Introduction to Data Science 1

Problem sheet 1

Ex 1

Determine the type of the following attributes in two ways:
1: Continuous, discrete, binary? 2: Nominal, ordinal, quantitative (interval, ratio)?

1. Altitude
2. Total number of rooms in a hotel
3. Military ranks
4. Distance from the center of Heroes Square
5. International Standard Book Number (ISBN)
6. Degree: measurement of plain angle (between 0 and 360)
7. Degree of transparency: transparent, translucent, opaque
8. Cloakroom ticket numbers
9. Grades (from 1 to 5)
10. Medals (bronze, silver, gold)
11. Sex (male, female)
12. Age (in years)
13. pH (acidity or basicity of an aqueous solution)

Ex 2

Prove that the following two definitions of the Chebyshev distance are equivalent:

$$d_{c_1}(p,q) = \lim_{r\to\infty} \sqrt[r]{\sum_{k=1}^{n} |p_k - q_k|^r}$$

$$d_{c_2}(p,q) = \max_{k\in\{1,\ldots,n\}} |p_k - q_k|$$

Ex 3

Prove the following statements:

1. $L_1(x,y) = \sum_{i=1}^{d} |x_i - y_i|$ is a distance metric,
2. $L_2^2(x,y) = \sum_{i=1}^{d} (x_i - y_i)^2$ is **not** a distance metric.

Ex 4

We are given a dataset with two attributes (X and Y) and the following covariance matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

Calculate the following measures:

1. $\mathrm{Corr}(X,Y)$
2. $\mathrm{Mahal}(B,C)$, where $B = \begin{bmatrix} 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 1.5 & 1.5 \end{bmatrix}$

Ex 5

Let two feature vectors contain the following attributes:

- person's height (between 1.5 and 1.8 meters)
- person's weight (between 40 and 120 kilograms)
- person's annual income (between 10,000 and 1 million dollars)

How would you calculate the distance between the two vectors? What kind of transformations would you apply, and which distance would you choose?

Ex 6

Consider the following three documents:

- d_1: "ant bee"
- d_2: "dog bee hog ant"
- d_3: "cat gnu dog eel fox"

These documents can be represented by 8-dimensional vectors in a so-called *document-term matrix*, which describes the frequency of terms that occur in a collection of documents:

	ant	bee	cat	dog	eel	fox	gnu	hog
d_1	1	1	0	0	0	0	0	0
d_2	1	1	0	1	0	0	0	1
d_3	0	0	1	1	1	1	1	0

1. Calculate simple matching coefficient (SMC) and Jaccard-coefficient of d_1 and d_2!
2. Determine the distances derived from these coefficients! Which is the better coefficient to handle the problem? Why?
3. Can any of these approaches distinguish the "John is quicker than Mary" and "Mary is quicker than John" documents?

Ex 7

Let $tf_{i,j}$ denote the entry in the i-th row and j-th column of the document-term matrix from the previous task. For instance, $tf_{1,1} = 1$, where row = d_1 and column = ant. Consider the following tf–idf transformation.

- Let df_j be the number of non-zero elements in the j-th column, i.e., the number of documents containing the j-th word. For example, $df_1 = 2$.
- Let m be the number of documents.
- Then the transformation is defined as follows:

$$tf\text{–}idf_{i,j} = tf_{i,j} \cdot \log\left(\frac{m}{df_j}\right)$$

What is the impact of this transformation? Considering real document-word matrices, what could be its purpose? What could the abbreviation tf–idf stand for?

Ex 8

The following table's rows correspond to customers (A, B, C), and the columns correspond to products $(a, b, \ldots, h)$. The table contains 1 if a given costumer bought the given item, 0 otherwise. Determine the Jaccard similarity and the Cosine similarity of A and B!

	a	b	c	d	e	f	g	h
A	1	1	0	1	1	0	1	1
B	0	1	1	1	1	1	1	0
C	1	0	1	1	0	1	1	1

Ex 9

Assuming that the cost of compression/stretching is 0, determine the DTW distance of the following time series (let the inner distance function be the absolute distance)! Find optimal alignment between the two time series (the warping path)!

$$t_1 = (3, 2, 5, 7, 8, 9), \qquad t_2 = (2, 3, 2, 3, 6, 8)$$

Introduction to Data Science 1

Problem sheet 2

Ex 1

How many logical (Boolean, $f : \{0,1\}^N \to \{0,1\}$)) functions can be generated on N binary attributes? What are the possible functions for $N = 2$?

Ex 2

Can decision trees learn logical (Boolean) functions? How to represent the following functions with a decision tree: A **OR** B, A **AND** B, A **XOR** B, where A and B are logical variables?

Ex 3

The following table summarizes a data set with three attributes (A, B, C) and two class labels (+, −). Build a two-level decision tree.

1. Using misclassification error as the inhomogeneity measure, calculate the gains for each attribute! Which attribute gives the best split?
2. Repeat the previous step for the two children of the root node! Which nodes should be split, and which is the second splitting attribute?
3. Calculate Accuracy, Error rate, Precision, Recall, and F-measure!
4. Choose C as the first splitting attribute and continue building the tree! How would a tree of depth two look in that case?

A	B	C	Number of instances	
			class: +	class: −
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

Ex 4

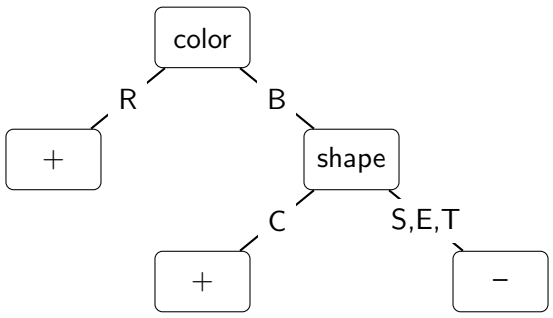
Construct the ROC curve of the following classifier and calculate the AUC! How do you interpret the AUC score? What would you suggest in terms of results? In the table, confidence scores increase from left to right.

Label:	+	+	−	+	−	−	+	+	−	+	
TP											
FN											
TN											
FP											
TPR											
FPR											

Ex 5

You can see a schematic diagram of a possible decision tree built on training data below.

1. Determine the confidence scores (ratio of positive observations) of the leaves based on the training data (train1, train2, ..., train7)!
2. Sort the confidence scores of the first three test instances (test1, test2, test3) in ascending order!
3. Construct an ROC curve using the first three instances of the test data (test1, test2, test3) and calculate the AUC!
4. Construct the ROC curve after adding two new test data (test4, test5)! Note that if more instances have the same confidence scores, the ROC curve may change diagonally!



ID	Shape	Color	Size	Class
train1	S	R	L	+
train2	C	R	H	+
train3	C	B	H	+
train4	T	R	L	+
train5	S	B	M	–
train6	E	B	L	–
train7	C	R	M	–

ID	Shape	Color	Size	Class
test1	C	R	H	+
test2	C	B	L	–
test3	E	B	H	–
test4	C	R	L	+
test5	E	R	H	–

Ex 6

Classify the following record X = (Marital status = Single, Annual income = 90K) using the Naive Bayes classifier based on the training data in the following table, where Default is the class label! Discretize annual income by 20K intervals: $[60\text{K}, 80\text{K}), [80\text{K}, 100\text{K}), \ldots$.

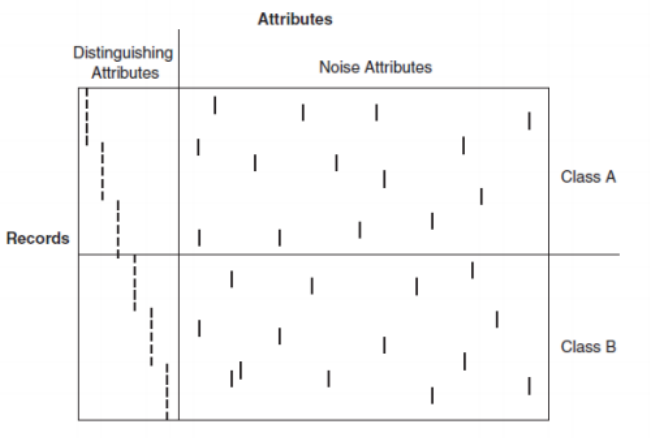
1. Use the original estimates!
2. Use Laplace smoothing!

Marital status	Annual income	Default
Single	125K	No
Married	95K	No
Single	70K	No
Married	120K	No
Divorced	75K	Yes
Married	60K	No
Divorced	220K	No
Single	85K	Yes
Married	75K	No
Single	90K	Yes

Ex 7

Assume that the following data set contains 1000 records with class label A and 1000 records with label B. There are some binary variables with distinctive power: $X_1, X_2, \ldots$. In addition, there are many noisy binary attributes that take the value 1 or 0 at random.

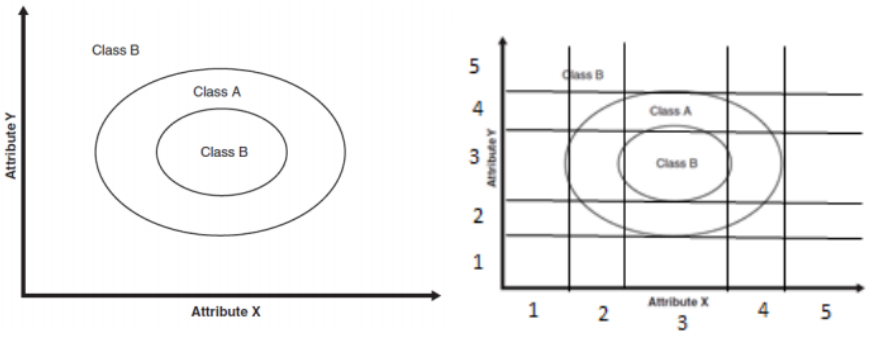
1. Sketch a decision tree that learns such data! How would the decision tree classifier perform on this data?
2. Which records are close to the first record? How would the kNN classifier perform on this data?
3. Outline the conditional probabilities! How would the Naive Bayes classifier perform on this data? Consider the first row as an example.



Ex 8

Consider the following data with two attributes (X and Y) and two possible labels (A and B). The position of class A and class B records in the X-Y space is illustrated below.

1. How would a decision tree work on such data? Indicate decision boundaries!
2. How would the kNN classifier perform on this data? What does its performance depend on?
3. How would the Naive Bayes classifier perform? Outline the conditional probabilities! We assume that the two classes have the same number of records and that the instances are distributed uniformly. Use the possible discretization given below, i.e., both attribute X and Y are discretized into 5 bins!



Introduction to Data Science 1

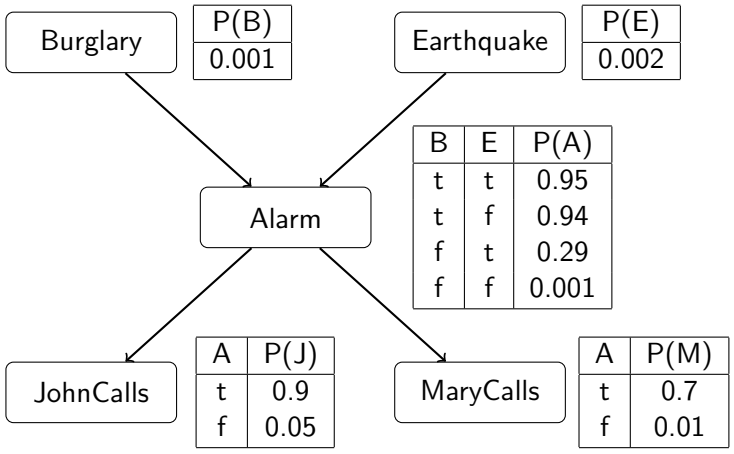
Problem sheet 3

Ex 1

The following binary-valued random variables are given:

- B = burglary in your home
- E = earthquake near your home
- A = the alarm sounds
- J = John calls because of the alarm
- M = Mary calls because of the alarm

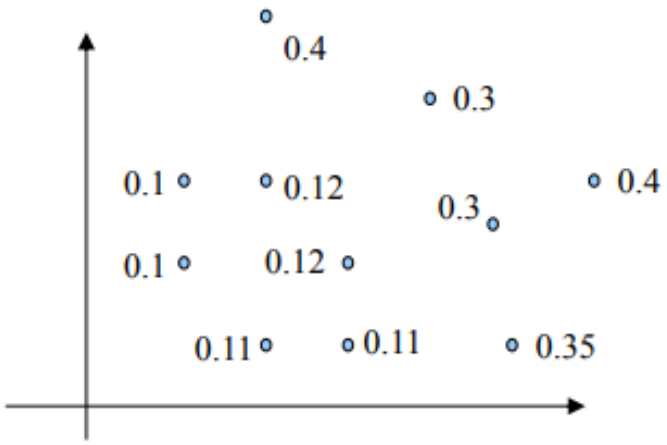
The relationship between the variables and the individual probabilities are illustrated by the following Bayesian network.



1. With how many parameters can the full joint distribution be described this way? How many parameters would be needed if the dependency network between the variables was not known?
2. Are Burglary and Earthquake independent of each other? Are they conditionally independent given the value of Alarm?
3. Are JohnCalls and MaryCalls independent of each other? Are they conditionally independent given the value of Alarm?
4. Outline how the conditional probability $\mathbb{P}(B = t \mid M = t,\, J = f)$ should be determined!

Ex 2

We would like to solve a regression problem using a decision tree. The maximum number of leaves is set to 3. How would it split the data given in the coordinate system below? Sketch the splits on the figure. (No precise calculation is required.)



Ex 3

Let (0, 0, –2); (0, 1, 1); (1, 0, 2) be three records on the x_1–x_2 plane, where the third coordinate is the y target variable. Determine the coefficients of the following linear regression that minimizes the squared error: $y = w_1x_1 + w_2x_2 + w_0$.

a) Determine the optimal coefficients analytically!

b) Approximate the optimal coefficients using the gradient descent method (a few steps are enough)!

c) Approximate the optimal coefficients using the stochastic gradient descent (a few steps are enough)!

For gradient methods use the following initialization of the weights: $w_1 = w_2 = w_0 = 1$. Let the learning rate be 0.25.

Ex 4

Consider the following data, where the first and the second coordinates are binary attributes, and the third is the class label: (1, 1, –); (1, –1, +); (–1, 1, +); (–1, –1, –). Which Boolean function do you recognize in the data? Is it linearly separable? If so, give the equation of the separating line with maximal margin. If the function is not linearly separable, then transform the data into the following three-dimensional feature space: $(x_1, x_2, x_1 \cdot x_2)$. Furthermore, find the equation of the separating plane with the maximum margin in the transformed space!

Ex 5

How does the $2 + x + y = 0$ line separates the 2-dimensional plane? Draw the line in a coordinate system. Which of the following records are support vectors of the given line: (–3, 0); (0, –3); (–1, 0); (0, –1); (0, 0)?

Ex 6

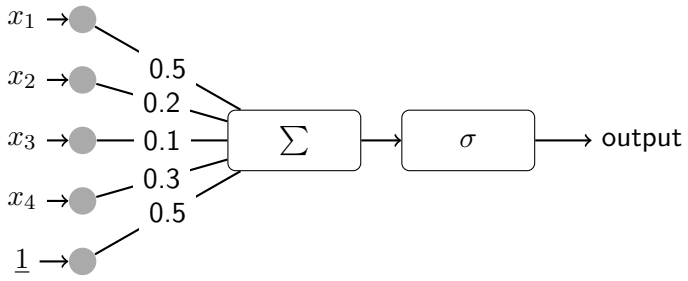
Represent the following logical functions with a perceptron or show that it is not possible to do so. In the latter case, construct a neural network with one hidden layer.

a) A **AND** B **AND** C

b) (A **XOR** B) **AND** (A **OR** B)

Ex 7

The following perceptron is given, which we use to solve a binary classification problem. The two class labels are denoted by $+1$ and -1, and the activation function is the sigmoid function.



1. How does this perceptron classify the following instance (the threshold is 0.5)?
   $(x_1, x_2, x_3, x_4) = (1, -0.8, -0.3, 1.5)$
2. Assume that the true class label of the instance above is 0, and this record is used for training the perceptron. The learning rate is $\lambda = 0.1$, and the objective function to be minimized is the logarithmic cost (log loss). Perform one update step! How do the weights of the edges change?

Ex 8

Consider a chain of two neurons. The input of the first neuron is x_1 and the output is $y_1 = ax_1 + b$. The input of the other neuron is x_2 and the output is $y_2 = cx_2 + d$ (the activation function is the identity function in both cases). Connect the two neurons so that the second neuron's input is y_1, i.e. $x_2{=}y_1$.

a) Draw this neural network!

b) Give the final output y_2 as a function of x_1!

c) Let the input of the ANN be x and the output be y. Using the gradient descent method, show how the weights (a, b, c, d) are updated after one training step if the squared error is minimized!

Introduction to Data Science 1.

Practice Problem Set 4.

Exercise 1

Consider the following market baskets (transactions).

1. Determine the frequent, maximal, and closed itemsets if the minimum support is 0.3 (minsupp = 0.3). Illustrate the operation of the Apriori algorithm on this example!
2. Determine the confidence and lift of the following rules!
   a) bread $\rightarrow$ milk
   b) {bread, coffee} $\rightarrow$ milk

TID	Basket
1	milk, beer, coffee
2	milk, coffee, bread
3	milk, coffee, beer
4	bread, beer
5	apple, coffee
6	bread
7	apple, bread
8	apple
9	coffee
10	milk

Exercise 2

The k-means clustering algorithm is initialized with the marked centroids. Perform one iteration and provide the new centroids at the end of the $i+1$-th iteration. Show your calculations!

Exercise 3

Given the following rating matrix, where rows represent users and columns represent movies. We want to estimate the missing ratings using a latent factor model. We assume two latent factors, set the learning rate to $\varepsilon = 0.1$, and omit regularization!

1. Perform one update step using user A and movie TW! Initialize the factor matrices for users and items with 1s.
2. Show that this update step corresponds to a step in a gradient descent method!
3. Perform the update step again, but this time using a regularization term!
4. *Bonus*: Which movies might the abbreviations represent?

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

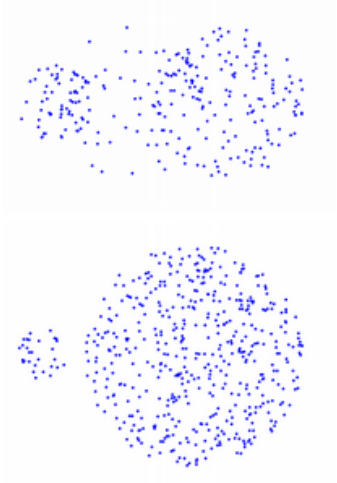
Exercise 4

Show how the k-means clustering algorithm and hierarchical clustering using MIN (single linkage) and MAX (complete linkage) distances would partition the following datasets into two clusters.

Exercise 5

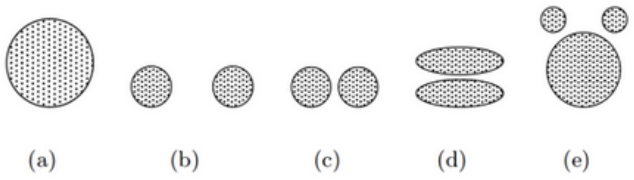
Given the similarity matrix below, show using dendrograms how a hierarchical clustering algorithm would cluster the instances using single linkage and complete linkage!

	1	2	3	4	5
1	1	0.15	0.6	0.15	0.95
2	0.15	1	0.5	0.2	0.2
3	0.6	0.5	1	0.05	0.7
4	0.15	0.2	0.05	1	0.85
5	0.95	0.2	0.7	0.85	1



Exercise 6

For the following two-dimensional data points, sketch the results that the k-means clustering algorithm would yield using Euclidean distance! Also indicate approximately where the final centroids would be located! If you think there is a case with multiple possible solutions, consider which gives a local vs. global optimum! (a) $k = 2$ (b) $k = 3$ (c) $k = 3$ (d) $k = 2$ (e) $k = 3$

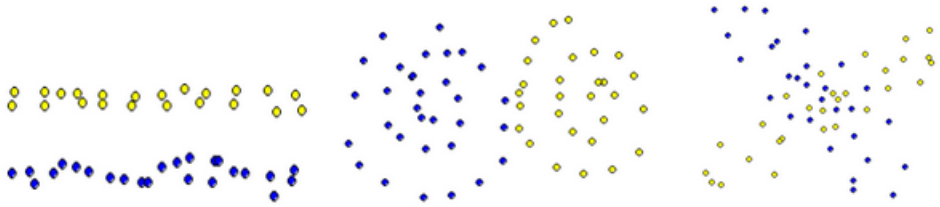


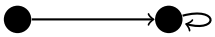
Exercise 7

Which clustering algorithm would best partition the following data into two clusters if the goal was to find the clusters indicated by colors by an expert? Consider the following: hierarchical clustering (with MIN and MAX distance), k-means, Gaussian Mixture Model.

Exercise 8

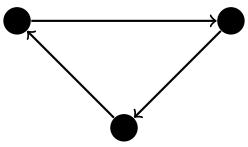
Calculate the PageRank values for the following graph by determining the stationary distribution for the cases α=1 and α=0.8!





Exercise 9

Calculate the first few iteration steps of the PageRank algorithm starting from the state vector $q^0 = (1, 0, 0)$ for the following graph. Determine this for both α=1 and α=0.7. What is the main difference? What problem arises if the teleportation probability is 0?



Exercise 10

What is the probability that the PageRank random walk performs exactly l steps by traversing edges (i.e., mimicking clicking on links) between two teleportations? Determine the probability as a function of α!