Exam: Artificial Intelligence – Algorithms and Application

Module Exam

Winter 2023/2024 Date: 04.04.2024

TECHNISCHE UNIVERSITÄT DARMSTADT

WIRTSCHAFTS INFORMATIK

Important Information

- Please check your exam copy for completeness.
  It covers **20 pages** (cover sheet included).
- Fill out the cover sheet immediately after receiving the exam.
- Use only the examination paper to solve the tasks. If you do not have enough space, you can receive additional paper during the examination. Additional papers must also be marked with your name and matriculation number.
- Please leave a **correction margin of 3 cm**.
- You have a total of **90 minutes** to complete the exam.
- Except for a **non-programmable calculator, no other aids** are allowed in the exam.

We wish you much success!

Please fill out clearly in block letters.

First Name ..... Last Name ..... Seat No. .....

Matr. No. ..... Course of Study ..... ❑ Master ❑ Diplom

Repeater:
❑ yes ❑ no

Section	Max. Points	Achieved Points
1	36	
2	24	
3	30	
Sum	90	

Exam Review („Klausureinsicht“):

(do not fill out before the review)

I have reviewed the corrected exam:

☐ There are no complaints about the correction.

☐ Complaints about the correction exist (see additional sheet).

Date: ……..

Signature: ……....

First Name..... Last Name..... Matr. No.....

1 Basic Concepts and Algorithms (36 Points)

1.1 Is **feature scaling** in general **required after** a **normalization** has been applied? Please briefly **explain** your **decision**. (2 P)

1.2 **Please** briefly **explain** what kind of problem the ***traveling salesman problem*** represents using the **national park example** from the lecture. (2 P)

1.3 Please provide an example of a "***CAPTCHA***". Please also **explain how** a "***CAPTCHA***" is **related** to **Artificial Intelligence** and **name three ability domains** of **Artificial Intelligence** that are **needed** to **solve it**. (3 P)

1.4 Suppose your favorite burger chain wants to open a new restaurant. To minimize interference with existing burger restaurants, the new restaurant should be as far away from the nearest restaurant as possible. **Which** (data visualization) **plot** or **model component** would you use to **solve this problem**? Please briefly **explain** your **decision**. (3 P)

1.5 Please **explain**: **What** tends to **happen** to the **training error** in a **linear model**, (e.g., a linear regression) as the **polynomial degree increases**? (1 P)

1.6 Please **explain** in your own words the **benefit** of the **bias** in **perceptron modeling**. (1 P)

1.7 Please **explain** the problem of "**Model Autophagy Disorder (MAD)**" in the **context** of **LLM**/**ChatGPT**. (3 P)

First Name..... Last Name..... Matr. No.....

1.8 Please **explain**: What is the **difference** between **Data Understanding** and **Business Understanding**? **Why** is there an **interaction between** these **two phases**? (2 P)

1.9 Please **explain** the **exploitation-exploration dilemma** in the **context** of how you decided between different **study programs** and **university courses**. (4 P)

1.10 Please briefly **define it** and **explain its structure**: What is a ***data frame*** in Python? (3 P)

1.11 Please briefly **explain** the **difference** between ***random sampling*** and ***random walk*** in search algorithms by comparing both concepts. (2 P)

1.12 Please briefly **explain** the **central statements** of the ***Moravec's Paradox***.
Please **also provide an example** of your own of today's Artificial Intelligence. (4 P)

1.13 Please **describe** the **task**, **performance measure**, and **experience** for the following machine learning problems (6 P)

a) Build an information system that detects broken components of an assembly line.
b) Build a bot that can win an Age of Empires 2 match against you.
c) Build a fraud detection system for your online shop.
d) Build a trading bot for a crypto market.

First Name..... Last Name..... Matr. No.....

2 Application of Machine Learning Algorithms (30 Points)

Consider Table 1. Table 1 represents various **customers** of a **financial services company** that **assesses creditworthiness**. Table 1 includes **two features**: "**SCORE_A**" and "**SCORE_B**". Each row in Table 1 also has a **class label** that is either "**TRUE**" or "**FALSE**" and is stored in the third **column** called "**CREDIT**".

Table 1. Customer Data.

ID	SCORE_A	SCORE_B	CREDIT
1	40	20	FALSE
2	50	50	TRUE
3	60	90	TRUE
4	10	25	FALSE
5	70	70	TRUE
6	60	10	FALSE
7	25	80	TRUE

2.1 The company has asked you to prepare a management presentation. Please **visualize** the data captured in Table 1 in a two-dimensional scatterplot. You can use symbols to visualize the CREDIT label. (5 P)

2.2 The company then asks you to **predict** the **creditworthiness** of the following new **customer** "$\mathbf{c_1}$".

c_1 = {SCORE_A: 20, SCORE_B: 35, CREDIT: ?}

Please **use** the **KNN algorithm** with ***k = 5*** and the ***Euclidean distance*** to **predict** the **creditworthiness** of the **above customer** based on the **data** in **Table 1**. Please **explain** your **calculations**. (5 P)

First Name..... Last Name..... Matr. No.....

2.3 Next, you apply a pre-trained classification tree to the data in Table 1 to predict each customer's CREDIT label again. The predictions produced by the classification tree are shown in the following Table 2.

Table 2. Predicted CREDIT Labels.

ID	Predicted CREDIT
1	FALSE
2	FALSE
3	TRUE
4	TRUE
5	TRUE
6	FALSE
7	FALSE

To evaluate the tree's prediction performance based on the predicted labels in Table 2, please **compute** a **confusion matrix** and the following measures: **Accuracy**, **Precision**, **Recall**. (10 P)

2.4 Please **explain**: **Why** is it **not** a **good idea** to **evaluate** your **classification model** on **training data** like in task 2.3? (2 P)

2.5 As a next step, you want to improve your classification model. **How many models** will be **built** and **tested** when you use **grid search**, assuming that you consider the following **three hyperparameters?**

- *max_depth* = {2, 3, 5, 10, 20}
- *measure* = {gini, entropy}
- *min_samples_leaf* = {5, 10, 20, 50}

Please also **explain** your **calculations**. (2 P)

First Name..... Last Name..... Matr. No.....

2.6 Please **fill** the six **missing parts** of the following **Python code** to **run** the **grid search** mentioned in the previous task. (3 P)

```
GridSearchCV(
    cv=_____________________,
    estimator=DecisionTreeClassifier(random_state=_____________),
    n_jobs=_____________________,
    param_grid={
        'criterion': [_____________________],
        'max_depth': [_____________________],
        'min_samples_leaf': [_____________________]},
    scoring='accuracy', verbose=1)
```

2.7 Please **name one method** other than grid search that can be used for **parameter tuning**. (1 P)

2.8 Please **explain** ***Wolpert's free lunch theorem*** using the **classification problem** of this section as an example. (2 P)

First Name..... Last Name..... Matr. No.....

3 Programming with Python (24 Points)

One of your colleagues prepared the following Python notebook for an analysis of car failures. Please consider the following code when answering the next questions:

Python Code

```
1 from pandas import read_excelll
2 from sklearn.tree import DecisionTreeClassifier
3 from sklearn.ensemble import RandomForestClassifier
4 from sklearn.model_selection import train_test_split
5 from sklearn.metrics import confusion_matrix
6 from sklearn.metrics import accuracy_score
7
8 dataset = read_excel(open("car_maintenance.xlsx", "rb"))
```

Out[3]:

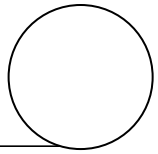
	PART_1023	PART_99	PART_02	OIL	CHECK_STATUS	FOLLOW-UP
0	1	2	1	0.25	acc	yes
1	2	1	1	0.15	unacc	no
2	1	2	1	0.25	acc	yes
3	1	1	1	0.15	good	yes
4	2	1	1	0.15	vgood	no
5	1	1	1	0.50	acc	yes

Python Code

```
1 dataset.describe()
```

Out[5]:

	PART_1023	PART_99	PART_02	OIL
count	31.000000	31.000000	31.0	31.000000
mean	1.580645	33.387097	1.0	0.180645
std	0.672022	179.210617	0.0	0.160040
min	1.000000	1.000000	1.0	0.050000
25%	1.000000	1.000000	1.0	0.100000
50%	1.000000	1.000000	1.0	0.150000
75%	2.000000	1.000000	1.0	0.200000
max	3.000000	999.000000	1.0	0.800000



First Name..... Last Name..... Matr. No.....

Note: The notebook continues on the next page.

Python Code

```
1  # remove outliers
2  # thr = dataset["PART_99"].quantile(0.999)
3  thr = 3
4  dataset = dataset[dataset["PART_99"] < thr]
5  dataset.describe()
6
7
8
9  # analytics
10 X = dataset.loc[:, "PART_1023":"CHECK_STATUS"]
11 Y = dataset.loc[:, "FOLLOW-UP"]
12 X_train, X_test, Y_train, Y_test = train_test_split(…)
13
14 clf = DecisionTreeClassifier()
15 clf = clf.fit(X_train, Y_train)
16
17 prediction = clf.predict(X_test)
18
19 confusion_matrix(Y_test, prediction)
20
21
```

3.1 **Which machine learning model** is **trained** in the above **code**? (1 P)

First Name..... Last Name..... Matr. No.....

3.2 Please **explain what** the **`"rb"`** in the **`open()` function**, which is part of the `read_excel()` command, **means**. (1 P)

3.3 Please **explain** the **result** of the **`dataset.describe()` command** displayed by **`OUT[5]`**. (4 P)

3.4 Unfortunately, some part of the code is missing in the third code cell. Please complete the code by **adding** a **command** to **delete column `"PART_02"`** in **line 7**. (2 P)

Note: You can write the code directly above in line 7 or in the space below this task (assuming it would be written in line 7).

3.5 Please **write** the **code** for the **missing parameters** of the **`train_test_split()` function** in line 12 of the third code cell (indicated by "…") to **split** your **dataset** with a **`test_size`** of **0.3.** (2 P)

3.6 When you run the above code cells, you get an `ImportError`. Please **explain** or **provide corrected code** on how you can fix the error in the above code cells. (2 P)

First Name..... Last Name..... Matr. No.....

3.7 Please **write code** to **implement another classification model** of your choice (e.g., from `scikit` learn). Your **code** should **include** the code to **create** an **object**, **train** it, and **use** the **`predict()`** function on it to **predict** the **class variable**. Please feel free to use the model you used in the lectorial. (6 P)

3.8 Please e**xplain** the **purpose** of **`confusion_matrix()`** in the **code**. **Why** do you **need** it and **what** does **it tell you**? (4 P)

3.9 Please **explain**: **What** could be a **possible step** to **improve** the **performance** of the above **classification model**? Please **provide** some **Python**/**pseudo code** to **implement** this **step**. (2 P)