Introduction to Data Science

Final Exam, VU University, Department of Computer Science

October 24th 2018, 12:00 - 14:45

This exam consists of 21 questions worth 90 points in total. Answer the open questions in full sentences. Unless stated otherwise, every multiple choice question has exactly one correct answer.
The exam determines 50% of your final grade for the course. Your exam grade must at least be 5.5 to pass the course.
Good luck!

Name: ______________________________

Student number: ______________________________

1. (8 points) In the CRISP-DM model of a data science project life cycle, there are 6 stages. For each of the activities below, fill in the stage that it fits most to. (Multiple activities can be connected to the same stage.)

   ____________ Making sure that a feature containing prices all use the same currency.

   ____________ Talking to the end users of your output to understand what they will want to do with it.

   ____________ Visualizing a feature to get a sense of its spread.

   ____________ Scaling your model from a pilot in one department, to the entire company.

   ____________ Talking to the end users of the pilot version of your model to see it makes a difference to them.

   ____________ Combining data sets into one.

   ____________ Talking to someone with substantive expertise to understand the outliers better.

   ____________ Asking the collector of the data for the exact procedure with which the data was collected, in order to check for potential biases.

2. (4 points) You are preparing a data set for a project that looks into the effect of population density on economic growth. Part of your data set is shown in table 1. Name two things you have to clean in this data, and propose a way to solve these problems.

countryid	countryname	continent	populationsize	surface
01	The Netherlands	NA	17,259,111	NA
02	Germany	europe	82.792.351	357.385,71
03	Croatia	evropa	4.190.66	56.594
04	Spain	EU	46.549.045	505.970
05	Greece	Europe	10,768,477	131,957

Table 1: Part of the data set *Population density in Europe*

3. (3 points) Which of the following statements about normal distributions are true.

□ **Correct** □ **Incorrect** A normal distribution's mean is also its median.

□ **Correct** □ **Incorrect** A normal distribution's mode is also its mean.

□ **Correct** □ **Incorrect** Positive and negative deviations from this central value are equally likely.

□ **Correct** □ **Incorrect** There is a linear relationship between a value's probability and its distance from the mean.

4. (2 points) Which of these statements is true for data following a log-normal distribution?

A. Your data frequency histogram roughly follows an exponential curve

B. Your data frequency histogram roughly follows a hyperbolic curve

C. Your data frequency histogram roughly looks like a bell, with a median lower than the mean

D. Your data frequency histogram roughly looks like a bell, with a median higher than the mean

5. (2 points) In figure 1 you see four images of a bull's eye, that represents the true mean in the population that you want to find. The dots in each of the images represent the data you have collected. Each image represents a type of data with low/high variance and/or bias. The images are organized along two axes. One from low to high variance, and one from low to high bias. Fill in these four items in the empty boxes in the figure, so that they accurately describe the four images:

   - Low variance
   - Low bias
   - High variance
   - High bias

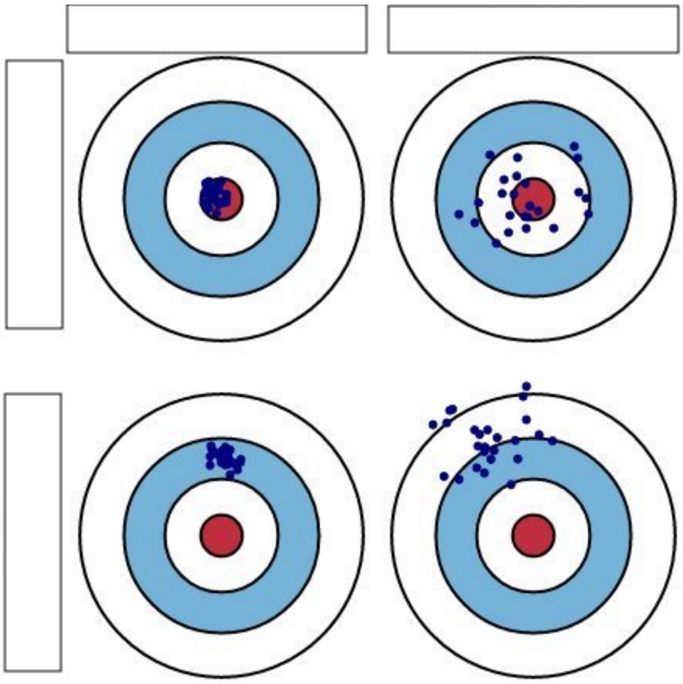
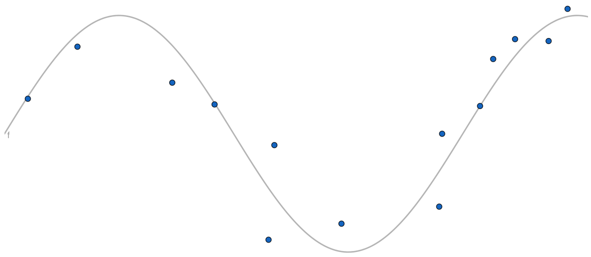
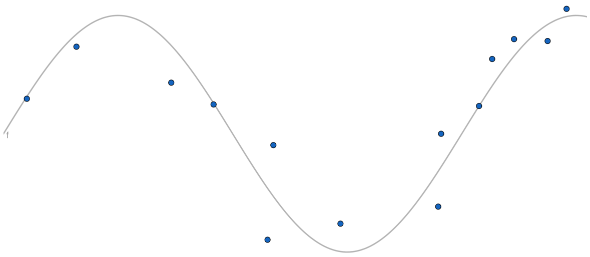


Figure 1: question 5

6. (3 points) Draw a model that is underfitting and one that is overfitting in the plots below.



(a) Example of underfitting.



(b) Example of overfitting.

7. (10 points) For each model, fill in the applicable properties in each cell of the table below. List all that apply.

	Type: supervised/ unsupervised	Input: numerical/ categorical/ordinal	Output: class/ value/other
Linear Regression			
Logistic Regression			
k-NN			
k-means			
Hierarchical Clustering			
Decision Tree			

8. (6 points) You want to predict whether a visitor to your website will purchase one of your products. In Table 2 you find an example data entry of two visitors. Of the models below, which would you choose? There are multiple acceptable answers, choose one. Discuss one advantage and one disadvantage of the method you picked.

ID	Time spent on page (mins)	Location	Browser	Reference	Bought product
2153	10	NA	Firefox Quantum	-	True
2154	3	EU	Chrome 3	Facebook ad	False

Table 2: Example data entry for question 8

A. k-means
B. k-NN
C. Neural network
D. Random Forests

Advantage:

Disadvantage:

9. (3 points) What is an advantage of random forests over a decision tree?
   - A. Random forests can handle missing values.
   - B. Random forests are less likely to overfit.
   - C. Random forests are more transparent.

10. (3 points) K-means uses iteration to reach a division of the data in clusters. What is the stopping criterion for this iteration?

11. (3 points) Which of the following statements about hierarchical clustering are true?

□ **Correct** □ **Incorrect**	The runtime of most hierarchical clustering algorithms scales linearly with n.
□ **Correct** □ **Incorrect**	Top-down and bottom-up hierarchical clustering will always produce the same tree.
□ **Correct** □ **Incorrect**	You need to pick the number of clusters you are looking for beforehand.

12. (3 points) Derive all 2-grams from this sentence. And from this one.

13. (3 points) What must a neural network absolutely have for it to handle classification problems where the classes are not linearly separable? (Pick one.)
   - A. A large number of input features
   - B. Inputs which were transformed non-linearly
   - C. More than two layers of nodes
   - D. Ordinal output labels
   - E. Neural networks can not handle these problems

14. (3 points) Which machine learning algorithm uses trial and error of pseudo-random combinations to find a good solution?
   - A. Evolutionary computing
   - B. Logistic regression
   - C. k-Nearest neighbours
   - D. Neural networks

15. (10 points) Explain one potential ethical concern when automatically predicting whether someone will default (not pay) on a mortgage, with a model trained on data of previously hand-selected people who were granted a mortgage.

16. (2 points) Give an example of a black box algorithm and of a glass box algorithm.

Black box:

Glass box:

17. (5 points) Explain how the choice for a black or glass box algorithm relates to accountability.

18. (3 points) Which charts should you use to display which data? (Connect each of the data sets on the left with a chart on the right, with a line.)

Average monthly temperatures in Amsterdam in the years 1980-2017	●	●	Stacked bars
Proportion of household expenses (housing, food, transportation, education, investments and entertainment) in the Netherlands, by income bracket	●	●	Scatter plot
Number of minutes of weekly physical activity and BMI measures of a sample of 1500 adults	●	●	Line graph

19. (4 points) For each of the following objectives, fill in an **F**, **A**, **I** or **R** to indicate which part of FAIR they belong to.

______ (Meta)data are assigned globally unique identifiers.

______ (Meta)data meet domain-relevant community standards.

______ There is a standardized communications protocol that is open, free, and universally implementable.

______ (Meta)data include qualified references to other (meta)data

20. (6 points) There are three main challenges for automatic language comprehension: Ambiguity (A), Variation (V) and Pragmatics (P). For each of the situations below, choose which challenge it is an example of, or whether it isn't an example of any of them (N).

□**A** □ **V** □ **P** □ **N** All these words can refer to sleeping: "doze", "nap", "hibernate", "slumber", "repose", "snooze", "siesta".

□**A** □ **V** □ **P** □ **N** The word 'like' can be used as verb, a preposition, an adjective, an adverb or an interjection.

□**A** □ **V** □ **P** □ **N** In order to apply a supervised learning algorithm, a linguist has to assign word categories (verb, noun, pronoun, preposition, etc.) to every single word in a newspaper by hand.

□**A** □ **V** □ **P** □ **N** In the following sentence, 'the red sweater' doesn't refer to a piece of clothing, but to a person: "Could the red sweater in the back please be silent now?".

□**A** □ **V** □ **P** □ **N** The sentence 'Time flies like an arrow' can be parsed in many ways. (For example in the same way as you would parse "Fruit flies like a banana".)

□**A** □ **V** □ **P** □ **N** Within the right context, it makes sense to say that peanuts fall in love.

21. (4 points) In his lecture about the data science department of Heineken, Ciaran mentioned four reasons why in his experience many data science projects fail. Mention two of them.