Exam Data Science II Final

Neighbours

Lecturer: Daniel T. McGuiness, Ph.D

.....

SEMESTER: WS 2024

DATE: 16.12.2024

TIME: 13:00 - 14:30

.....  .....

First and Last Name

.....

.....

Student Registration Number

.....

Grading Scheme	$\geq 90\%$	1
	$\leq 80\%$ and $\geq 90\%$	2
	$\leq 70\%$ and $\geq 80\%$	3
	$\leq 60\%$ and $\geq 70\%$	4
	$\leq 60\%$	5

Result:

___/ max. 100 points

Grade:

Student Cohort BA-MECH-22

Study Programme B.Sc Mechatronics, Design and Innovation

Permitted Tools Nothing is allowed.

Important Notes

Unnecessary Items

Place all items not relevant to the test (including mobile phones, smartwatches, etc.) out of your reach.

Identification (ID)

Lay your student ID or an official ID visibly on the table in front of you.

Examination Sheets

Use only the provided examination sheets and label each sheet with your name and your student registration number. The sheets be labelled on the front. Do not tear up the examination sheets.

Writing materials

Do not use a pencil or red pen and write legibly.

Good Luck!

Question	Maximum Point	Result
On the Topic of Machine Learning	100	
Sum	100	

[Q1] On the Topic of Machine Learning — 100

a. What type of algorithm would you use to segment your customers into multiple groups ? Please justify your reasoning. (20)

b. What are the main applications of clustering algorithms? Please give examples of where clustering algorithms would be the preferred options over other machine learning algorithms. (20)

c. What is the fundamental idea behind support vector machines and what is a support vector? Please explain both these questions with sufficient detail and draw diagrams or write equations if necessary. (20)

d. What is precision and recall ? Is it possible to have a system which has both perfect precision and recall? Please give an example of a perfect recall and a perfect precision. (20)

e. When a dataset dimension has been reduced from n to $n - 1$ is it possible to go back (i.e., has information been lost) ? If so, why? If not why? (20)

a. If you don't know how to define the groups, then you can use a clustering algorithm (unsupervised learning) to segment your customers into clusters of similar customers. However, if you know what groups you would like to have, then you can feed many examples of each group to a classification algorithm (supervised learning), and it will classify all your customers into these groups. (20)

b. The main applications of clustering algorithms include data analysis, customer segmentation, recommendation systems, search engines, image segmentation, semi-supervised learning, dimensionality reduction, anomaly detection, and novelty detection. (20)

c. The fundamental idea behind Support Vector Machines is to fit the widest possible "street" between the classes. In other words, the goal is to have the largest possible margin between the decision boundary that separates the two classes and the training instances. When performing soft margin classification, the SVM searches for a compromise between perfectly separating the two classes and having the widest possible street (i.e., a few instances may end up on the street). Another key idea is to use kernels when training on nonlinear datasets. SVMs can also be tweaked to perform linear and nonlinear regression, as well as novelty detection. (20)

d. Precision is a metric evaluating the ability of a model to correctly predict positive instances. This reduces the number of false positives in the process. False positives are cases in which a machine learning model incorrectly labels as positive when they're actually negative. (4)

   Recall is a metric evaluating the ability of a machine learning model to correctly identify all of the actual positive instances within a data set. True positives are data points classified as positive by the model that are actually positive (correct), and false negatives are data points the model identifies as negative that are actually positive (incorrect). (4)

   Example for Perfect Precision: In medical diagnosis, precision is paramount to ensure accurate identification of serious conditions like cancer. False positive diagnoses can lead to unnecessary treatments, procedures, and psychological distress for patients. (4)

   Example for Perfect Recall: Consider a call centre for insurance claims. Most fraudulent claims are phoned in on Mondays. What's the best thing for an insurance company to do on Mondays? . It is far better to flag more claims as positive (likely fraud) for further investigation than to miss some of the fraud and pay out cash that should have never been paid. A false positive (flagged for additional scrutiny as possibly fraud, but the customer loss was real) can likely be cleared up by assigning an experienced worker, who can insist on a police report, request building security video, etc. A false negative (accepting a fraudsters false claim and paying out in cash) is pure loss to the insurance company, and encourages more fraud. (4)

   It possible, theoretically, for a system to have both **perfect recall and precision**. However for all practical purposes, it is not possible . (4)

e. Once a dataset's dimensionality has been reduced using an algorithm, **it is almost always impossible to perfectly reverse the operation**, as some information gets lost during dimensionality reduction. While some algorithms (such as PCA) have a simple reverse transformation procedure that can reconstruct a dataset relatively similar to the original, other algorithms do not. (20)