Introduction to Data Science: Exam

University of Gothenburg
Department of Computer Science and Engineering
January 2023 (DIT852)

Information:

- The exam takes place from 08:30–12:30 on Wednesday 4 January 2023.
- Speak with an invigilator if you want to ask the examiner a question. The examiner will be available between 09:30 and 10:00.
- You can earn a total of 50 points from 7 questions in the exam.
- Bonus points will be added to the exam points based on Zoom poll results. The maximum will be 2.5 bonus points (5 % of the points available in the written exam) if you gave correct answers for all Zoom poll questions.
- Grades are normally determined as follows: $\geq$ 40 % for grade G.

Instructions:

- You may use one A4-sized sheet with hand-written notes (front and back), but all work must be your own. No photocopies or print-outs of slides, books, or material off the web. If you bring a sheet of notes to the exam, it must be handed in with your exam solutions. *Please, write the exam code on it.*
- Begin the answer to each question on a new page. Write page number and question number on **every** page.
- Write clearly; unreadable = wrong!
- Fewer points are given for unnecessarily complicated solutions.
- Indicate clearly if you make any assumptions that are not given explicitly in the question.
- Show **ALL** your work. You will get little or no credit for an unexplained answer. Please indicate why a specific computation or transformation is appropriate. The points of each question appear in parentheses; use this for guiding your time.
- There is no need to compute numerical answers; you may leave binomials, factorials and fractions, should they arise, as is.
- **No electronic devices of ANY kind!** Please store **all** your devices in your bag and not on your person. Any device found at your seat, **even if it is turned off**, will be considered cheating and reported!
- Printed English language dictionaries—including dictionaries translating to and from another language to English—are allowed. Electronic dictionaries are not.

Question 1 [*3 points total*]

In the context of analysing data, what is *stratification*? Why is stratification done? Give an example of a use case where stratification is useful.

Question 2 [*10 points total*]

(a) [*2 pts*] Explain how a *k-nearest neighbours* classifier works in the case where $k = 5$.

(b) [*2 pts*] When calling the scikit-learn function `neighbors.KNeighborsClassifier` the user can specify whether the parameter *weights* has the value "uniform" or "distance". Explain how the choice of value for this parameter affects how classification is done.

(c) [*4 pts*] Describe how 5-fold cross-validation can be carried out to evaluate the performance of a classifier.

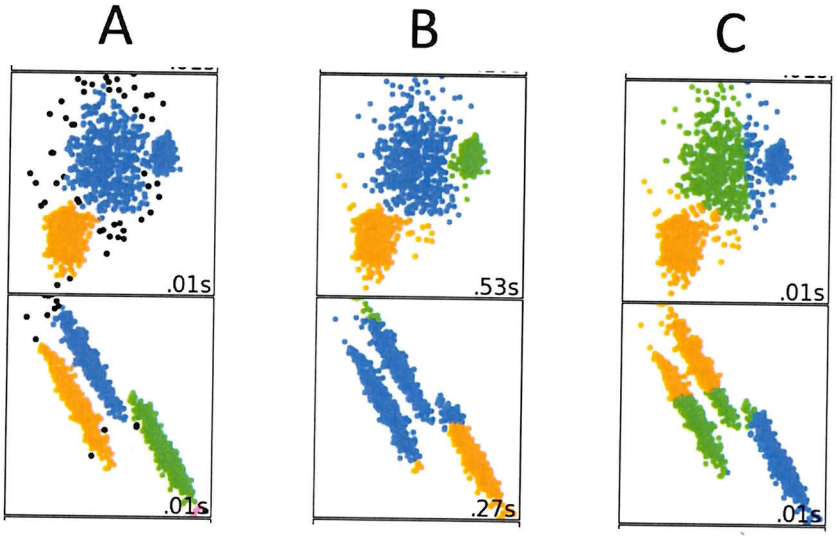
(d) [*2 pts*] The following confusion matrix relates to a scenario where a truck driver might be a smuggler or might be innocent, and a customs officer decides either to stop and check ("control") a truck, or to allow it to pass through customs without being checked.

		Actual Class	
		Smuggler	Innocent
Predicted Class	Control	Controlled Smuggler	Controlled Innocent
	Pass	Passed Smuggler	Passed Innocent

What are the possible errors that might be made by the customs officer? Explain which of these errors you consider to be the most serious.

Question 3 [*8 points total*] 6

(a) [*4 pts*] This diagram shows three pairs of clustering results. One pair of results was produced using k-means clustering. One pair of results was produced using density-based clustering. One pair of results was produced using hierarchical clustering.



Which pair of clustering results were produced using which method? Motivate your answer.

(b) [*4 pts*] In the DBSCAN clustering algorithm, what are *core points*?

Apart from core points, what two other kinds of points are identified by the DBSCAN clustering algorithm? Describe each of these kinds of points.

Question 4 [*10 points total*]

(a) [*3 pts*] Consider the following Python statement:

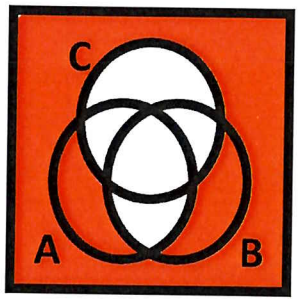
```
D = set([x for x in range(1,10) if x % 3 == 0])
```

i) What is the value of set D?

ii) What is the *cardinality* of D?

iii) What is the *power set* of D?

(b) [*2 pts*] Give a set notation expression that corresponds to the region shown in red in this Venn diagram.



(c) [*3 pts*] Show which, if any, of the following expression is/are equivalent to $p \wedge q$

i) $\neg(p \rightarrow q)$

ii) $\neg(p \rightarrow \neg q)$

iii) $\neg p \vee \neg q$

(d) [*2 pts*] Suppose x and y are both positive integers. State, with reasons, whether each of the following statements is true or false.

i) $\forall x \forall y(x + y \geq x \cdot y)$ ✗

ii) $\forall x \exists y(x + y \geq x \cdot y)$ ✓

iii) $\exists x \forall y(x + y \geq x \cdot y)$ ✓

iv) $\exists x \exists y(x + y \geq x \cdot y)$ ✗

$1+1 \geq 1$

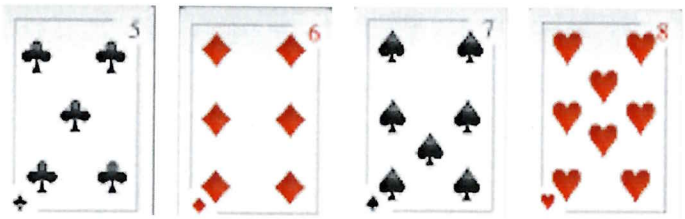
$1+2 \geq 2 \quad 2+1 \geq 2$

$2+2 \not\geq 4$

$1+3 \geq 3 \quad 4+1 \geq 4$

Question 5 [*7 points total*]

(a) [*5 pts*] Suppose that the four cards shown below (5 of clubs, 6 of diamonds, 7 of spades, 8 of hearts) are mixed randomly.



Suppose that cards are drawn randomly from the set of four cards.

i) What is the probability that the first two cards drawn are both red cards?

ii) What is the probability that the second card drawn has a higher value than the first?

iii) Suppose that the first card drawn is a red card. What is the probability that the second card drawn has a higher value than the first?

(b) [*2 pts*] Suppose we draw five cards from a full deck of 52 cards.
What is the probability that the 8 of hearts is among the drawn cards? Express this probability both as a fraction and using binomial coefficients.

Question 6 [*6 points total*]

(a) [*2 pts*] Explain the goal of a generative adversarial network (GAN).

(b) [*2 pts*] Draw the architecture of a GAN and explain how it is trained.

(c) [*2 pts*] What kind of an autoencoder can achieve the same goal as a GAN? Explain why.

Question 7 [*6 points total*]

(a) [*2 pts*] What is the formal definition of a graph problem (or path-finding problem)?

(b) [*2 pts*] Explain the main steps of the generic search algorithm.

(c) [*2 pts*] How are different search algorithms (such as BFS and DFS) obtained from the generic search algorithm?

CHALMERS
EXAMINATION / TENTAMEN

Course code/kurskod	Course name/kursnamn			
DIT 852	Intro to Data Science			
Anonymous code Anonym kod		Examination date Tentamensdatum	Number of pages Antal blad	Grade Betyg
DIT852-0002-FWT		04/01/23	7	9

* I confirm that I've no mobile or other similar electronic equipment available during the examination.
Jag intygar att jag inte har mobiltelefon eller annan liknande elektronisk utrustning tillgänglig under eximinationen.

Solved task Behandlade uppgifter No/nr		Points per task Poäng på uppgiften	Observe: Areas with bold contour are to completed by the teacher. Anmärkning: Rutor inom bred kontur ifylles av lärare.
1	X	2	
2	X	9	
3	X	6	
4	X	8	
5	X	5	
6	X	1	
7	X	2	
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
Bonus poäng			
Total examination points Summa poäng på tentamen		33	+ 2 BONUS POINTS

CHALMERS | Anonymous code: DIT852-0002-FWT | Anonym kod | Points for question (to be filled in by teacher): 2 | Poäng på uppgiften (ifylles av lärare) | Consecutive page no. Löpande sid nr: 1 | Question no. Uppgift nr: 1

Question 1: Stratification is a method used to analyse data by dividing the data into smaller subgroups (or layers). This is done in an effort to find similarities between ~~data~~ data in a particular groups in hopes of finding more useful insight, or finding insight more efficiently ~~between~~ on data ~~with~~ that have similarities ~~amongst~~ between ~~them~~ ok. 1

Stratification is useful for example when trying to predict the test score of students on an exam given its standing among its classmates if different subgroups were created ~~for~~ based on their standings. For example top 25% -> high group, middle 50% -> Mid group, last 25% percentile -> low group.

Example could be clearer.

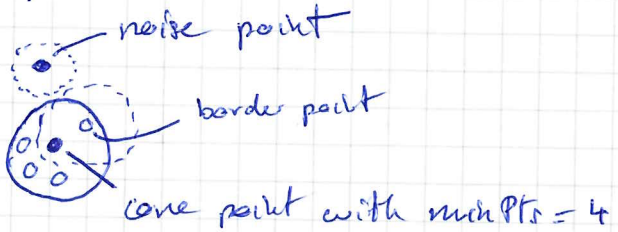
2

CHALMERS | Anonymous code: DIT852-0002-FWT | Points for question (to be filled in by teacher): 6 | Consecutive page no. Löpande sid nr: 3
Anonym kod | Poäng på uppgiften (ifylles av lärare) | Question no. Uppgift nr: 3

Question 3:

(a) B was clustered using ~~hierarchical~~ k-means clustering as this method took the longest time to compute but also worked best on spherical shapes. C was clustered using hierarchical clustering as it was faster then k-means. And finally A was clustered using density-based clustering as it was faster then k-means but also because it left many outliers behind that are not included in the groups which is typical of that clustering method. ✓ 2

(b) Core points are datapoints that have at least minPts neighbouring datapoints, minPts being a parameter set by the user. Other points include border points which are non-core points that have at least one core point as its neighbour; and also noise points which are non-core and non-border points (also called outliers).



Neighbouring points could be defined as any two points within distance ϵ, ϵ being a parameter set by the user.

4

3

Question 4:

(a) i) $D = \{1, 2\}$, $|D| = 2$ (✓) ✗

ii)

iii) the power set is $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. ✗

1

(b) $((A \cap B) \cup C)^c$

2

(c)

p	q	$p \wedge q$	$p \to q$	$\neg(p \to q)$	$\neg q$	$\neg(p \to q)$	$\neg(p \to \neg q)$	$\neg p$	$\neg p \vee \neg q$
T	T	T	T	F	F	F	T	F	F
T	F	F	F	T	T	T	F	F	T
F	T	F	T	F	F	T	F	T	T
F	F	F	T	F	T	T	F	T	T

∴ only $\neg(p \to \neg q)$ is equivalent to $p \wedge q$.

3

(d) ~~if True as ref $x=4$ and $y=1 \Rightarrow x+1$~~

~~as~~ i) False as for x or $y > 1$ the equality does not hold.

ii) ~~True as long as $y=1 \Rightarrow x+y \geq x \times y$ for $\forall x$~~ False because even if $y=1$, x can still be 1 under $\forall x$ which x-

iii) ~~True as long as $x=1 \Rightarrow x+y \geq x \times y$ for all $\forall y$~~ False

iv) True as long as $x=1$ and $y \geq 2$ or $y=1$ and $x \geq 2$

i) False, as if ~~both~~ either x ~~and~~ or y >2 the ^in equality does not hold.

i.e: $x=3$ & $y=2$ $5 \not\geq 6$

✓

ii) True, as long as $y=1$, the ^equality holds for $\forall x$. i.e: $x=4$ $y=1$ $5 \geq 4$.

iii) True, as long as $x=1$, the ^equality holds for $\forall y$.

iv) True, as long as $x=1$ or 2 and $y=1$ or 2, the ^equality holds.

2

CHALMERS | Anonymous code: DIT852-0002-FWT | Anonym kod | Points for question (to be filled in by teacher): 5 | Poäng på uppgiften (ifylles av lärare) | Consecutive page no. Löpande sid nr: 5 | Question no. Uppgift nr: 5

Question 5: (a) i) $S = \{(red, red), (red, black), (black, red), (black, black)\}$

↳ $P(\text{Both red}) = 1/4$ ✗

ii) $S = \{(5,6),(5,7),(5,8),(6,5),(6,7),(6,8),(7,5),(7,6),(7,8),(8,5),(8,6),(8,7)\}$

Let's call event A the event where the second card pulled is higher than the first: $P(A) = 6/12 = 1/2$

iii) $S = \{(6,5),(6,7),(6,8),(8,5),(8,6),(8,7)\}$. Let's call the event that the second card drawn has a higher value than the first after the first card is red. Event B: $P(B) = 2/6 = 1/3$ 4

(b) Let's call the event C: $P(C) = \dfrac{\binom{1}{1}\binom{51}{4}}{\binom{52}{5}}$ ✓

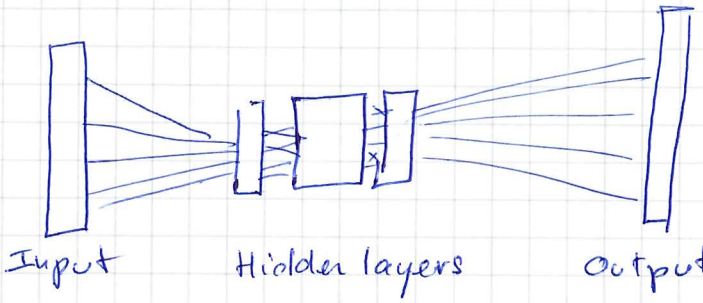
And as a fraction? 1

CHALMERS	Anonymous code: DIT852-0002-FWT	Points for question (to be filled in by teacher)	Consecutive page no. Löpande sid nr: 6
	Anonym kod	Poäng på uppgiften (ifylles av lärare): 1	Question no. Uppgift nr: 6

Question 6:

(a) The goal of a GAN is to efficiently code unlabeled data by regenerating the input. 1

(b)



0

(c)

0

3

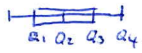
Question 7: (a) The graph problem consists of finding a certain path from a given starting point to a certain end point along a network of nodes which are connected to each other through edges with given weights. Usually the graph problem consist of finding that path given a ~~the~~ condition for its cost function, which is a function that sums the weight of the nodes that path traverses (usually try to minimize this cost function).

(b) The main steps of the general search algorithm consists of # entering the word as the input, then given an activation formula give the input through the hidden layer in an attempt to predict a context as the output.

(c) Different search algorithm can be obtained depending on the depth of the neural network, namely the number of hidden layers in the algorithm.

Regression: predicting a numerical quantity

Exam Code: DIT852-0002-FWT

Box plot: 

- interquantile = $Q_3 - Q_1$

Linear Regression: $y = kx + m$ → Find "fit" that min residual errors

Residual Plots: diff. btw prediction & actual 

Correlation: quantifies the strength of linear trend
- mean: measures the center of data: $\frac{\Sigma x_i}{n} = \bar{x}$
- variance: " how the data varies $= \frac{\Sigma(x_i - \bar{x})^2}{n-1}$
- std dev: " the spread from expected values $= \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$

Classification: assigning label from discrete set of possibilities

K-nearest: likelihood a point joins a group based on the groups around him based on frequencies of label around it (k-nearest datapoints).
- Pro: no training, no ass. about shape - Cons: comp. expensive, no insight

Log. Regression: predict prob. of a binary event occuring $\frac{1}{5} \Rightarrow 6 = \frac{1}{1+e^x}$

Support Vector machines: find line that seperates two classes with the highest margins

Conf. matrix: looks at the performance of a classification alg.
- Accuracy: $\frac{TP+TN}{TP+FP+TN+FN}$ - Precision: % of result which are relevant $\frac{TP}{TP+FP}$
- Recall: % of relevant result correctly classified → $\frac{TP}{TP+FN}$
- specificity: ability to predict TN of each category → $\frac{TN}{TN+FP}$
- F-score: harmonic mean of Precision & Recall $\frac{2TP}{2TP+FP+FN}$

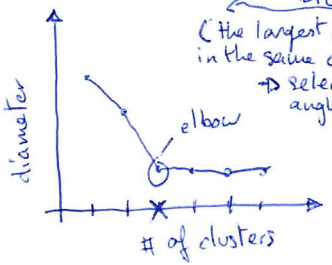
cross-validation: partition the data → choose test block, rest is training → train & test → repeat until all blocks have been test once.

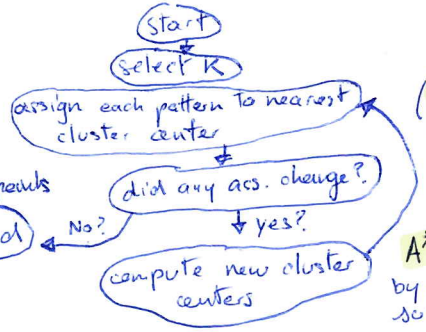
Clustering: grouping data by similarities

Purpose?
- underlying structure: gain insight in data, generate hypothesis
- natural classification: find degree of similarities in the data
- compression: organize data, summarize in clusters

K-means: "greedy" algorithm, choose "K"
- works best for: • spherical clusters • equal variance • equal cluster size

to find "K": elbow plot → longest dist. btw cluster diameter (the largest dist. between 2 data points in the same cluster) → select K where the angle is smallest "elbow".





DBSCAN: Pros: can fit any shapes - doesn't require a # of cluster to be defined
- distance measure: Euclidian
- number defining neighbours: EPS, max distance btw two points that are neighboors
- number defining clusters: minPts, min number of points in clusters.

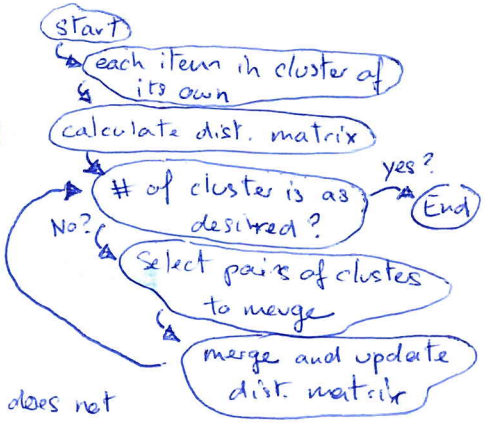
Points: - core point: has at least minPts neighbours: minPts = 4
- border point: non-core point that has at least one core point as its neighbour - neighbor: any two points within EPS.
- noise point: or outliers, is non-core & non-borders

Algorithm: ① label all points as core, border, noise
② eliminate noise points
③ put an edge btw all core points with EPS of each other
④ make each group of connected core points into a separate cluster
⑤ assign each border points to one of the clusters of its associated core point

Hierarchical Clustering: build bigger clusters by joining smaller clusters together that are close

merging clusters:
- neighbor joining: if input distance is correct, then output tree is correct
- Bi-clustering: Bi clustering simultaneously clusters rows and columns of a data matrix

A* returns a min. cost solution if there is a solution and:
- the branching factor is finite
- the arc cost are uniformly bounded (there exist a $\epsilon > 0$ such that all arc costs are $> \epsilon$).
- the heuristic f is admissible.



Validating clustering:
- stability: if removing random points does not change the clusters fundamentally
- stability over repetitions: stable if the same points end up in the same clusters from random initialization
- silhouette coefficient: used to calculate the goodness of a clustering technique using mean

• heuristic: cost of the cheapest path from end node to goal node.

(Def of a Neuron in NN): consists of a weight vector "w", and bias "b" and an activation function $f: R \to R$. The output of a neuron is $f(w \cdot x + b)$. The act. f. generalizes the firing behavior in the biological neuron's model (perceptron), the bias sets the firing level and the weights are the gains of diff. inputs.

A*: A* is an instance of the best-first-search algorithm, where the elements of the frontier are sorted by a grade and the path with the smallest grade is selected, in each iteration. The grade is the sum of the total cost of the path and its heuristic function. The greedy b-f-s alg. only considers the first value, the cost of the path.

Admissible Heuristic function: if its value for any given path never overestimates the cost of reaching the goal node from the end point of p. For any path p' from the end point to the goal node $c(p') >= h(p)$ where $c(p')$ is the actual cost and $h(p)$ is the heuristic f.