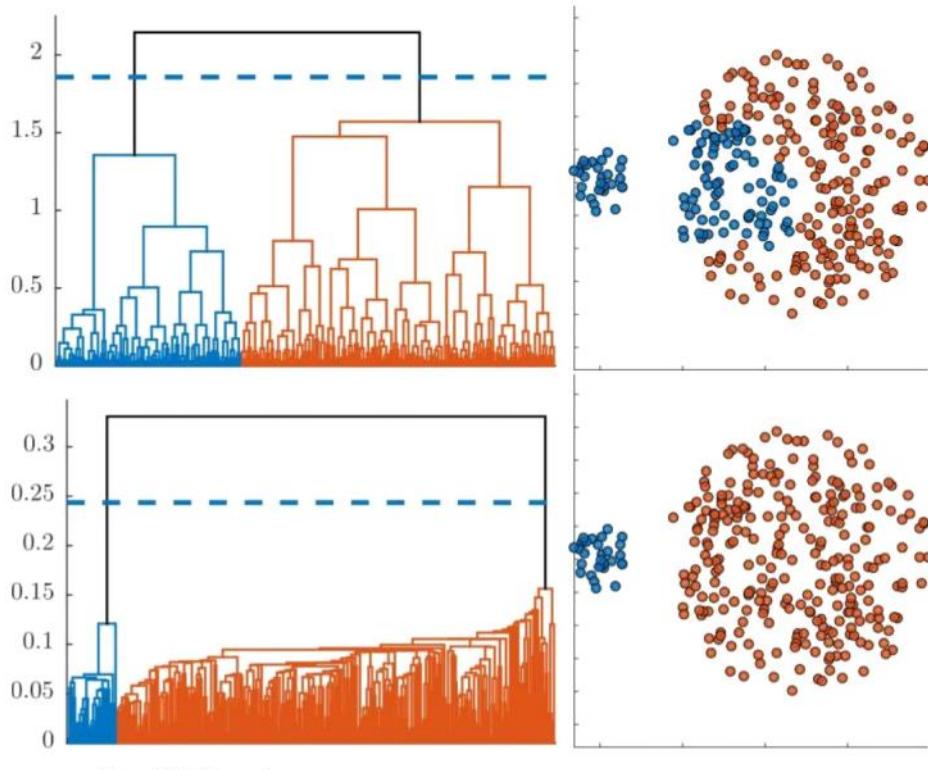
64 DTU Compute

Lecture 10 9 November, 2021

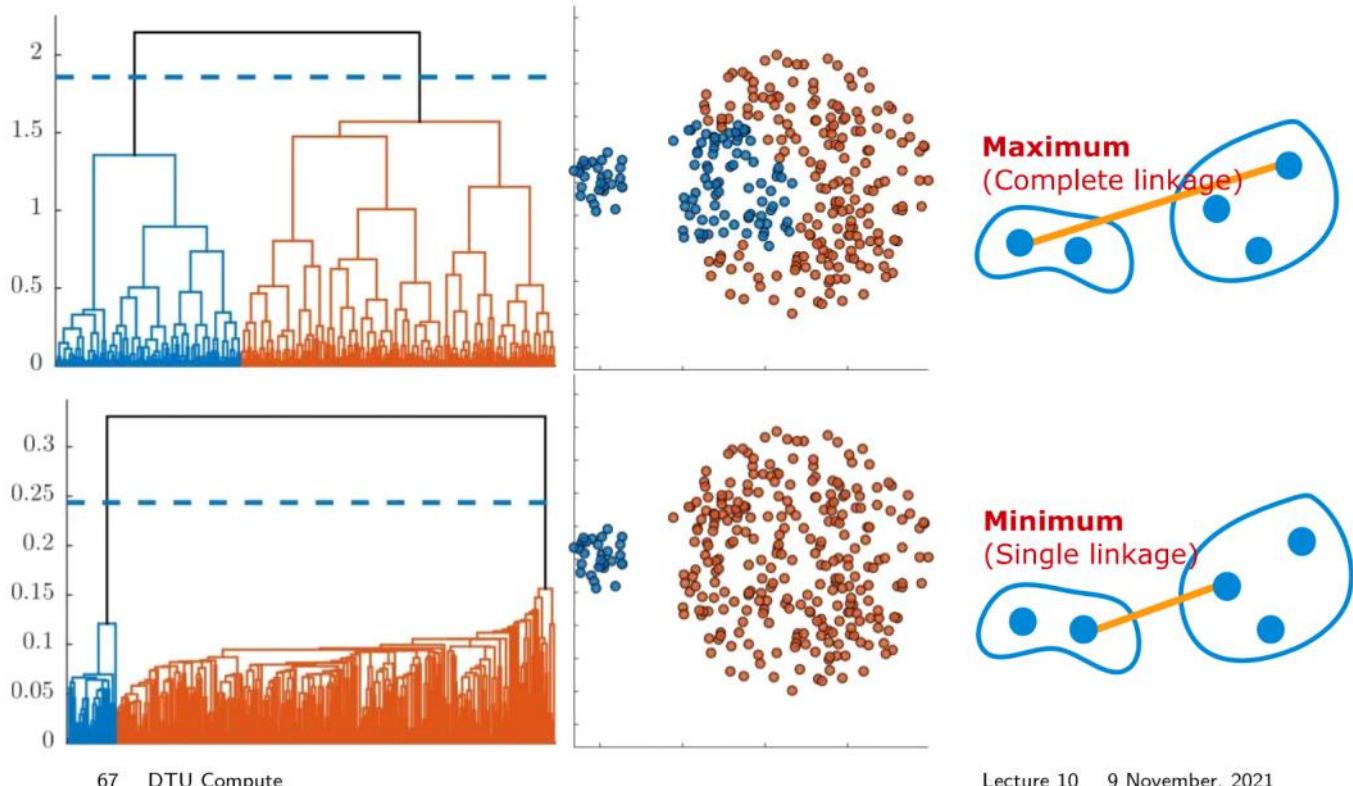
Clusterings and linkage function



Clusterings and linkage function



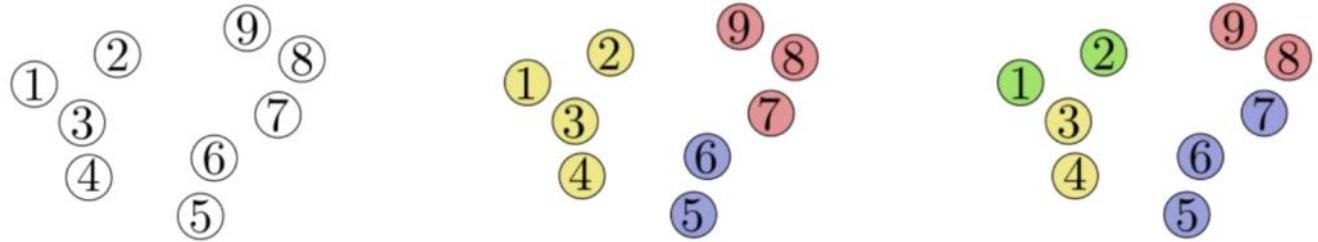
Clusterings and linkage function



Comparing partitions

- How similar are Q and Z

We have two cluster methods and we want to evaluate how similar they are



$$Z = [1 \ 1 \ 1 \ 1 \ 2 \ 2 \ 3 \ 3 \ 3]$$

$$Q = [4 \ 4 \ 1 \ 1 \ 2 \ 2 \ 2 \ 3 \ 3]$$

- Note encoding is (and should be!) arbitrary

$$Q' = [10 \ 10 \ 3 \ 3 \ 8 \ 8 \ 8 \ 1 \ 1]$$

Encoding

$$n_{km} = \{\text{Observations assigned to cluster } k \text{ in } Z \text{ and } m \text{ in } Q\} = \sum_{i=1}^N \sum_{j=1}^N \delta_{z_i, k} \delta_{z_j, m}$$

No matter how we label the data

$$\mathbf{n}^Z = \{\text{Number of observations assigned to cluster } k \text{ in } Z\} = \sum_{m=1}^M n_{km}$$

$$\mathbf{n}^Q = \{\text{Number of observations assigned to cluster } m \text{ in } Q\} = \sum_{k=1}^K n_{km}$$

Rows correspond to Z cluster are columns to Q
We have the first 2 because we have 2 entries (3,4) Class in Z and second in Q

That belongs to cluster 1 in Z and to

Cluster 1 in Q

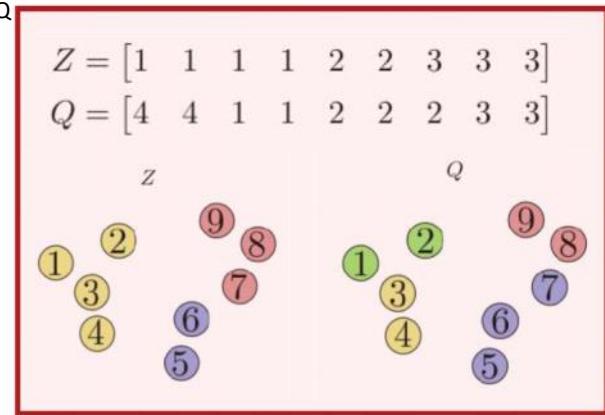
$$\mathbf{n} = \begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}$$

n(1,2) : how many points belong to cluster 1 in Z and to cluster 2 in Q

Note the horizontal/vertical sums of \mathbf{n} :

$$\mathbf{n}^Z = \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{n}^Q = [2 \ 3 \ 2 \ 2]$$

Total number belonging in Z To class1 class2 and class3



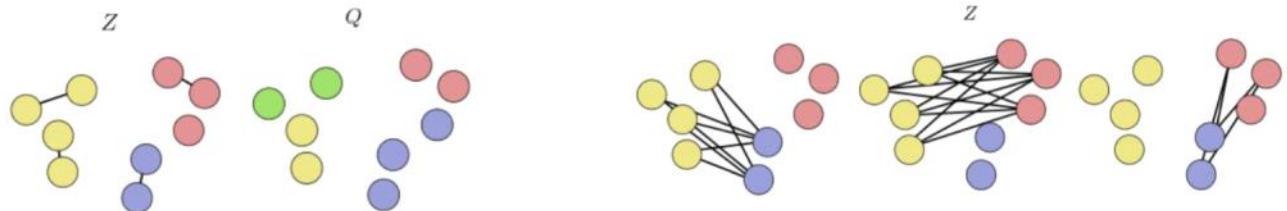
Jaccard and SMC

We need to use some metric that doesn't depend on the labels we gave

- Any two observations i, j can either be in the same cluster, or in different clusters
- There are $\frac{1}{2}N(N - 1)$ pairs total
- We get two $\frac{1}{2}N(N - 1)$ -long binary vectors corresponding to each pair i, j

$$S = \{ \text{ Number of pairs } i, j \text{ in the same cluster in } Z, Q \}$$

$$D = \{ \text{ Number of pairs } i, j \text{ in different clusters in } Z, Q \}$$



$$\text{Rand index: } R(Z, Q) = \frac{S + D}{\frac{1}{2}N(N - 1)} = \frac{4 + 24}{\frac{1}{2}9 \cdot 8} = \frac{7}{9},$$

Simple Match Coefficient: SMC

$$\text{Jaccard similarity: } J(Z, Q) = \frac{S}{\frac{1}{2}N(N - 1) - D} = \frac{4}{\frac{1}{2}9 \cdot 8 - 24} = \frac{1}{3}$$

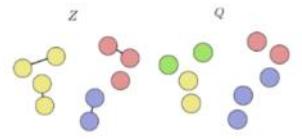
Jaccard and rand index in general

Recall

$$\mathbf{n} = \begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \quad \mathbf{n}^Z = \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{n}^Q = \begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

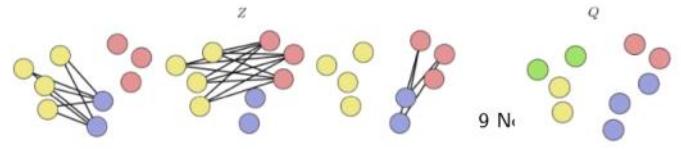
$S = \{ \text{ Number of pairs } i, j \text{ in the same cluster in } Z, Q \}$

$$\begin{aligned} &= \sum_{k=1}^K \sum_{m=1}^M \frac{n_{km}(n_{km} - 1)}{2} \\ &= \frac{2(2-1)}{2} + \frac{2(2-1)}{2} + \frac{2(2-1)}{2} + \frac{1(1-1)}{2} + \frac{2(2-1)}{2} = 4 \end{aligned}$$



$D = \{ \text{ Number of pairs } i, j \text{ in different clusters in } Z, Q \}$

$$\begin{aligned} &= \frac{N(N-1)}{2} - \sum_{k=1}^K \frac{n_k^Z(n_k^Z - 1)}{2} - \sum_{m=1}^M \frac{n_m^Q(n_m^Q - 1)}{2} + S \\ &= 36 - 10 - 6 + 4 = 24 \end{aligned}$$



Quiz 04: Cluster overlap

	<i>o</i> ₁	<i>o</i> ₂	<i>o</i> ₃	<i>o</i> ₄	<i>o</i> ₅	<i>o</i> ₆	<i>o</i> ₇	<i>o</i> ₈	<i>o</i> ₉	<i>o</i> ₁₀
<i>o</i> ₁	0.0	2.0	5.7	0.9	2.9	1.8	2.7	3.7	5.3	5.1
<i>o</i> ₂	2.0	0.0	5.6	2.4	2.5	3.0	3.5	4.3	6.0	6.2
<i>o</i> ₃	5.7	5.6	0.0	5.0	5.1	4.0	3.3	5.4	1.2	1.8
<i>o</i> ₄	0.9	2.4	5.0	0.0	2.7	2.1	2.2	3.5	4.6	4.4
<i>o</i> ₅	2.9	2.5	5.1	2.7	0.0	3.5	3.7	4.0	5.8	5.7
<i>o</i> ₆	1.8	3.0	4.0	2.1	3.5	0.0	1.7	5.3	3.8	3.7
<i>o</i> ₇	2.7	3.5	3.3	2.2	3.7	1.7	0.0	4.2	3.1	3.2
<i>o</i> ₈	3.7	4.3	5.4	3.5	4.0	5.3	4.2	0.0	5.5	6.0
<i>o</i> ₉	5.3	6.0	1.2	4.6	5.8	3.8	3.1	5.5	0.0	2.1
<i>o</i> ₁₀	5.1	6.2	1.8	4.4	5.7	3.7	3.2	6.0	2.1	0.0

Table 1: The pairwise distances between $N = 10$ observations from the travel review dataset. the colors indicate classes

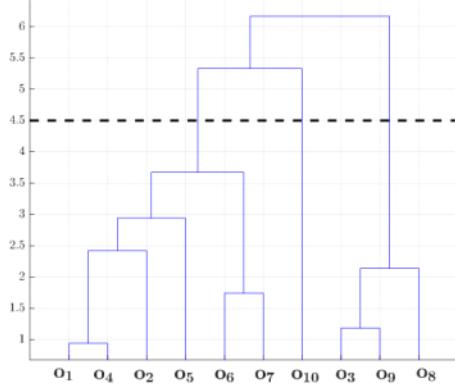


Figure 1: Dendrogram with a cutoff generating 3 clusters.

Consider the dendrogram in Figure 1. Suppose we apply a cutoff (indicated by the black line) thereby generating three clusters. We wish to compare the quality of this clustering, Q , to the ground-truth clustering, Z , indicated by the colors in Table 1. Recall the *Jaccard similarity* of the two clusterings is

$$J[Z, Q] = \frac{S}{\frac{1}{2}N(N-1) - D}$$

in the notation of the lecture notes. What is the Jaccard similarity of the two clusterings?

- A. $J[Z, Q] \approx 0.104$
- B. $J[Z, Q] \approx 0.143$
- C. $J[Z, Q] \approx 0.174$
- D. $J[Z, Q] \approx 0.153$
- E. Don't know.

Calculate n, then S and D, and plug into the Jacard

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i, i = 1, \dots, n$

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i, i = 1, \dots, n$
- **Information** obtained from observing x_i is

$$I = -\log p_i$$

If it is very likely, then the information is low and
I'm not very surprised by see this outcome.
If it's unlike, the the information is high

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i, i = 1, \dots, n$
- **Information** obtained from observing x_i is

$$I = -\log p_i$$

- Average information obtained is called the **entropy**

$$H[p_X] = \mathbb{E}[I] = - \sum_{i=1}^n p_i \log p_i$$

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i, i = 1, \dots, n$
- **Information** obtained from observing x_i is

$$I = -\log p_i$$

- Average information obtained is called the **entropy**

$$H[p_X] = \mathbb{E}[I] = - \sum_{i=1}^n p_i \log p_i$$

- Entropy is defined for general densities $P(X = x_i, Y = y_j) = p_{ij}$

$$H[p_{XY}] = - \sum_{i=1}^n \sum_{j=1}^m p_{ij} \log p_{ij}$$

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i, i = 1, \dots, n$
- **Information** obtained from observing x_i is

$$I = -\log p_i$$

- Average information obtained is called the **entropy**

$$H[p_X] = \mathbb{E}[I] = - \sum_{i=1}^n p_i \log p_i$$

- Entropy is defined for general densities $P(X = x_i, Y = y_j) = p_{ij}$

$$H[p_{XY}] = - \sum_{i=1}^n \sum_{j=1}^m p_{ij} \log p_{ij}$$

- The **Mutual information** is defined as

$$\text{MI}[X, Y] = H[P_X] + H[P_Y] - H[P_{XY}]$$

The amount of information that we obtain about one variable
Through the other variable

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i, i = 1, \dots, n$
- **Information** obtained from observing x_i is

$$I = -\log p_i$$

- Average information obtained is called the **entropy**

$$H[p_X] = \mathbb{E}[I] = - \sum_{i=1}^n p_i \log p_i$$

- Entropy is defined for general densities $P(X = x_i, Y = y_j) = p_{ij}$

$$H[p_{XY}] = - \sum_{i=1}^n \sum_{j=1}^m p_{ij} \log p_{ij}$$

- The **Mutual information** is defined as

$$\text{MI}[X, Y] = H[P_X] + H[P_Y] - H[P_{XY}]$$

- The **Normalized mutual information** is defined as

$$\text{NMI}[X, Y] = \frac{\text{MI}[X, Y]}{\sqrt{H[P_X]} \sqrt{H[P_Y]}}$$

Comparing using mutual information

If we have 2 cluster, then we can use information theory to say something about how similar they are. We define $P_{ZQ}(i, j) = \frac{1}{N} n_{ij}$, $P_Z(i) = \frac{n^Z}{N}$ and $P_Q(j) = \frac{n^Q}{N}$. Example:

$$P_{ZQ} = \frac{1}{9} \begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \quad P_Z = \frac{1}{9} \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad P_Q = \frac{1}{9} \begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

We get a joint probability distribution over two cluster, dividing by the number of elements: if I randomly draw i,j what's the probability that they lie in the cluster both in Z and Q

Similarly, I can define a prob distribution over Z and Q (P_z and P_q): if I take a random node what's the prob that it belong to the clusters

Comparing using mutual information

We define $P_{ZQ}(i, j) = \frac{1}{N}n_{ij}$, $P_Z(i) = \frac{n_i^Z}{N}$ and $P_Q(j) = \frac{n_j^Q}{N}$. Example:

$$P_{ZQ} = \frac{1}{9} \begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \quad P_Z = \frac{1}{9} \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad P_Q = \frac{1}{9} \begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

- **Entropy** computed as $H[p_X] = -\sum_{i=1}^n p_i \log p_i$:

$$\text{Entropy of } Z: \quad H[Z] = -\frac{4}{9} \log \frac{4}{9} - \frac{1}{3} \log \frac{1}{3} - \frac{2}{9} \log \frac{2}{9} \approx 1.06$$

$$\text{Entropy of } Q: \quad H[Q] = -\frac{2}{9} \log \frac{2}{9} - \frac{2}{9} \log \frac{2}{9} - \frac{2}{9} \log \frac{2}{9} - \frac{1}{3} \log \frac{1}{3} \approx 1.37$$

$$\text{Entropy of } Z \text{ and } Q: \quad H[ZQ] = -4 \times \frac{2}{9} \log \frac{2}{9} - \frac{1}{9} \log \frac{1}{9} = 1.58.$$

Comparing using mutual information

We define $P_{ZQ}(i, j) = \frac{1}{N}n_{ij}$, $P_Z(i) = \frac{n_i^Z}{N}$ and $P_Q(j) = \frac{n_j^Q}{N}$. Example:

$$P_{ZQ} = \frac{1}{9} \begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \quad P_Z = \frac{1}{9} \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad P_Q = \frac{1}{9} \begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

- **Entropy** computed as $H[p_X] = -\sum_{i=1}^n p_i \log p_i$:

$$\text{Entropy of } Z: \quad H[Z] = -\frac{4}{9} \log \frac{4}{9} - \frac{1}{3} \log \frac{1}{3} - \frac{2}{9} \log \frac{2}{9} \approx 1.06$$

$$\text{Entropy of } Q: \quad H[Q] = -\frac{2}{9} \log \frac{2}{9} - \frac{2}{9} \log \frac{2}{9} - \frac{2}{9} \log \frac{2}{9} - \frac{1}{3} \log \frac{1}{3} \approx 1.37$$

$$\text{Entropy of } Z \text{ and } Q: \quad H[ZQ] = -4 \times \frac{2}{9} \log \frac{2}{9} - \frac{1}{9} \log \frac{1}{9} = 1.58.$$

- **Mutual information:**

$$\text{MI}[Z, Q] = H[Z] + H[Q] - H[Z, Q] \approx 1.06 + 1.37 - 1.58 \approx 0.85.$$

Comparing using mutual information

We define $P_{ZQ}(i, j) = \frac{1}{N}n_{ij}$, $P_Z(i) = \frac{n_i^Z}{N}$ and $P_Q(j) = \frac{n_j^Q}{N}$. Example:

$$P_{ZQ} = \frac{1}{9} \begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \quad P_Z = \frac{1}{9} \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad P_Q = \frac{1}{9} \begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

- **Entropy** computed as $H[p_X] = -\sum_{i=1}^n p_i \log p_i$:

$$\text{Entropy of } Z: \quad H[Z] = -\frac{4}{9} \log \frac{4}{9} - \frac{1}{3} \log \frac{1}{3} - \frac{2}{9} \log \frac{2}{9} \approx 1.06$$

$$\text{Entropy of } Q: \quad H[Q] = -\frac{2}{9} \log \frac{2}{9} - \frac{2}{9} \log \frac{2}{9} - \frac{2}{9} \log \frac{2}{9} - \frac{1}{3} \log \frac{1}{3} \approx 1.37$$

$$\text{Entropy of } Z \text{ and } Q: \quad H[ZQ] = -4 \times \frac{2}{9} \log \frac{2}{9} - \frac{1}{9} \log \frac{1}{9} = 1.58.$$

- **Mutual information:**

$$\text{MI}[Z, Q] = H[Z] + H[Q] - H[Z, Q] \approx 1.06 + 1.37 - 1.58 \approx 0.85.$$

- **Normalized mutual information:**

$$\text{NMI}[Z, Q] = \frac{\text{MI}[Z, Q]}{\sqrt{H[Z]}\sqrt{H[Q]}} \approx \frac{0.85}{\sqrt{1.06}\sqrt{1.37}} \approx 0.70.$$

A number between 0 and 1 which tells how similar are the 2 split clusters

$$\text{NMI}[Z, Q] = \frac{\text{MI}[Z, Q]}{\sqrt{H[Z]}\sqrt{H[Q]}} \approx \frac{0.85}{\sqrt{1.06}\sqrt{1.37}} \approx 0.70.$$

A number between 0 and 1 which tells how similar are the 2 split clusters



02450: Introduction to Machine Learning and Data Mining

Mixture models and density estimation

Tommy Sonne Alstrøm

DTU Compute, Technical University of Denmark (DTU)

DTU Compute

Department of Applied Mathematics and Computer Science

$$\int_a^b \Theta^{17} \delta e^{i\pi} = -1$$

$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$

 $\infty = \{2.7182818284\}$

 $\chi^2 \approx \Sigma!$

Today Remember you can evaluate the course+TAs on DTU Inside

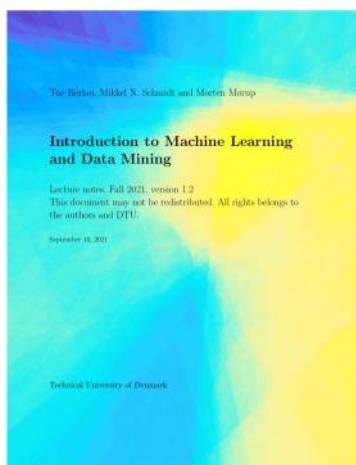


Feedback Groups of the day:

Abdulrahman Ramadan, Akos Pilis, Amira Mohamed Mohamoud Omar, Andrias Poulsen, Casper Rasmussen, Christian Kento Rasmussen, Christian Merithz Uhrenfeldt Nielsen, Djem Rizalar, Elisha Hirushani Peiris, Emmely Sofia Blom Vigsø, Frederik Sartov Olsen, Frederikke Uldahl Martensen, Giorgia Moranzoni, Guillaume Ratier, Gustav Troelsen Vindelev, Jake Hoang Viet Minh Pham, Johan Jens Kryger Larsen, Jonas Høj Jensen, Jonas Reeves Ussing, Jonathan Gilbert Ribergaard Mai, Jonathan Sørup Lund, Josep Maria Pujol March, Juliette Marie Victoire Vlieghe, Junaid Ahmed Qazi, Kasper Buur Vistesen, Kasper Helverskov Petersen, Kasper Kruse Lange, Kerstin Johanna Felicitas von Borries, Kevin Sonne, Line Selma Meinert Petersen, Mads Vibe Ringsted, Magnus Guldberg Pedersen, Magnus Vilhjalm Barrett Levinse, Marco Saretta, Marcus Roberto Nielsen, Mathias Daniel Frosz Nielsen, Milad Taghikhani, Morten Kielsgaard Dziegieł Sørensen, Niels Krätzel Ørdam, Oline Zachariassen, Oliver Low Petersen, Oliver Zacho, Ousama Mhadden, Rasmus Holmgård Nielsen, Ricky Hyldgaard Pedersen, Sara Riegels Trængbæk, Sarthak Trehan, Søren Nielsen, Taras Zykow Gordiyuk, Teodor Manne de Val Weywadt, Tobias Holmegaard Schwarze, Tommy Vu, William Henrik Klingsten Peytz

Reading material:

Chapter 19, Chapter 20



Lecture Schedule



- ① Introduction
31 August: C1
Data: Feature extraction, and visualization
- ② Data, feature extraction and PCA
7 September: C2, C3
- ③ Measures of similarity, summary statistics and probabilities
14 September: C4, C5
- ④ Probability densities and data visualization
21 September: C6, C7
Supervised learning: Classification and regression
- ⑤ Decision trees and linear regression
28 September: C8, C9
- ⑥ Overfitting, cross-validation and Nearest Neighbor
5 October: C10, C12 (Project 1 due before 13:00)
- ⑦ Performance evaluation, Bayes, and Naive Bayes
12 October: C11, C13
- ⑧ Artificial Neural Networks and Bias/Variance
26 October: C14, C15
- ⑨ AUC and ensemble methods
2 November: C16, C17
Unsupervised learning: Clustering and density estimation
- ⑩ K-means and hierarchical clustering
9 November: C18
- ⑪ Mixture models and density estimation
16 November: C19, C20 (Project 2 due before 13:00)
- ⑫ Association mining
23 November: C21
Recap
- ⑬ Recap and discussion of the exam
30 November: C1-C21

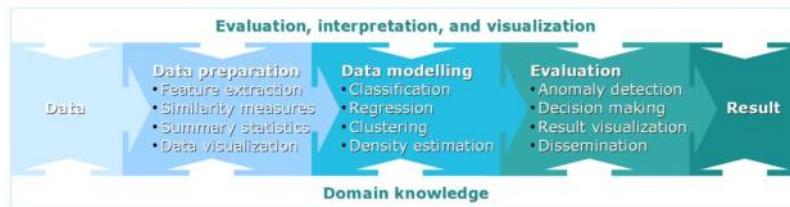
Online help: Forum on DTU Learn

Videos of lectures: <https://video.dtu.dk>

Streaming of lectures: Zoom (link on DTU Learn)

3 DTU Compute

Lecture 11 16 November, 2021



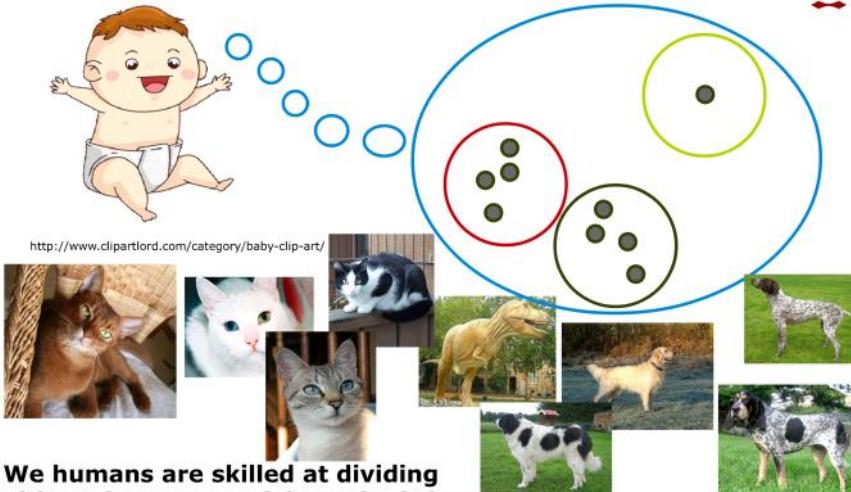
Learning Objectives

- Explain the role of the parameters in the Gaussian Mixture Model (GMM) and how the parameters are updated using the EM-algorithm
- Explain how cross-validation can be used for GMM
- Understand and apply kernel density, K-nearest neighbour density and average relative density estimation for outlier detection

4 DTU Compute

Lecture 11 16 November, 2021

Imagine you observe the world for the first time! DTU

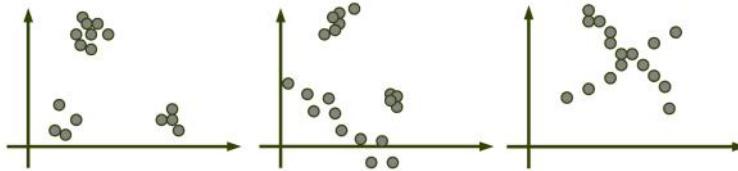


We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?

5 DTU Compute

http://www.clipartlord.com/category/baby-clip-art/
http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Cat_which_cat_on_fence.jpg
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
http://commons.wikimedia.org/wiki/File:MasPin-Astro-SVE.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthR_wb.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthairPaw.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthairPaw02.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthairPaw03.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthairPaw04.jpg
http://commons.wikimedia.org/wiki/File:Sauveterre2.jpg

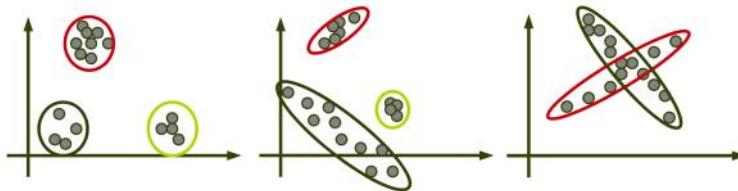
- What are the clusters below and what characterize each cluster?
- Is **k-means** well suited for modeling the clusters below?
 - Will it always find the optimum solution?
 - Can it model the sizes of the clusters?
 - Can it model the shape of the clusters?
 - How can we determine the number of clusters?



6 DTU Compute

Lecture 11 16 November, 2021

- What are the clusters below and what characterize each cluster?
- Is **k-means** well suited for modeling the clusters below?
 - Will it always find the optimum solution?
 - Can it model the sizes of the clusters? We want a model that can shape the cluster
 - Can it model the shape of the clusters? And determine the number of K nicely
 - How can we determine the number of clusters?



7 DTU Compute

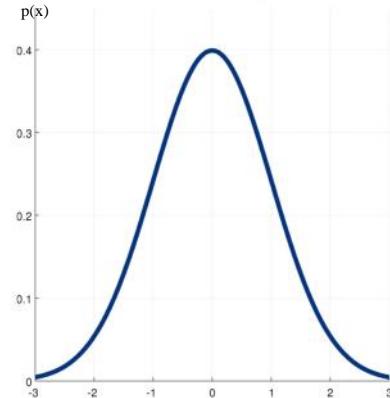
Lecture 11 16 November, 2021

The density estimation problem: we want to estimate this density(curve) based on some observation.

Normal distribution

- Probability density function describes the relative chance of a given value to occur
- Normal distribution characterized by
 - Mean
 - Variance

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



8 DTU Compute

Lecture 11 16 November, 2021

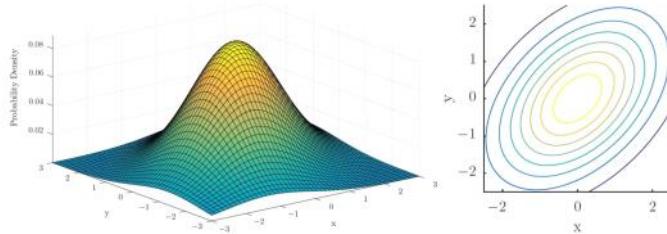
Multivariate Normal distribution

2d-case

It returns the density of a specific observation

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

In the 2d case, it's returning a scalar



Multivariate Normal distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- Example: 2-dimensional Normal distribution

$$\boldsymbol{\mu} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

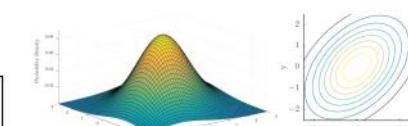
$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{bmatrix}$$

Cov(x1,x2) : if I observe x1, what I learn about x2?

- If cov(x1,x2) = 0, then we don't get anything and viceversa
- If cov(x1,x2) > 0:
- If cov(x1,x2) < 0:



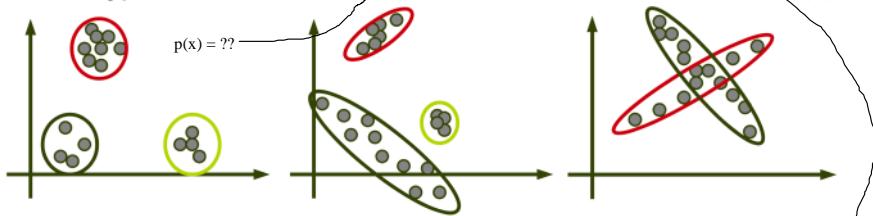
N.B. the COV can be negative, but not the VAR



We want to find a model that can give us the Gaussian density for each cluster

We want to estimate the density $p(x)$ in the graph, which is the sum of the individual density (red, gray, yellow)

Prototypical mixture model



- We want a **density** $p(x)$ of our observations $x \in \mathbb{R}^M$
- Suppose we have K clusters and let $z = k$ if x belongs to cluster k
- According to the basic rules of probability:

We can use the marginalization eq (5.11), introducing a new variable

$$p(\mathbf{x}) = \sum_{k=1}^K p(\mathbf{x}, z=k) = \sum_{k=1}^K p(\mathbf{x}|z=k)p(z=k)$$

We want to have the joint distribution, $p(\mathbf{x}) = \sum_{k=1}^K p(\mathbf{x}, z=k) = \sum_{k=1}^K p(\mathbf{x}|z=k)p(z=k)$ multiply by the density for a specific observation given cluster k

Between the observations and the cluster \rightarrow Joint distr.

Membership \rightarrow we pass from the prob of obs x , to Then using product rule $\rightarrow p(a,b) = p(a|b)p(b)$

To the prob of x to belong to a specific cluster

- If we specify $p(\mathbf{x}|z=k)$ and $p(z=k) = w_k$ we have a model
- The density for each cluster, if we are in a specific cluster, $p(\mathbf{x}|z=k)$ is a gaussian distr
 - $p(z=k)$: that's a kind of prob of a specific cluster pairing, we just parameterize With the prob of being in a specific cluster, w_k

11 DTU Compute

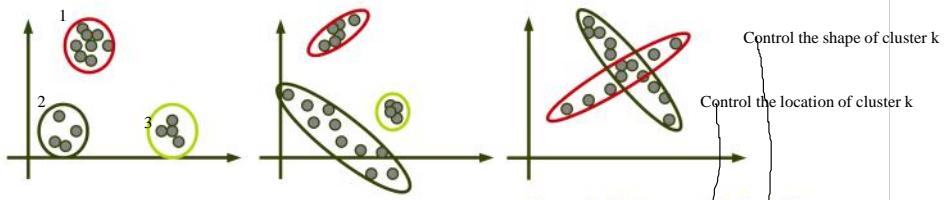
Lecture 11 16 November, 2021



Suppose we are in cluster 1, and we want to compute the density of that specific cluster ($p(\mathbf{x}|z=1)$). For every cluster we will have a specific mean value μ_k and a specific Σ_k that control the shape of the cluster



The Gaussian Mixture Model (GMM)



- Different locations $\mu_{(k)}$
- Different shape $\Sigma_{(k)}$
- Different sizes w_k

All clusters Data density \downarrow Sum of cluster specific densities assumed normal distributed \downarrow Prob $p(x|z=k)$

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{x} | \mu_{(k)}, \Sigma_{(k)})$$

s.t. $\sum_{k=1}^K w_k = 1, w_k \geq 0$ Choice we need to make in order to have a true density

12 DTU Compute

Lecture 11 16 November, 2021

W1 = 0.5, it means that cluster 1 it's dominant and it's most likely to happen

- k=3 it's a very dense cluster because it has a very low standard deviation on x and y (0.2 and 0.5)
- k=2 a variance in x and y larger, so the density on each specific point



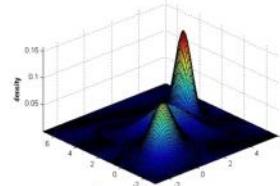
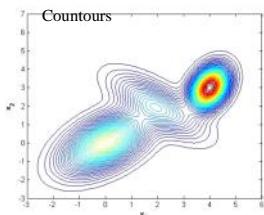
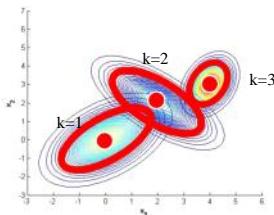
N.B: we should interpret the value w_k as the prob. of being in a cluster

GMM example

K = 3

The value w are closely related to the var value, because the normal distribution has to sum up to 1 so it kind of model the width
N.B. the higher the prob or the more datapoint of a specific cluster, the higher the weight

$$p(\mathbf{x}) = \frac{w_1}{0.5} \mathcal{N}(\mathbf{x} | \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}) + 0.2 \mathcal{N}(\mathbf{x} | \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}) + \frac{w_3}{0.3} \mathcal{N}(\mathbf{x} | \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.5 \end{bmatrix})$$



All the hills, have to sum up to 1

The counter plot should always be read relatively: the density in the red region is for sure higher than elsewhere, but we don't know the actual value.

$\mu_{(k)}$: Cluster center (prototypical example in cluster)

$\Sigma_{(k)}$: Shape of the cluster

w_k : Relative size/density of the cluster

13 DTU Compute

Lecture 11 16 November, 2021

In K=1, I can say that the variance along x1 is greater than x2 $\rightarrow \text{var}(x1) >> \text{var}(x2)$ because the cluster is much wider in the x1 direction

Quiz 01 (please answer on Piazza): GMM

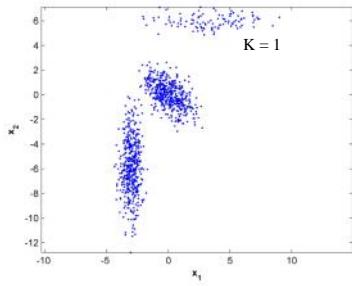


Figure 1: 1000 data observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

In Figure 1 is shown 1000 observations drawn from a density defined by a Gaussian Mixture Model (GMM) with three clusters. Suppose $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is the multivariate normal distribution, which one of the following GMM densities was used to generate the data?



Var(x1)

$$A \quad p(x) = 0.1 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) \text{Var}(x2) \\ + 0.45 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) \\ + 0.45 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix})$$

$$B \quad p(x) = 0.45 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) \\ + 0.1 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) \\ + 0.45 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix})$$

$$C \quad p(x) = 0.1 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) \\ + 0.45 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) \\ + 0.45 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix})$$

$$D \quad p(x) = 0.1 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) \\ + 0.45 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) \\ + 0.45 \cdot \mathcal{N}(\mathbf{x} | \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix})$$

E Don't know.

14 DTU Compute

Lecture 11 16 November, 2021

Sanity check time:



- Consider the Gaussian mixture model (GMM)

Eq 19.3 : in the book w_k is $\prod_{k=1}^K$

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{x} | \mu_{(k)}, \Sigma_{(k)}) \quad \text{s.t. } \sum_{k=1}^K w_k = 1, w_k \geq 0$$

- What is the value of the integral?

$$\begin{aligned} \int p(\mathbf{x}) d\mathbf{x} &= \int \sum_{k=1}^K w_k \mathcal{N}(\mathbf{x} | \mu_{(k)}, \Sigma_{(k)}) d\mathbf{x} \\ &= \sum_{k=1}^K w_k \underbrace{\int \mathcal{N}(\mathbf{x} | \mu_{(k)}, \Sigma_{(k)}) d\mathbf{x}}_1 \\ &= \sum_{k=1}^K w_k = 1 \end{aligned}$$

--> The Gaussian Mixture Model is a density function -> sum to 1 and non negative

15 DTU Compute

Lecture 11 16 November, 2021

EM algorithm is used to derive the parameter w_k , because we want to learn them from the data, μ_k and COV_k



Gaussian mixture models, EM algorithm

$p(z_n=k | x_n)$ is the prob of a specific point x_n belonging to class k

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

Computing and updating the probabilities

- Expectation** Of the model

- For each object, calculate the probability of belonging to each distribution

- Maximization** Updating the parameter of the model

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

We want to learn the parameter μ_k , w_k , and COV_k

Ex: we have a cluster membership of 1 always, it means any given point can only belong to 1 cluster

$p(z_n=k|x_n) = 1$ if the point belongs to the cluster,
Or 0 otherwise

Then $p(z_n=k|x_n)$ can only be 1 or 0

- N_k is summing the number of observations that belong to cluster k
- μ_k : calculate the mean value of the distribution
- w_k = the fraction of points out of the total that belong to cluster k
- $COV_k = \frac{1}{N_k} \sum_{n=1}^N (x_n - \mu_k)(x_n - \mu_k)^\top p(z_n=k|x_n)$

E-step

$$p(z_n=k|x_n) = \frac{w_k \mathcal{N}(x|\mu_{(k)}, \Sigma_{(k)})}{\sum_{k=1}^K w_k \mathcal{N}(x|\mu_{(k)}, \Sigma_{(k)})}$$

It's the entire distribution $p(x_n)$



This is a density, so if we are in a specific cluster, we have the mean and COV param, then the $N(\cdot)$ tells the likelihood that x_n was generated by that mean and cov

Instead, w_k is the prior: $p(z_n=k)$

In the general case:

- Gives the SOFT membership: points can belong to clusters to a specific degree (point can belong 90% to cluster 1 and 10% to cluster 2)
- In the mean and cov_k, point that have a high membership ($p(z_n=k|x_n)$) will account for more, have a higher weight

Quiz 02 (please answer on Piazza): GMM

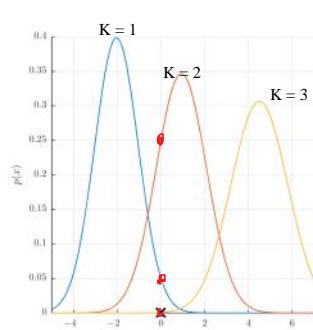


Figure 1: Mixture components in a GMM mixture model with $K = 3$

Consider a 1D GMM mixture model where each of the $K = 3$ (Gaussian) mixture components are illustrated in Figure 1 as the colored curves and the figure also shows a new observation indicated by the cross. Suppose we wish to apply the EM algorithm to this mixture model beginning with the E-step (i.e. assuming the mixture components has the means and variances indicated by Figure 1 and equal weights). According to the EM algorithm, what is the (approximate) probability the black cross is assigned to the blue (left-most) mixture component? $p(z_{x=1}|x_n) ??$

A. 0.05 it's the density of black cross for $k = 1$ but we need the probability!

B. 0.17 FIRST OF ALL : realize that the y-axis give $p(x)$ so each plot represent $p(x|k=1,2,3)$. Then we use the eq: and the 3 red points

C. 0.25 $p(z_{x=1}|x) = \frac{1/3 * 0.05}{1/3 * 0.005 + 1/3 * 0.25} = 0.17$

D. 0.02

E. Don't know. N.B. remember that the Normal distribution $N(x|\mu_k, \Sigma_k)$ is equal to $p(x|k=1)$



Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

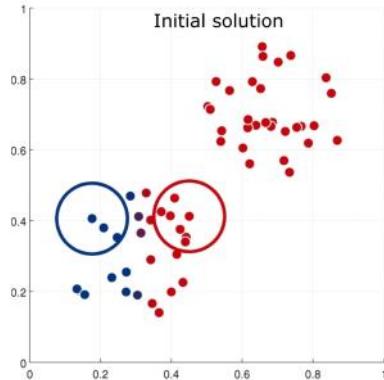
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

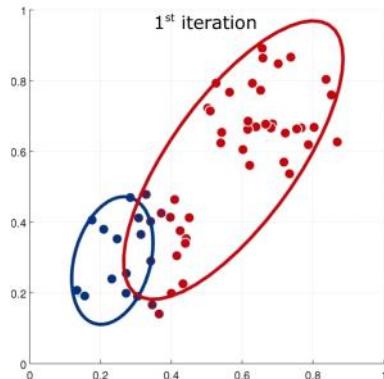
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

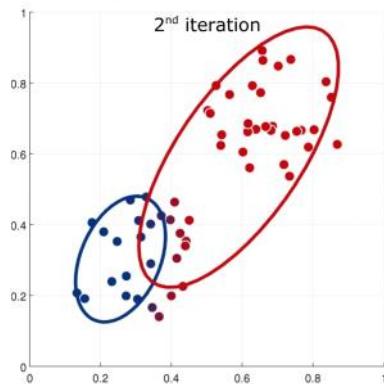
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

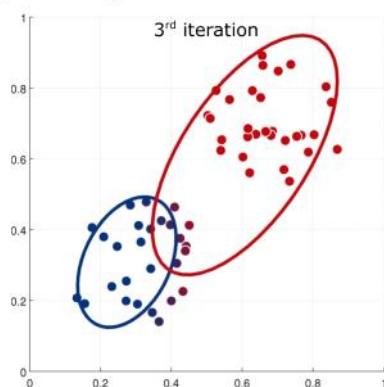
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

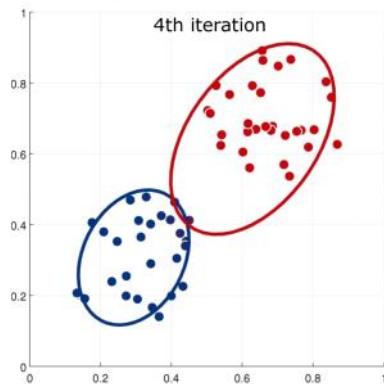
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

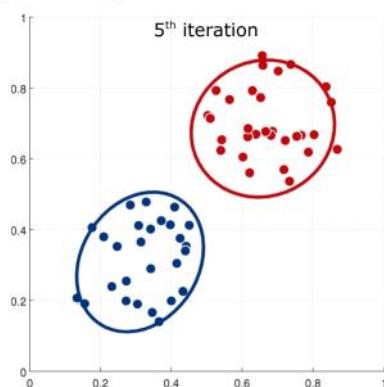
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

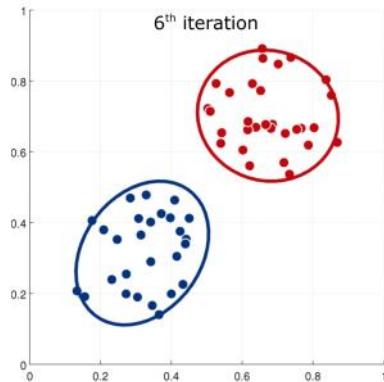
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

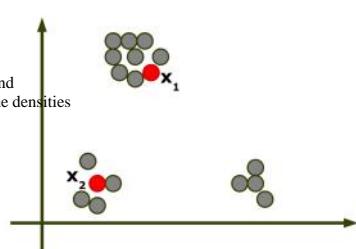
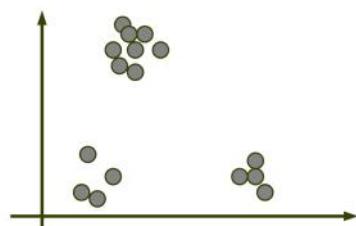
Until the parameters do not change



The covariance will be 0 -> meaning that all the densities will be located exactly at the point, bad solutions.

- Consider the data to the right with 16 observations.
 - What would ideally happen if we used a GMM with K=16 clusters to model the data?
- Imagine we have two **test observations** denoted x_1 and x_2 (red points) that are not used for training.
 - What happens to $p(x_1)$ and $p(x_2)$ if we use K=3 and K=16 clusters?

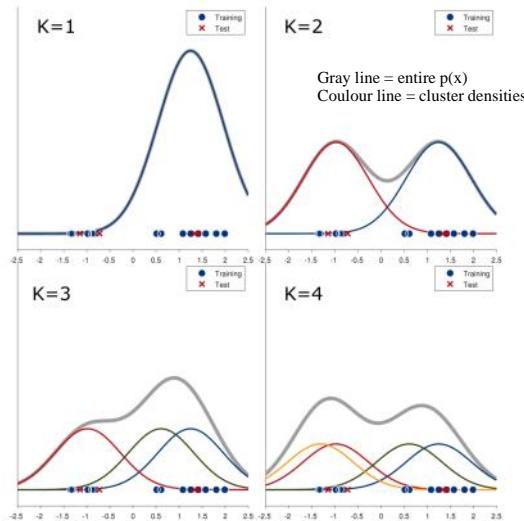
If we have 16 cluster, then the densities for the 2 test points will be 0, and Our model will be really really bad because we have test point where the densities is 0.



Mixture models

- Selecting complexity using crossvalidation

EM Initial solution



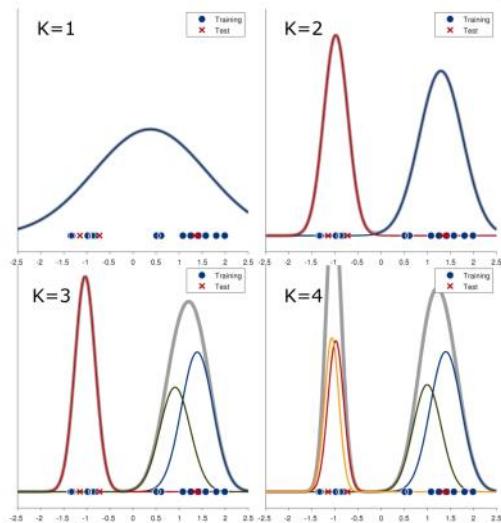
26 DTU Compute

Lecture 11 16 November, 2021

Mixture models

- Selecting complexity using crossvalidation

EM 1st iteration

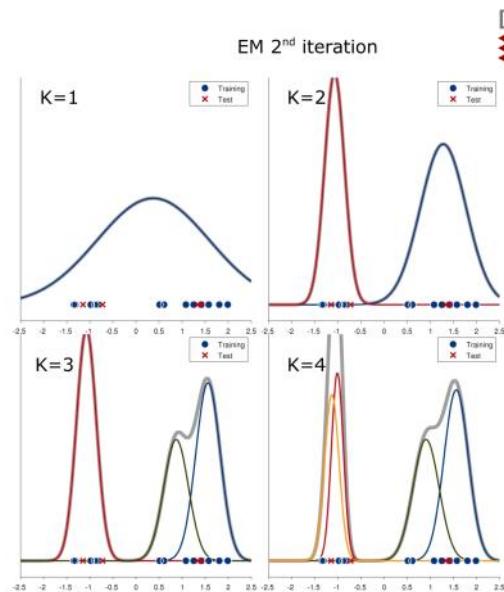


27 DTU Compute

Lecture 11 16 November, 2021

Mixture models

- Selecting complexity using crossvalidation

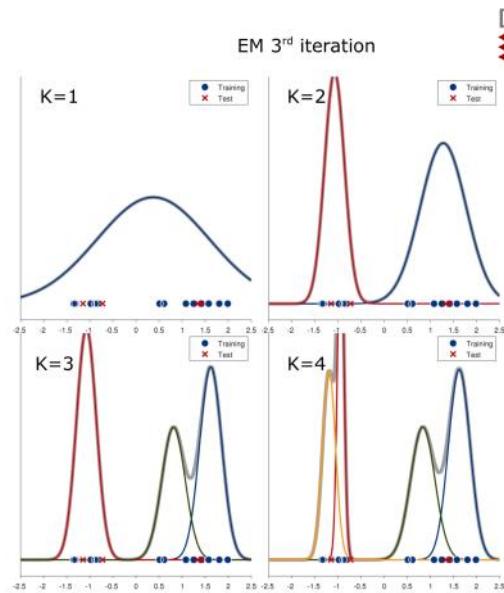


28 DTU Compute

Lecture 11 16 November, 2021

Mixture models

- Selecting complexity using crossvalidation

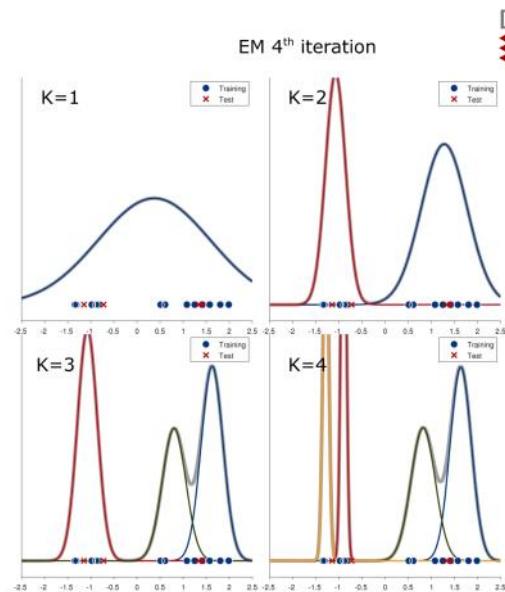


29 DTU Compute

Lecture 11 16 November, 2021

Mixture models

- Selecting complexity using crossvalidation



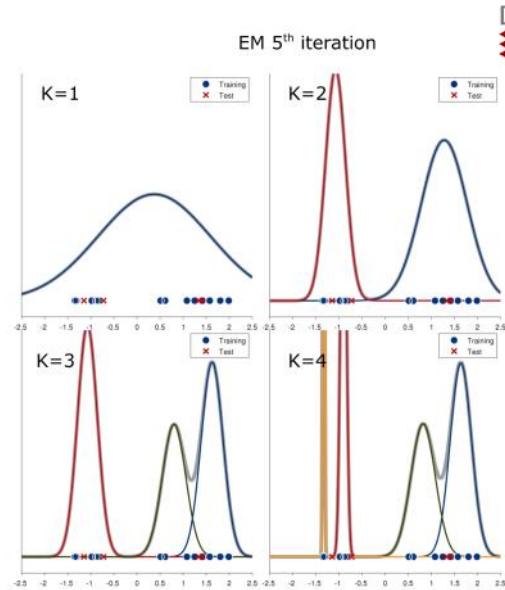
30 DTU Compute

Lecture 11 16 November, 2021

Mixture models

- Selecting complexity using crossvalidation

In k=4 we have a really high density (to avoid)
Because we have a cluster with only 1 point



31 DTU Compute

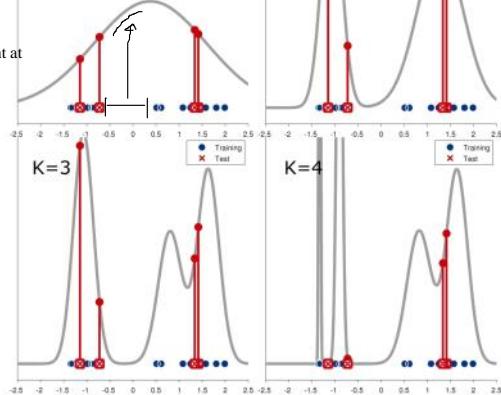
Lecture 11 16 November, 2021

Mixture models

- Selecting complexity using crossvalidation

$K=4$ is bad, because for test point we have densities = 0

$K=1$ is bad because in the space where we have no point at All, instead we have high densities



32 DTU Compute

Lecture 11 16 November, 2021

Mixture models

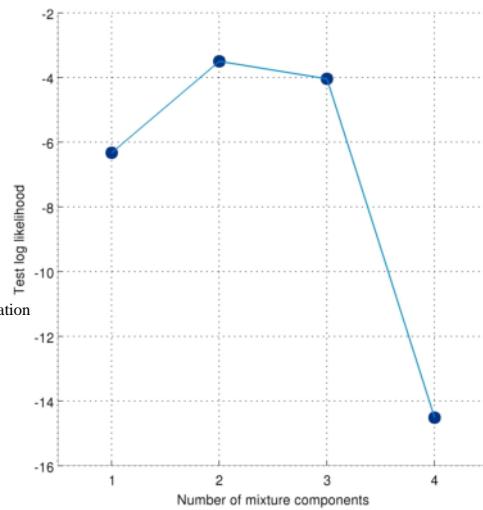
- Selecting complexity using crossvalidation

Compute the test likelihood

$\Theta = \{w_k, \mu_k, \Sigma_k\}$ our param

We compute the likelihood
 $p(x_{\text{test}}|\theta) \rightarrow \text{likelihood}$

Why we take the log of prob ?? It changes multiplication To summation, so it gives more numerical stability



33 DTU Compute

Lecture 11 16 November, 2021

K-means versus GMM

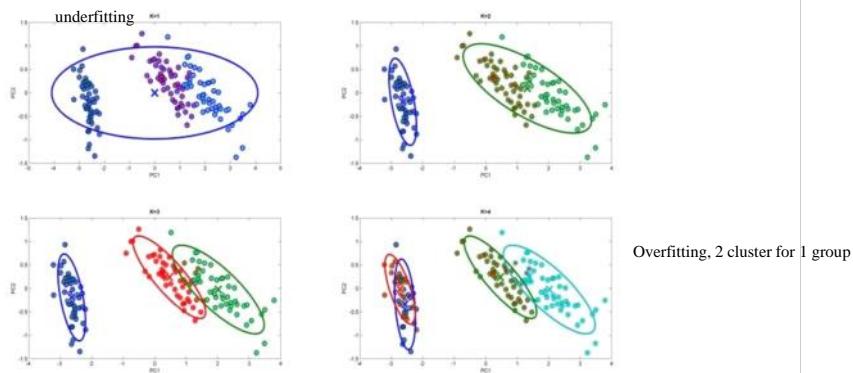
K-means

- No guarantee of optimal solution
- Does not model shape of clusters
- Does not model the size of clusters
- Difficult to assess the number of clusters to use particularly when there is no ground truth

Gaussian mixture model (GMM)

- No guarantee of optimal solution (even more local minima issues due to the additional model parameters)
- Models shape of cluster as ellipsoid
- Models the size of clusters
- Possible to estimate the number of components by cross-validation

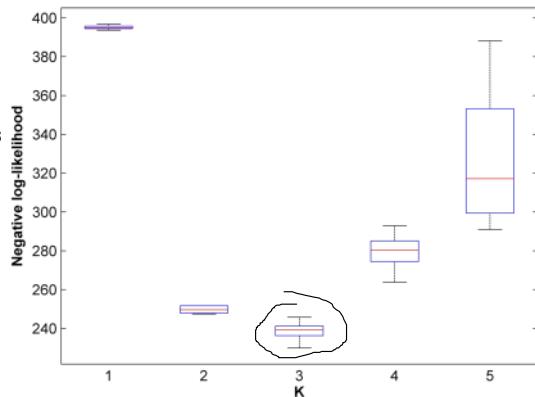
GMM on Iris data using 1,2,3 and 4 components



Recap of GMM on Iris data

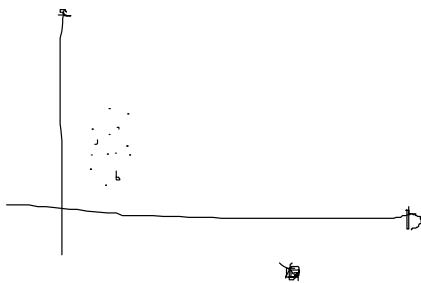
GMM 10 fold cross-validation on Iris data repeated five times where the five runs are plotted using box-plots.

The negative log: it means
That we are looking for the
lowest value as the best number
of cluster, which turn out to be 3



Anomaly detection: Definition

- Given a collection of data objects
 - Each object has associated a number of features
- Detect which objects **deviate from normal** behaviour

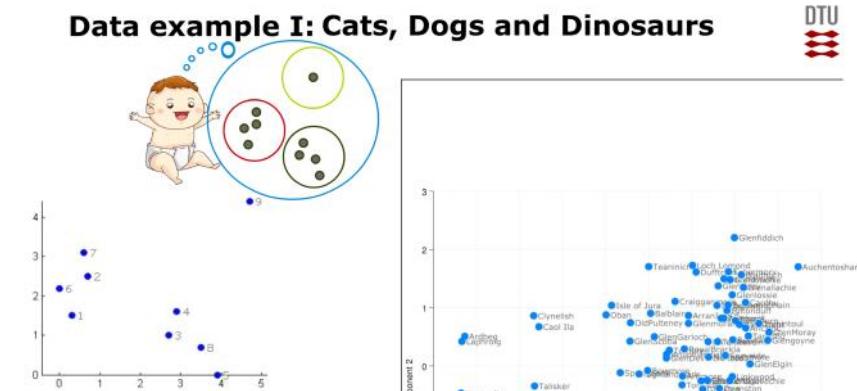


We want to detect this point(anomaly) -> using density threshold

Anomaly detection: Example

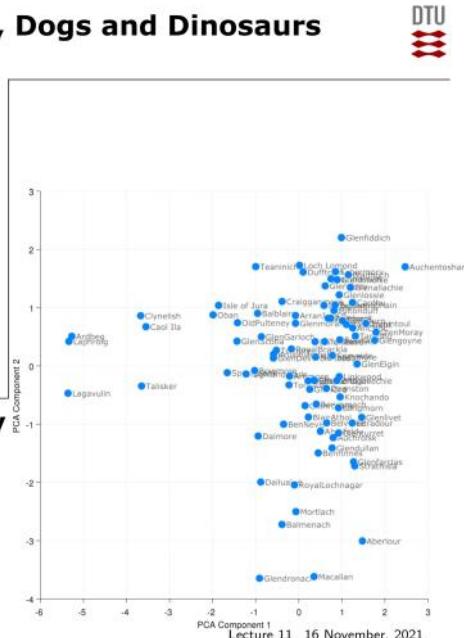
- Credit card **fraud detection**
 - Recognize dubious credit card transactions based on the transaction history of the card holder
- Network **intrusion detection**
 - Detect hacker attacks, web crawlers etc.
- **Ecosystem disturbances**
 - Detect hurricanes, floods droughts, heat waves and fires
- **Health and medicine monitoring**
 - Detect abnormal behaviour in populations and patients
- **Fault detection in industry systems**
 - Detect when a wind turbine performs poorly due to ice coating on blades
- Detection of **outliers** in data measurements
 - Remove erroneous measurements due to misreading from an instrument

Data example I: Cats, Dogs and Dinosaurs



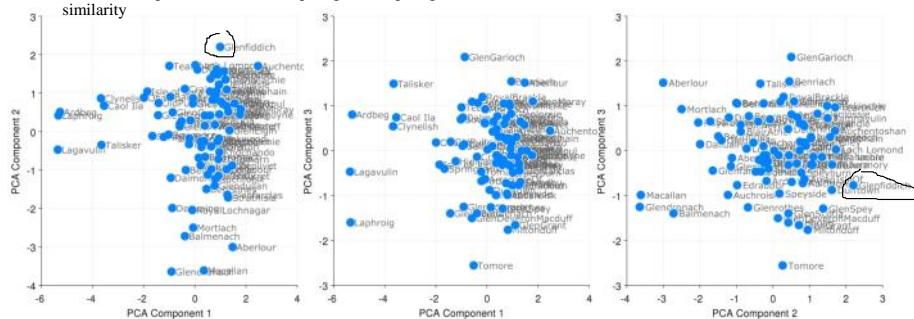
Data example II: Whisky

- 86 types of Scotch whisky
- Human ratings 1-5
- 12 taste categories
 - body, sweetness, smoky, medicinal, tobacco, honey, spicy, winey, nutty, malty, fruity, floral



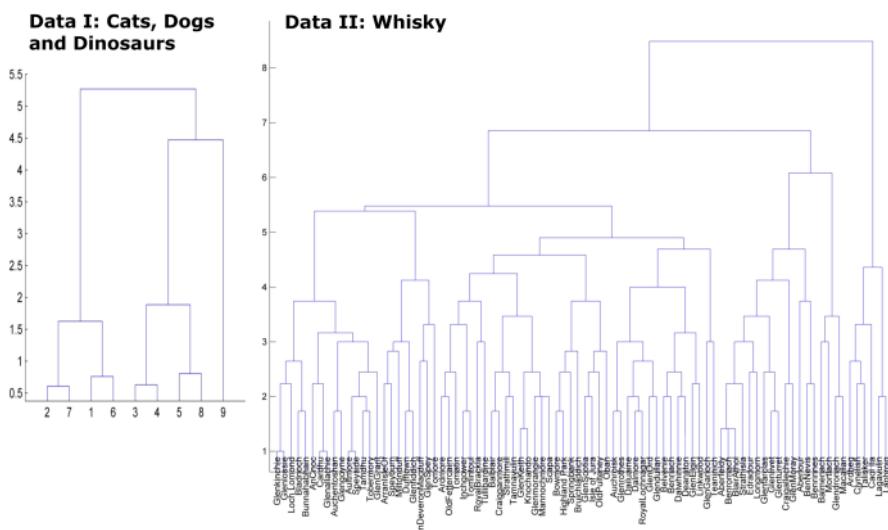
PCA plot

PCA is not a really good solution for detecting anomaly, because the point ends up in different region. PCA is instead quite good for giving information about similarity



Dendrogram

- Dendrograms can be used to visualize relative distances between the observations



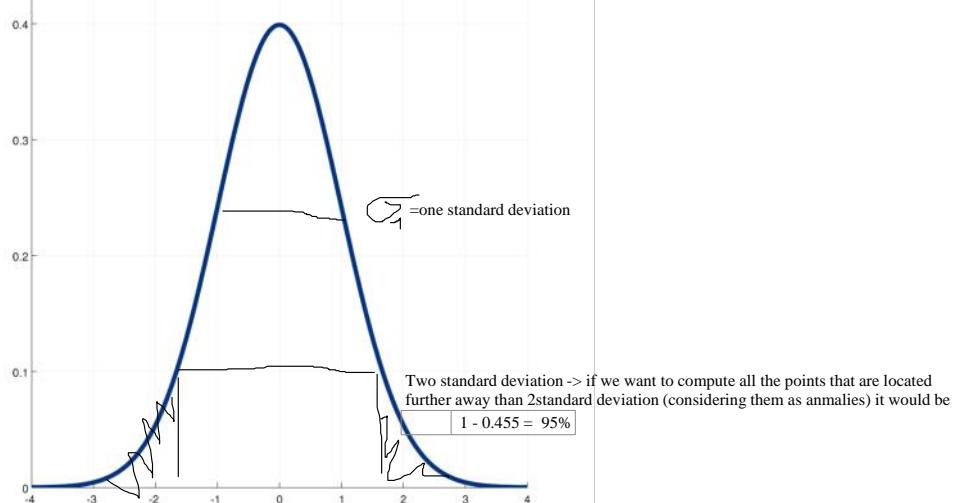
Density based techniques: Univariate normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	$p(z >c)$
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



42 DTU Compute

Lecture 11 16 November, 2021

Normal distribution

- Map attribute to standard Normal variable

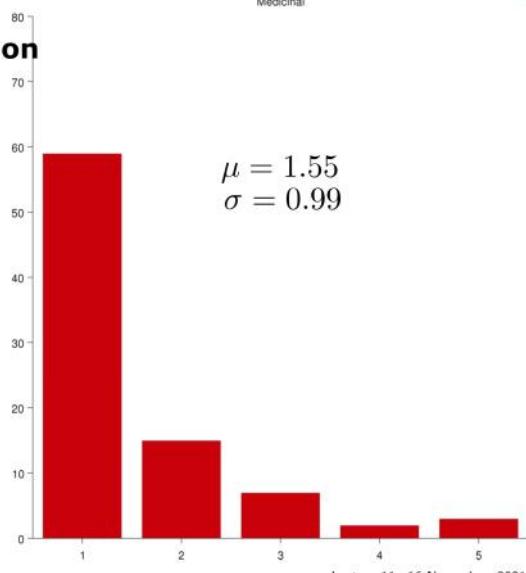
$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	$p(z >c)$
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001

$$\mu = 1.55$$

$$\sigma = 0.99$$



43 DTU Compute

Lecture 11 16 November, 2021

Normal distribution

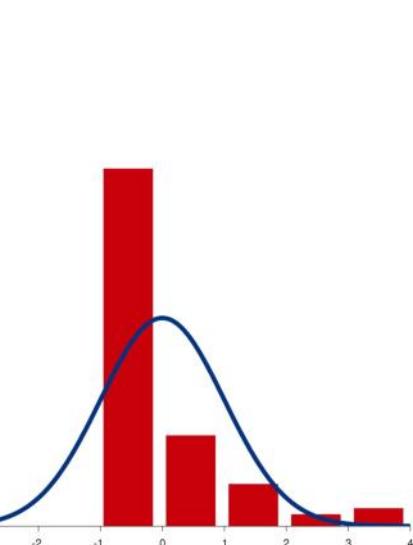
- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	$p(z > c)$
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001

Medicinal: z-score



44 DTU Compute

Lecture 11 16 November, 2021

Normal distribution

- Map attribute to standard Normal variable

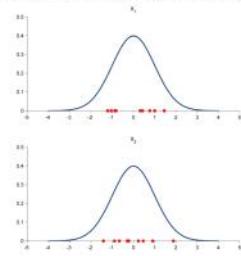
- Choose a threshold

$$z = \frac{x - \mu}{\sigma}$$

$$p(|z| > c) = 0.001$$

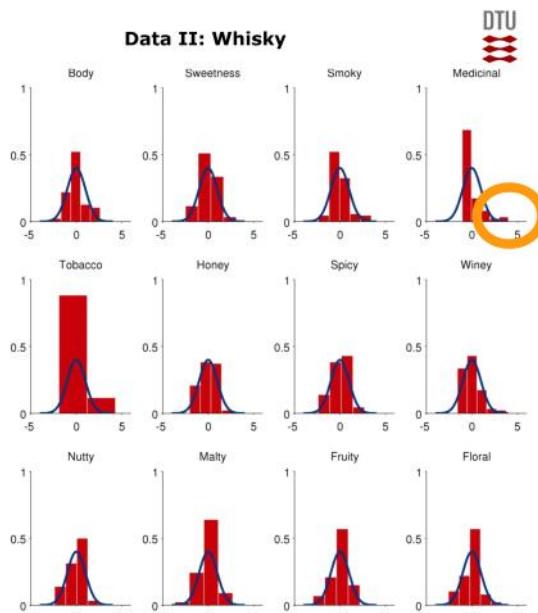
$$c = 3.2905$$

Data I: Cats, Dogs and Dinosaurs



45 DTU Compute

Data II: Whisky



Note: Assumes attributes follow a normal distribution which may not be a valid assumption!

Lecture 11 16 November, 2021

Normal distribution

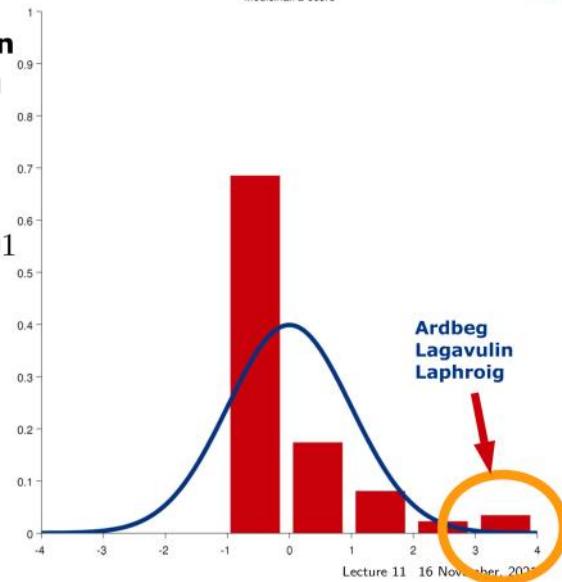
- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

$$p(|z| > c) = 0.001$$

$$c = 3.2905$$



46 DTU Compute

Lecture 11 16 November, 2021

Approaches to anomaly detection

- Density-based techniques**

- Estimate the density of data objects
- Outliers are:
 - Data objects in low density area

- We can of course use the GMM to evaluate the density of test data.**
- why not on the training data?

- Approaches we will presently also consider:**

- Kernel density estimation
- Inverse average distance to K nearest neighbours (KNN density)
- Average relative KNN density

47 DTU Compute

Lecture 11 16 November, 2021

If we don't want to have a Gaussian Mixture Model because it has a lot of param to estimate and it could be prone to local minima, we don't know how to set the threshold, then we can use Kernel Density Estimator



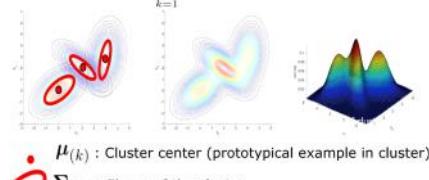
Density based techniques: Kernel Density Estimator

Recall the Gaussian Mixture Model (GMM)

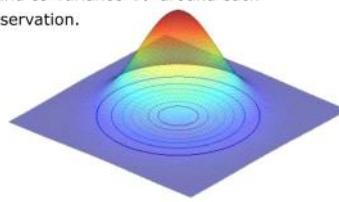
Data density Sum of cluster specific densities assumed normal distributed

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})$$

$$(s.t. \sum_{k=1}^K w_k = 1, w_k \geq 0)$$



Kernel Density estimation based on Gaussian Kernel:
 Consider the GMM and define a Gaussian with mean \mathbf{x}_n and co-variance $\sigma^2 \mathbf{I}$ around each Observation.



Let all observation weight the same, i.e. $w_n = 1/N$

$$p(\mathbf{x}) = \sum_{n=1}^N \frac{1}{N} \mathcal{N}(\mathbf{x} | \mathbf{x}_n, \sigma^2 \mathbf{I})$$

We take each data points, and on each of them we put a normal distr with a mean value of that specific datapoint and a certain std, and we do it for all our data points. > we put like a normal distr shape on top of the data points and then we add those up.

Only free parameter σ^2 !

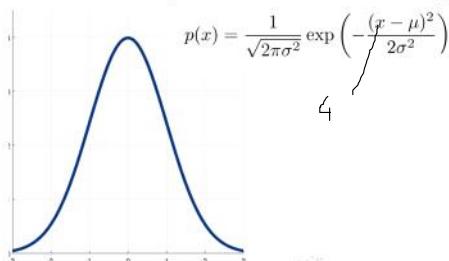
There is nothing special about the normal distribution. For a general mixture distribution p the general form of kernel density estimator is:

$$p(\mathbf{x}) = \sum_{n=1}^N \frac{1}{N} p(\mathbf{x} | \mathbf{x}_n, \theta)$$

This may be useful if \mathbf{x} is discrete or non-negative.

Calculate the Kernel Density estimator -> apply the $p(\mathbf{x})$

Piazza quiz 03: Kernel density (Spring 2013)



Consider five observations of an attribute x given by
 $X = (\text{mu1}, \text{mu2}, \text{mu3}, \dots, \text{mu5})$
 $X = \{2, 3, 5, 10, 12\}$.

Based on the five observations, what is the Gaussian kernel density estimate at $x = 4$ using $\sigma^2 = 4$?

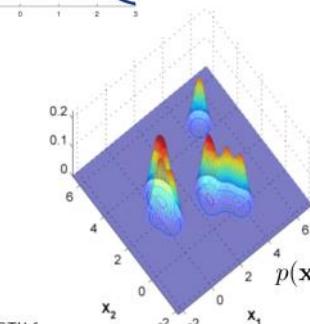
- A. $\frac{1}{\sqrt{8\pi}} \exp(-\frac{53}{4})$ For sure not A and B, because we should have a sum of 5 exponential
- B. $\frac{1}{5\sqrt{8\pi}} \exp(-\frac{53}{4})$
- C. $\frac{1}{5\sqrt{8\pi}} (\exp(-\frac{1}{2}) + 2 \cdot \exp(-\frac{1}{8}) + \exp(-\frac{9}{2}) + \exp(-8))$
- D. $\frac{1}{5\sqrt{8\pi}} (\exp(-1) + 2 \cdot \exp(-\frac{1}{4}) + \exp(-9) + \exp(-16))$
- E. Don't know.

$$p(x) = 1/5 * 1/\sqrt{2\pi\sigma^2} * (\exp(-(-2)^2/(2\cdot 4)) + \dots)$$

know.

$$\frac{1}{5} \left(\frac{1}{\sqrt{6\pi}} e^{-\frac{(4-2)^2}{2\cdot 4}} + e^{-\frac{(4-3)^2}{2\cdot 4}} + e^{-\frac{(4-5)^2}{2\cdot 4}} + e^{-\frac{(4-10)^2}{2\cdot 4}} + e^{-\frac{(4-12)^2}{2\cdot 4}} \right)$$

$$V(\mathbf{x} | \mathbf{x}_n, \sigma^2 \mathbf{I})$$



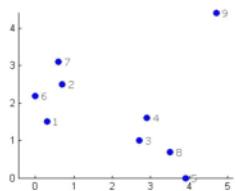
$$p(\mathbf{x}) = \sum_{n=1}^N \frac{1}{N} \mathcal{N}(\mathbf{x} | \mathbf{x}_n, \sigma^2 \mathbf{I})$$

Use this formula

How do we determine σ^2 ? Simply use cross-validation and then test each data point in the model



Data I: Cats, Dogs and Dinosaurs

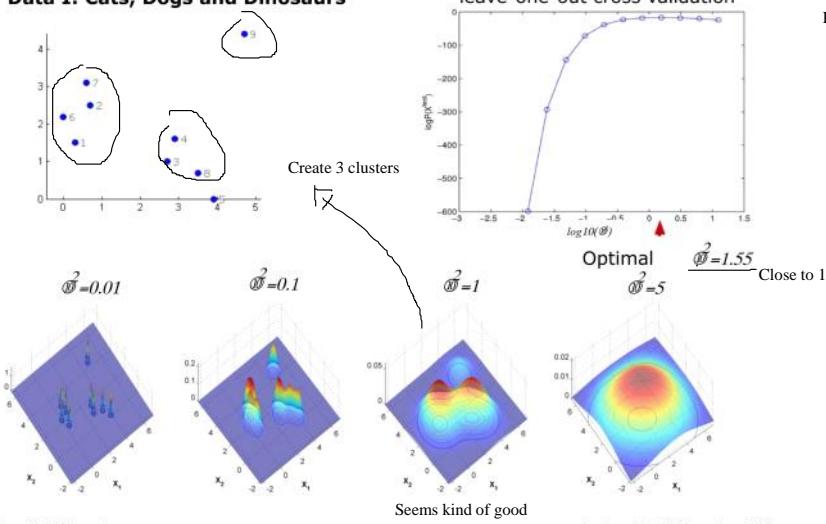


50 DTU Compute

Lecture 11 16 November, 2021

How do we determine σ^2 ? Crossvalidation!

Data I: Cats, Dogs and Dinosaurs



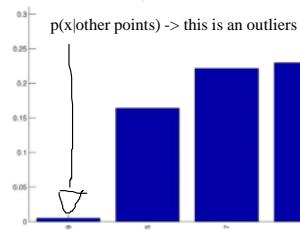
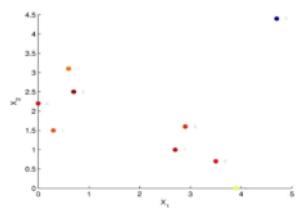
51 DTU Compute

HOW can I do anomaly distribution???

- Inspired by the cross validation, we remove each point every step and look at how the density change. If we remove point 9, for example, in sigma = 1, the small hill disappears, but if we remove point 2, it doesn't change much because there are lot of other neighbors

Lecture 11 16 November, 2021

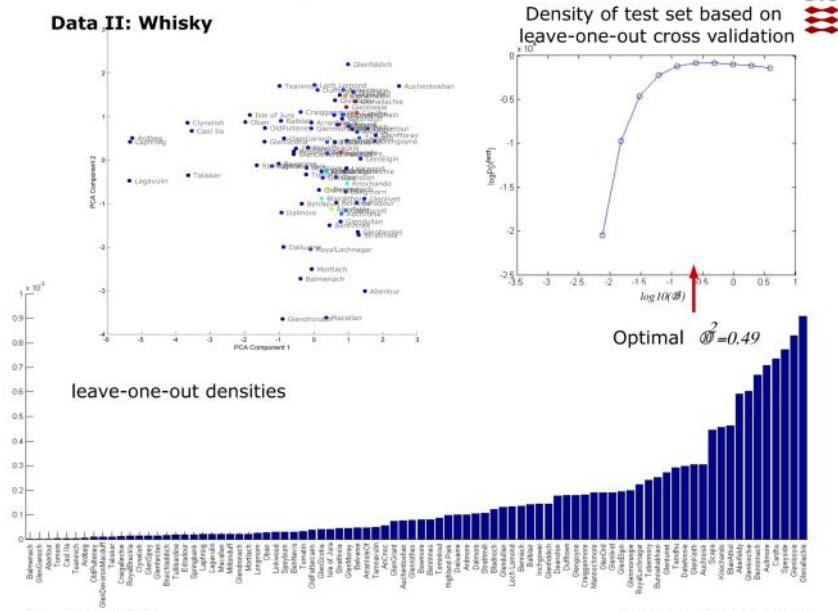
Estimated leave-one-out density evaluated at each observation



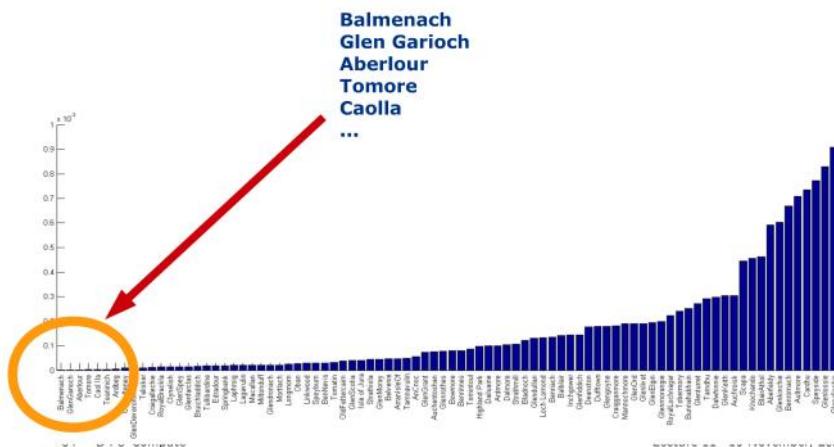
52 DTU Compute

Lecture 11 16 November, 2021

Estimated density evaluated at each observation



Data II: Whisky



We want to compute DENSITY

Inverse distance density estimation

• Distance based measure of density

- Density is inverse proportional to average distance to k nearest neighbors
- Density is low if nearest neighbors are far away

$$\text{density}_{\mathbf{X} \setminus i}(\mathbf{x}_i, K) = \frac{1}{\frac{1}{K} \sum_{\mathbf{x}' \in N_{\mathbf{X} \setminus i}(\mathbf{x}_i, K)} d(\mathbf{x}_i, \mathbf{x}')}$$

• Relative density

- Density compared to density at nearest neighbors

$$\text{ard}_{\mathbf{X}}(\mathbf{x}_i, K) = \frac{\text{density}_{\mathbf{X} \setminus i}(\mathbf{x}_i, K)}{\frac{1}{K} \sum_{\mathbf{x}_j \in N_{\mathbf{X} \setminus i}(\mathbf{x}_i, K)} \text{density}_{\mathbf{X} \setminus j}(\mathbf{x}_j, K)}$$

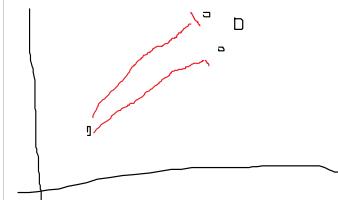
$N_{\mathbf{X}}(\mathbf{x}, K) = \{\text{The } K \text{ observations in } \mathbf{X} \text{ which are nearest to } \mathbf{x}\}$

$$\mathbf{X}_{\setminus i}^T = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \mathbf{x}_{i-2} \ \mathbf{x}_{i-1} \ \mathbf{x}_{i+1} \ \mathbf{x}_{i+2} \ \cdots \ \mathbf{x}_N]$$

55 DTU Compute

These measures are taken from:
"Introduction to Data Mining" by Pang-Ning Tan, Michael Steinbach, Vipin Kumar

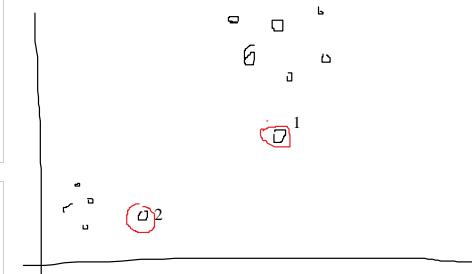
k=2 consider neighbors



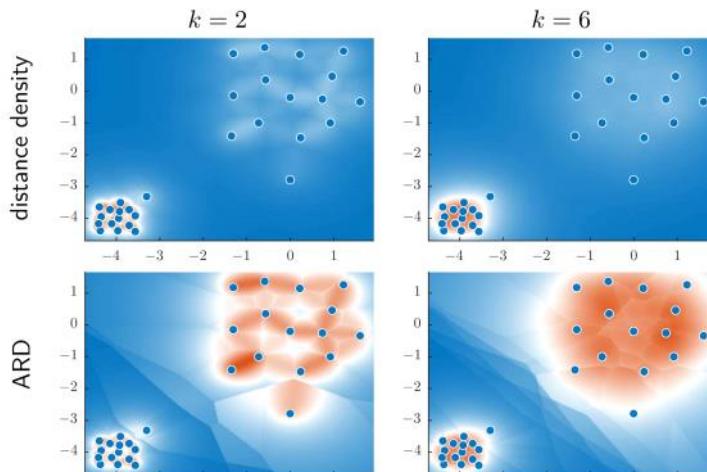
If the distance between the neighboring point is kind of large, then it should mean that the density should be very low

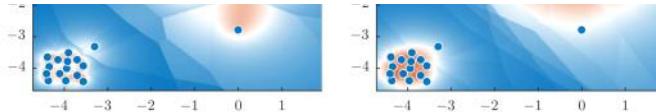
This is the idea of distance based measure of density: the density is simple 1/dist(k nearest point)

Relative density



- If we are in a cloud of points, where all the points are fairly apart, then the point that is closed to that cloud should not consider as anomaly (1)
- If we are a bit far away from a high density cloud, then the point should be considered as an anomaly(2)





Piazza quiz 4: ARD (Spring 2015)



	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
O1	0	393.5	68.1	165.4	271.8	200.6	210.9	206.1	166.3	365.0
O2	393.5	0	411.3	361.8	478.6	490.9	409.2	382.3	391.1	37.4
O3	68.1	411.3	0	119.8	208.4	136.6	152.8	154.3	111.1	387.1
O4	165.4	361.8	119.8	0	137.5	130.8	62.1	44.7	32.5	346.2
O5	271.8	478.6	208.4	137.5	0	99.0	76.8	101.0	116.4	468.5
O6	200.6	490.9	136.6	130.8	99.0	0	100.1	124.0	100.5	473.8
O7	210.9	409.2	152.8	62.1	76.8	100.1	0	29.5	45.2	396.8
O8	206.1	382.3	154.3	44.7	101.0	124.0	29.5	0	44.6	370.1
O9	166.3	391.1	111.1	32.5	116.4	100.5	45.2	44.6	0	375.1
O10	365.0	37.4	387.1	346.2	468.5	473.8	396.8	370.1	375.1	0

Table 1: Pairwise Euclidean distance between the 10 first observations in the PM10 (air pollution) dataset. Red/green observations corresponds to high/low pollution levels.

We suspect that observation O1 in Table 1 may be an outlier. In order to assess if this is the case we will calculate the average relative density (ARD) based on the distances in the table using the definitions:

$$\text{density}(\mathbf{x}, K) = \left(\frac{1}{K} \sum_{\mathbf{y} \in N(\mathbf{x}, K)} \text{distance}(\mathbf{x}, \mathbf{y}) \right)^{-1},$$

$$\text{a.r.d.}(\mathbf{x}, K) = \frac{\text{density}(\mathbf{x}, K)}{\frac{1}{K} \sum_{\mathbf{y} \in N(\mathbf{x}, K)} \text{density}(\mathbf{y}, K)},$$

where $N(\mathbf{x}, K)$ is the set of K nearest neighbors of observation \mathbf{x} and $\text{a.r.d.}(\mathbf{x}, K)$ is the average relative density of \mathbf{x} using K nearest neighbors. What is ARD for observation O1 for $K = 2$ nearest neighbors?

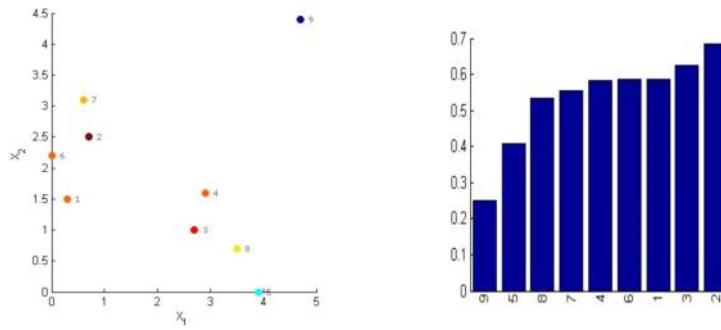
- A. 0.01
- B. 0.02
- C. 0.23
- D. 0.46
- E. Don't know.



Inverse distance density estimation

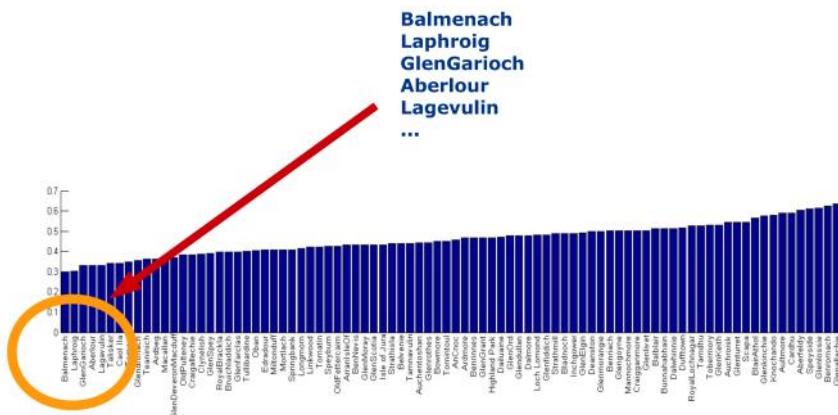
- KNN density (5 nearest neighbors)

Data I: Cats , dogs and dinosaurs



Inverse distance density estimation

- KNN density (5 nearest neighbors)

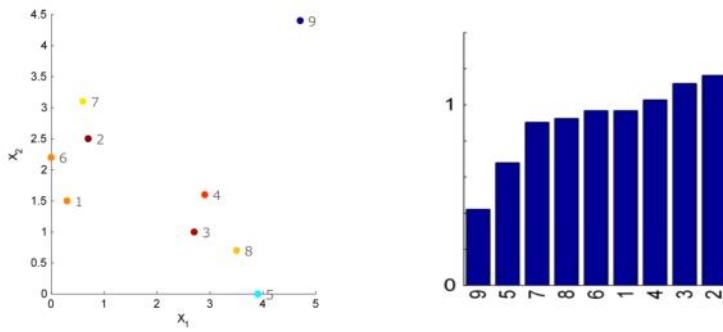


Lecture 11 16 November, 2021

Average Relative density

- Average Relative KNN density (5 nearest neighbors)

Data I: Cats , dogs and dinosaurs

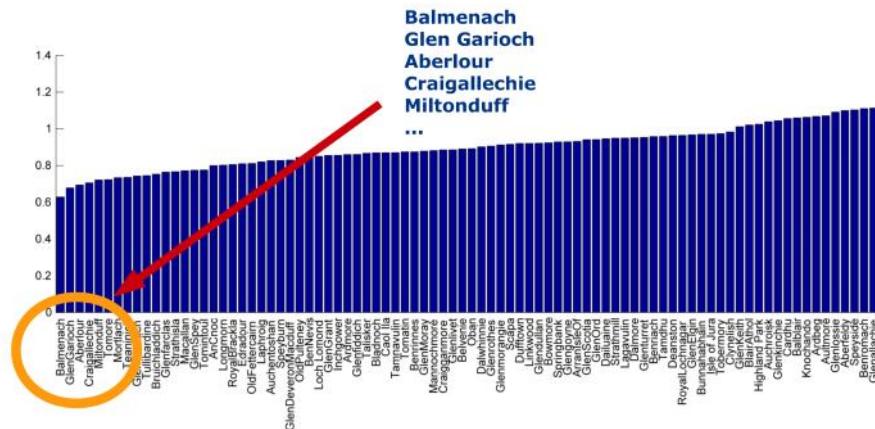


60 DTU Compute

Lecture 11 16 November, 2021

Average relative density

- Average relative KNN density (5 nearest neighbors)



61 DTU Compute

Lecture 11 16 November, 2021

Results using different methods

- Kernel Density Estimation**
 - Balmenach
 - Glen Garioch
 - Aberlour
 - Tomore
 - Caolla
- KNN density**
 - Balmenach
 - Laphroig
 - Glen Garioch
 - Aberlour
 - Lagavulin
- KNN average relative density**
 - Balmenach
 - Glen Garioch
 - Aberlour
 - Craigallechie
 - Miltonduff

Common: Balmenach, Glen Garioch,
Aberlour

62 DTU Compute

Lecture 11 16 November, 2021

Resources



<https://www.youtube.com> Nice explanation of expectation maximization for
the Guassian Mixture Model (<https://www.youtube.com/watch?v=WaKNSBeDLTw>)



02450: Introduction to Machine Learning and Data Mining

Association mining

Jes Frellsen

DTU Compute, Technical University of Denmark (DTU)

$$\int_a^b \Theta^{e^{\sqrt{17}}} \delta e^{i\pi} = -1$$

$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$

$\infty = \{2.7182818284\}$

$x^{\lambda} \approx \sum!$

DTU Compute

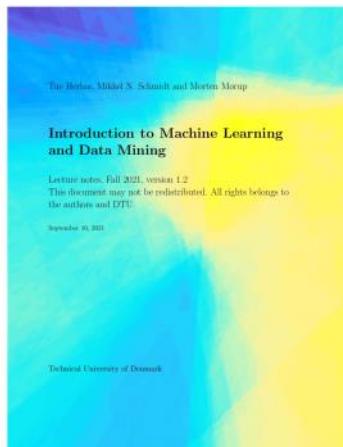
Department of Applied Mathematics and Computer Science

Today

Feedback Groups of the day:

Agnes Syshøj Lorenzen, Alessia Saccardo, Alex Zanetta, Alexander Patino Walbækken, Andreas Baltzer Skov, Andreas Christopher Theilgaard, Annachiara Rossi, Anne Marie Nørrelykke Rossen, Antarlina Mukherjee, Benjamin Werner-Griesau, Bianca Zamora, Carolina Lopez Olmos, Christian Ole Nielsen, Christopher Høeg W ejendorp, Christopher Emil Carlsen Wærenskjold Rosenørn, Chuansheng Liu, Edgars Treimanis, Erik Priest, Frederik Kirkeby Schjerning, Helene Brasch Lind Petersen, Henriette Clara Sophie Rilling, Huiyu Lu, Jack William Marshall Rose, Joana Malvar Fonseca Vilela, Johanne Skotte Steen-Hansen, Junesoo Shin, Kami Reza, Karl Meisner-Jensen, Karthik Sivakumar, Kitti Kovács, Lars Thor Sørensen, Lucía Carmen Marianne Vanhollebeke, Luis Gideon Pieschel, Maltse Emil Skytte, Matilde Uth, Mattias Erik Tammi, Max-Emil Scotten, Michelle Kirkmand Sølløk, Neha Teutloff, Nikolaj Bach Meineche, Niels Thonemann, Oskar Florian Hövenhoff Kristoffersen, Pablo Sarabia Fraile, Paweł Jarzynski, Rasmus Johansen Rieneck, Rasmus Linde Mønsted, Rebecca Viuff, Rebekka Wätzold Høgh Madsen, Rune Møller Nedergaard, Sandra Ýrr Sonjudóttir, Santiago Maldonado Hernandez, Sarah Lou, Simon Gerdes Pontoppidan, Simon Pontoppidan, Spyridon Vlachospyros, Thomas Rathsch Strange

Reading material: Chapter 21



Lecture 12 23 November, 2021

Lecture Schedule

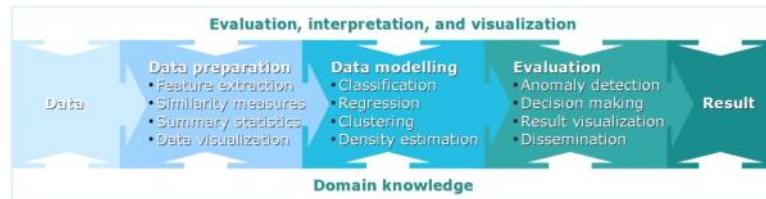


- ① Introduction
31 August: C1
Data: Feature extraction, and visualization
- ② Data, feature extraction and PCA
7 September: C2, C3
- ③ Measures of similarity, summary statistics and probabilities
14 September: C4, C5
- ④ Probability densities and data visualization
21 September: C6, C7
Supervised learning: Classification and regression
- ⑤ Decision trees and linear regression
28 September: C8, C9
- ⑥ Overfitting, cross-validation and Nearest Neighbor
5 October: C10, C12 (Project 1 due before 13:00)
- ⑦ Performance evaluation, Bayes, and Naive Bayes
12 October: C11, C13
- ⑧ Artificial Neural Networks and Bias/Variance
26 October: C14, C15
- ⑨ AUC and ensemble methods
2 November: C16, C17
Unsupervised learning: Clustering and density estimation
- ⑩ K-means and hierarchical clustering
9 November: C18
- ⑪ Mixture models and density estimation
16 November: C19, C20 (Project 2 due before 13:00)
- ⑫ Association mining
23 November: C21
Recap
- ⑬ Recap and discussion of the exam
30 November: C1-C21

Online help: Forum on DTU Learn
Videos of lectures: <https://video.dtu.dk>
Streaming of lectures: Zoom (link on DTU Learn)

Lecture 12 23 November, 2021

3 DTU Compute



Learning Objectives

- Calculate support and confidence of association rules
- Describe the Apriori algorithm for association mining and how it is used for efficient estimation of association rules

4 DTU Compute

Lecture 12 23 November, 2021

Association rule discovery: Definition

- Given a set of **records**
 - Each containing a number of **items from a set**
- Goal:** Produce dependency rules
 - Predict the occurrence of an item based on occurrences of other items

Association Mining

mining association rules 

About 1,850,000 results (0,05 sec)

[PDF] Fast algorithms for mining association rules
R.Agrawal, R.Srikant - Proc. 20th int. conf. very large data bases, VLDB, 1994 - it.uu.se
We consider the problem of discovering **association rules** between items in a large database of sales transactions. We present two new algorithms for solving this problem that are fundamentally different from the known algorithms. Experiments with synthetic as well as real ...
☆ 59 Cited by 26110 Related articles All 115 versions 80

Mining association rules between sets of items in large databases
R.Agrawal, T.Imielinski, A.Swami - Proceedings of the 1993 ACM ..., 1993 - dl.acm.org
We are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. We present an efficient algorithm that generates all significant **association rules** between items in the database. The algorithm incorporates ...
☆ 59 Cited by 23230 Related articles All 39 versions

An effective hash-based algorithm for mining association rules
JS.Park, MS.Chen, PS.Yu - Acm sigmod record, 1995 - dl.acm.org
In this paper, we examine the issue of **mining association rules** among items in a large database of sales transactions. The **mining of association rules** can be mapped into the problem of discovering large itemsets where a large itemset is a group of items which ...
☆ 59 Cited by 2465 Related articles All 18 versions

Source: Google Scholar (November, 2020)

Association rule discovery: Example

Market basket analysis

Training set	Rules discovered
1. {Bread, Soda, Milk} 2. {Beer, Bread} 3. {Beer, Soda, Diaper, Milk} 4. {Beer, Bread, Diaper, Milk} 5. {Soda, Diaper, Milk}	{Milk} \rightarrow {Soda} {Diaper, Milk} \rightarrow {Beer}

Market basket data

- Representation as

Transaction table

ID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

Data matrix

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Association analysis, rules and support



- **Itemset** What a customer bought
 - For example {Bread, Soda, Milk}, {Milk, Diaper}, {}

- **Support** for an itemset **X**
 - Percentage of transactions that contain **X**

Count how many times the itemset {milk, diaper} is listed

- **Association rule** Express : if i bought X, how likely is i also buy Y??
 - Expression of the form: **X** \rightarrow **Y**
where **X** and **Y** are disjoint item sets

- **Support** for an association rule **X** \rightarrow **Y**
 - Percentage of transactions that contain **X** \cup **Y**

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} = P(X, Y) \quad \text{It's a joint probability}$$

Piazza quiz 1: Support (Spring 2018)



	x_1^L	x_2^L	x_3^L	x_4^L	x_5^L	x_6^L	x_7^L	x_8^L	x_9^L	x_{10}^L	x_1^H	x_2^H	x_3^H	x_4^H	x_5^H	x_6^H
O1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
O2	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
O3	1	0	0	1	1	0	1	0	1	0	1	0	1	0	1	0
O4	1	0	1	0	1	0	0	1	0	1	1	0	1	0	1	0
O5	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0
O6	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
O7	0	1	1	0	1	0	0	1	0	1	0	1	0	1	0	1
O8	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1
O9	0	1	0	1	1	0	1	0	0	1	0	1	0	1	0	1
O10	1	0	0	1	0	1	0	1	0	1	1	0	1	1	0	0

Table 1: The ten first observations of the airline safety dataset binarized considering the attribute x_1 – x_6 .

We consider a dataset of airline safety binarized according to the median value. Values below median is referred to with the superscript *L* and above the median value using the superscript *H*. In Table 1 is

given the first 10 observations O1–O10. Consider the association rule:

$$\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}.$$

What is the support of the rule?

- A. 0.0 %
- B. 20.0 %
- C. 66.7 %
- D. 100.0 %
- E. Don't know.

Association analysis, confidence



- **Itemset**

- For example {Bread, Soda, Milk}, {Milk, Diaper}, {}

- **Support** for an itemset **X**

- Percentage of transactions that contain **X**

- **Association rule**

- Expression of the form: **X** \rightarrow **Y**
where **X** and **Y** are disjoint item sets

- **Support** for an association rule **X** \rightarrow **Y**

- Percentage of transactions that contain **X** \in **Y**

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} = P(X, Y)$$

- **Confidence** for an association rule **X** \rightarrow **Y** If I contain **X**, what's the chance that I contain **Y**??

- Percentage of transactions containing **X** that also contain **Y**

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{P(Y, X)}{P(X)} = P(Y|X)$$

We want to identify items that have a high support and confidence,

11 DTU Compute because with high support means that lot of customers bought that combination, and rules with high confidence means that if I bought **X**, it's also likely to buy **Y** Lecture 12 23 November, 2021

Piazza quiz 2: Confidence (Spring 2018)



X	x_1^L	x_2^H	x_2^L	x_3^H	x_3^L	x_3^H	x_4^L	x_4^H	x_5^L	x_5^H	x_6^L	x_6^H
Q1	1	0	1	0	1	0	1	0	1	0	1	0
Q2	0	1	0	1	0	1	0	1	0	1	0	1
Q3	1	0	0	1	1	0	1	0	1	0	1	0
Q4	1	0	1	0	1	0	0	1	0	1	1	0
Q5	0	1	1	0	1	0	1	0	1	0	1	0
Q6	0	1	0	1	0	1	0	1	0	1	0	1
Q7	0	1	1	0	1	0	0	1	0	1	0	1
Q8	1	0	1	0	1	0	1	0	0	1	0	1
Q9	0	1	0	1	1	0	1	0	0	1	0	1
Q10	1	0	0	1	0	1	0	1	0	1	1	0

We again consider the airline safety data and the rule
 $\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}$.

What is the confidence of the rule?

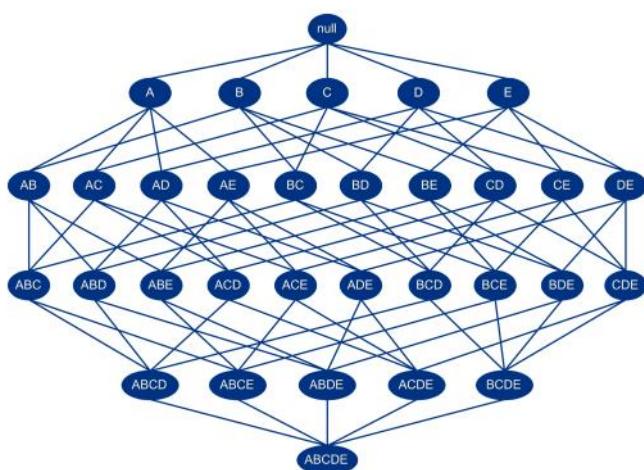
- A. 0.0 %
- B. 20.0 %
- C. 60.7 %
- D. 100.0 %
- E. Don't know.

Table 1: The ten first observations of the airline safety dataset binarized considering the attribute $x_1 \dots x_6$.

Association rule mining

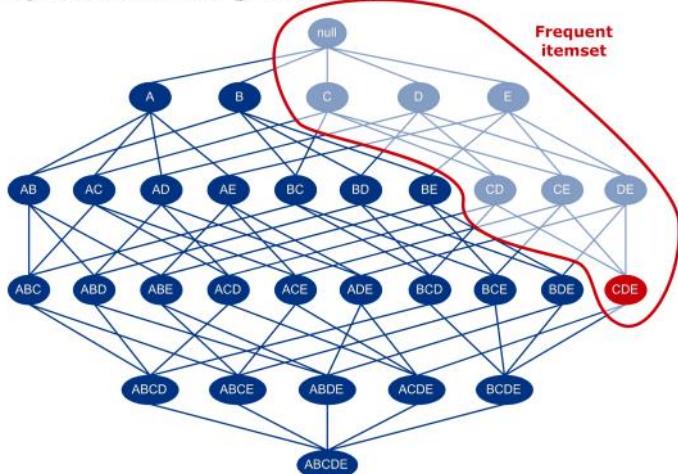
- Find all association rules that have We look for rules with High confidence or high support setting threshold
 - **Support $\geq minsup$**
 - **Confidence $\geq minconf$**
- Approach
 - **Frequent itemset generation**
 - Generate a list of all **itemsets** with **Support $\geq minsup$**
 - **Association rule generation**
 - Generate all **association rules** with **Confidence $\geq minconf$**

Frequent itemset generation



How many different itemsets can be created for a problem with a total of D items? 2^D

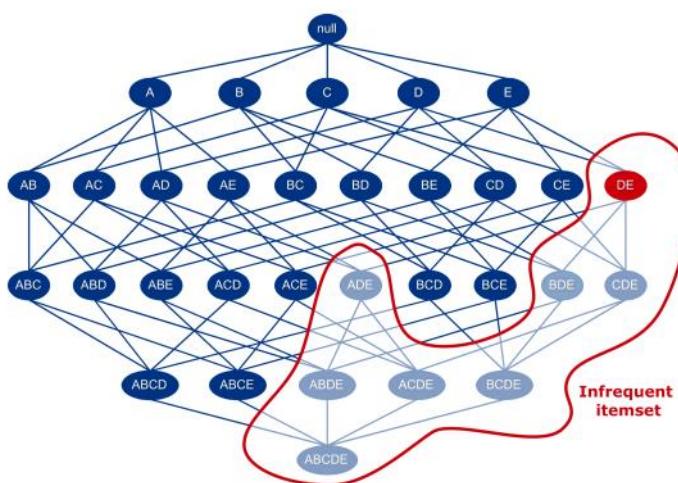
Frequent itemset generation



Downwards-closure
property

If an itemset is frequent, then all of its subsets must also be frequent

Frequent itemset generation



If an itemset is infrequent, then all of its supersets must also be infrequent

We have a transaction table, and we want to find all the itemset that have a support greater than ϵ .

- 1- Start from the single itemset, and rules out the infrequent one. If A is infrequent then all the itemset that contain A are infrequent as well.
- 2- Takes the single-frequent items and add 1 item and if it is not frequent, remove it

The Apriori Algorithm

Algorithm 8: Apriori algorithm

```

Find all 1-itemsets
1: Given N transactions and let  $\epsilon > 0$  be the minimum support count
2:  $L_1 = \{\{j\} | \text{supp}(\{j\}) \geq \epsilon\}$ 
3: for  $k = 2, \dots, M$  and  $L_k \neq \emptyset$  do
4:    $C'_k = \{s \cup \{j\} | s \in L_{k-1}, j \notin s\}$ 
5:   Set  $C_k = C'_k$ 
6:   for each  $c \in C'_k$  do
7:     for each  $s \subset c$  such that  $|s| = k-1$  do
8:       if  $s$  is not frequent, i.e.  $s \notin L_{k-1}$  then
9:          $C_k = C_k \setminus \{c\}$  (Remove c from  $C_k$ )
10:      end if
11:    end for
12:  end for
13:   $L_k = \{c | c \in C_k, \text{supp}(c) \geq \epsilon\}$  (compute support)
14: end for
15:  $L_1 \cup L_2 \cup \dots \cup L_k$  are then all frequent itemsets

```

Remove all the generated itemsets for which subsets are not part of the k-1-itemsets

Keep remaining k-itemsets with enough support.

Output all frequent itemsets



Piazza quiz 3: A-priori (Fall 2018)

Running Apriori algo.

Columns are feature1,2,... and rows are itemsets

The matrix tells us that the itemset containing feature1 and feature2 is frequent, so it has support at least $\geq \epsilon$

The matrix shows all the frequent itemset with at least 2 features

We will consider a binary dataset consisting of the $M = 6$ features $f_1, f_2, f_3, f_4, f_5, f_6$. We wish to apply the Apriori algorithm to find all itemsets with support greater than $\epsilon = 0.15$. Suppose at iteration $k = 3$ we know that:

$$L_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

dide itemsets C'_3 , and then rule out some of them using the downwards-closure principle thereby saving many (potentially costly) evaluations of support. Suppose L_2 is given as above, which of the following itemsets does the Apriori algorithm *not* have to evaluate the support of?

- A. $\{f_2, f_3, f_4\}$
- B. $\{f_1, f_2, f_6\}$
- C. $\{f_2, f_3, f_6\}$
- D. $\{f_1, f_3, f_4\}$ [1 0 1 0 0 0] is not frequent in L_2
- E. Don't know.

Recall the key step in the Apriori algorithm is to construct L_3 by first considering a large number of can-

Which itemset is not going to be part of the candidates in line 13



Solution:

Recall the Apriori algorithm obtain L_3 from L_2 in three steps. First, the Apriori algorithm construct C'_3 by, for each itemset I in L_2 , loop over all items not already in I and consider all such combinations where I is enlarged by a single item as a candidate itemset in C'_3 . Specifically we get:

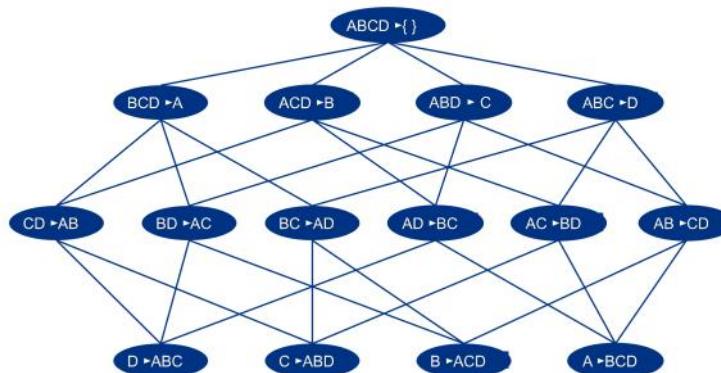
$$C'_3 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The downwards closure principle is then applied by removing itemset I in C'_3 if I contains a subset of 2 items not found in L_2 . We thereby get:

$$C_3 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Finally, L_3 is constructed from C_3 by removing those itemsets with a support lower than ϵ . Thus, the itemsets we don't have to compute support from are those itemsets found in C_3 but not in C_2 , or as an even simpler criteria, those which have a subset of size 2 not found in L_2 . This rules out all options except D.

Association rule generation

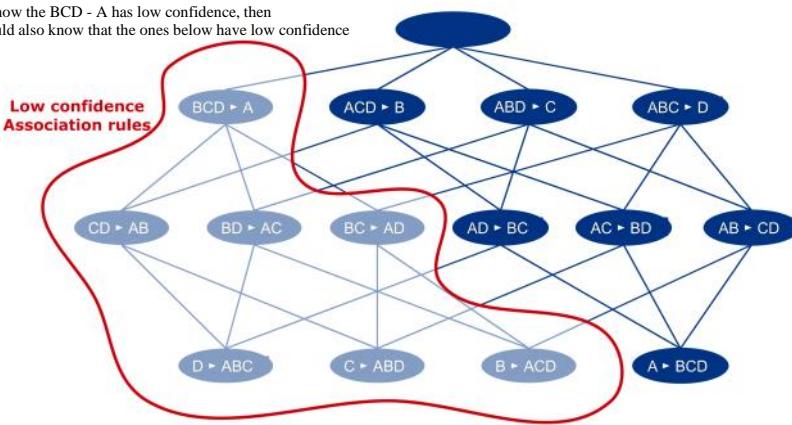




What's the confidence of $BCD \rightarrow A$? = $\text{supp}(ABCD) / \text{supp}(BCD)$
 Then, what's the confidence of $CD \rightarrow AB$ = $\text{supp}(ABCD) / \text{supp}(CD)$.
 We know that the first quantity is low, so we say something about the second one?
 \rightarrow downwards closure property : $\text{supp}(C,D) \geq \text{supp}(B,C,D)$ (if I bought BCD than I also bought CD, then the prob if CD is higher) \rightarrow the denominator of the 2° confidence is greater, then the confidence of the second one will be smaller

Association rule generation

If I know the $BCD - A$ has low confidence, then
 I should also know that the ones below have low confidence



Due to the downward closure property, the smaller itemset the higher the support

Results for market basket example

Itemset	Support	Association rule	Support	Confidence
Milk	80%	{ } \rightarrow Milk	80%	80%
Bread	60%	Soda \rightarrow Milk	60%	100%
Soda	60%	Diaper \rightarrow Milk	60%	100%
Beer	60%	Soda, Diaper \rightarrow Milk	40%	100%
Diaper	60%	Beer, Diaper \rightarrow Milk	40%	100%
Diaper Milk	60%	Beer, Milk \rightarrow Diaper	40%	100%
Soda Milk	60%			
Bread Beer	40%			
Bread Milk	40%			
Soda Diaper	40%			
Beer Diaper	40%			
Beer Milk	40%			
Soda Diaper Milk	40%			
Beer Diaper Milk	40%			

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

- How can we do association mining for continuous data?

Yes, we can binarized the data

	Attribute 1	Attribute 2	Attribute 3
	0.3689	0.9827	0.6999
	0.4607	0.7302	0.6385
	0.9816	0.3439	0.0336
	0.1564	0.5841	0.0688
	0.8555	0.1078	0.3196
	0.6448	0.9063	0.5309
	0.3763	0.8797	0.6544
	0.1909	0.8178	0.4076
	0.4283	0.2607	0.8200
X=	0.4820	0.5944	0.7184
	0.1206	0.0225	0.9686
	0.5895	0.4253	0.5313
	0.2262	0.3127	0.3251
	0.3846	0.1615	0.1056
	0.5830	0.1788	0.6110
	0.2518	0.4229	0.7788
	0.2904	0.0942	0.4235
	0.6171	0.5985	0.0908
	0.2653	0.4709	0.2665
	0.8244	0.6959	0.1537

Binarize data according to percentiles

AttributeName	Attribute 1	Attribute 2	Attribute 3	AttributeNameBins	Attribute 1 0-50 %	Attribute 1 50-100 %	Attribute 2 0-50 %	Attribute 2 50-100 %	Attribute 2 33.3-66.7 %	Attribute 2 66.7-100 %	Attribute 3 0-50 %	Attribute 3 50-100 %
	0.3689	0.9827	0.6999		1	0	0	0	1	0	0	1
	0.4607	0.7302	0.6385		0	1	0	0	1	0	1	1
	0.9816	0.3439	0.0336		0	1	0	1	0	1	0	0
	0.1564	0.5841	0.0688		1	0	0	1	0	1	0	0
	0.8555	0.1078	0.3196		0	1	1	0	0	1	0	0
	0.6448	0.9063	0.5309		0	1	0	0	1	0	1	1
	0.3763	0.8797	0.6544		1	0	0	0	1	0	1	1
	0.1909	0.8178	0.4076		1	0	0	0	1	1	1	0
	0.4283	0.2607	0.8200		0	1	1	0	0	0	0	1
X=	0.4820	0.5944	0.7184	Xbinary=	0	1	0	1	0	0	0	1
	0.1206	0.0225	0.9686		1	0	1	0	0	0	0	1
	0.5895	0.4253	0.5313		0	1	0	1	0	0	0	1
	0.2262	0.3127	0.3251		1	0	1	0	0	1	0	0
	0.3846	0.1615	0.1056		1	0	1	0	0	0	1	0
	0.5830	0.1788	0.6110		0	1	1	0	0	0	0	1
	0.2518	0.4229	0.7788		1	0	0	1	0	0	0	1
	0.2904	0.0942	0.4235		1	0	1	0	0	1	0	0
	0.6171	0.5985	0.0908		0	1	0	1	0	1	0	0
	0.2653	0.4709	0.2665		1	0	0	1	0	1	0	0
	0.8244	0.6959	0.1537		0	1	0	0	1	1	1	0

Recap of association rule discovery on Iris data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	Sepal Length	Sepal Width	Petal Length	Petal Width	Type											
1	5.1	3.5	1.4	0.2	Iris-setosa											
2	4.9	3.0	1.4	0.2	Iris-setosa											
3	4.7	3.2	1.3	0.2	Iris-setosa											
4	4.6	3.1	1.5	0.2	Iris-setosa											
5	5.0	3.6	1.4	0.2	Iris-setosa											
6	5.4	3.9	1.7	0.4	Iris-setosa											
7	4.6	3.4	1.4	0.3	Iris-setosa											
8	5.0	3.4	1.5	0.2	Iris-setosa											
9	4.5	2.3	1.3	0.3	Iris-setosa											
10	4.5	2.3	1.3	0.3	Iris-setosa											
11	4.9	3.1	1.5	0.2	Iris-setosa											
12	4.7	3.2	1.3	0.2	Iris-setosa											
13	5.8	4.0	1.2	0.2	Iris-setosa											
14	5.1	3.5	1.4	0.2	Iris-setosa											
15	5.9	3.0	1.5	0.2	Iris-setosa											
16	6.0	3.4	1.4	0.2	Iris-setosa											
17	5.4	3.9	1.5	0.4	Iris-setosa											
18	7.0	3.2	4.7	1.4	Iris-versicolor											
19	6.4	3.0	4.3	1.3	Iris-versicolor											
20	6.9	3.1	4.4	1.4	Iris-versicolor											
21	5.5	2.3	4.0	1.3	Iris-versicolor											
22	6.5	2.8	4.5	1.5	Iris-versicolor											
23	6.5	3.0	4.2	1.5	Iris-versicolor											
24	6.5	3.0	4.2	1.5	Iris-versicolor											
25	7.7	3.8	5.5	1.5	Iris-versicolor											
26	6.9	3.1	5.1	1.8	Iris-versicolor											
27	6.5	2.9	4.6	1.4	Iris-versicolor											
28	6.5	3.0	4.6	1.4	Iris-versicolor											
29	6.5	3.0	4.6	1.4	Iris-versicolor											
30	7.0	3.2	4.7	1.4	Iris-versicolor											
31	7.6	3.0	5.4	1.5	Iris-versicolor											
32	4.3	3.0	1.1	0.1	Iris-virginica											
33	5.1	3.5	1.5	0.1	Iris-virginica											
34	5.9	3.0	4.2	1.5	Iris-virginica											
35	6.4	3.2	4.5	1.5	Iris-virginica											
36	7.9	3.8	5.4	1.4	Iris-virginica											
37	7.3	3.2	6.0	1.8	Iris-virginica											
38	6.7	3.0	5.4	1.7	Iris-virginica											
39	6.9	3.1	5.4	1.5	Iris-virginica											
40	6.5	3.0	5.0	1.6	Iris-virginica											
41	6.5	3.0	5.0	1.6	Iris-virginica											
42	7.1	3.0	5.9	1.8	Iris-virginica											
43	7.0	3.2	5.4	1.5	Iris-virginica											
44	7.6	3.0	6.0	1.2	Iris-virginica											
45	7.3	3.2	6.4	1.2	Iris-virginica											
46	6.9	3.1	5.5	1.8	Iris-virginica											
47	6.5	3.0	5.1	1.8	Iris-virginica											
48	6.5	3.0	5.1	1.8	Iris-virginica											
49	7.0	3.0	5.9	1.8	Iris-virginica											
50	7.6	3.0	6.5	1.5	Iris-virginica											

Xbinary =

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	Sepal Length Low	Sepal Width Low	Sepal Length High	Sepal Width High	Petal Length Low	Petal Width Low	Petal Length High	Petal Width High								
1	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
2	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
3	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
4	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
5	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
6	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
7	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
8	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
9	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
10	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
11	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
12	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
13	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
14	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
15	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
16	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
17	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
18	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
19	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
20	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
21	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
22	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
23	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
24	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
25	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
26	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
27	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
28	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
29	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
30	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
31	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
32	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
33	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
34	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
35	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
36	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
37	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
38	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
39	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
40	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
41	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
42	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
43	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
44	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
45	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
46	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
47	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
48	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
49	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	
50	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	

<https://towardsdatascience.com> Alternative guide to association rule learning (<https://towardsdatascience.com/association-rules-2-aa9a77241654>)

<http://www.cse.msu.edu> Key reference for association rule learning, "Fast algorithms for mining association rules" (Agrawal & Srikan) (<http://www.cse.msu.edu/~cse960/Papers/MiningAssoc-AgrawalIAS-VLDB94.pdf>)

<https://rakesh.agrawal-family.com> Other key reference "Mining association rules between sets of items in large databases" (Agrawal et. al.) (<https://rakesh.agrawal-family.com/papers/sigmod93assoc.pdf>)