Lab Class ML:I, ML:II

Until Wednesday, Nov. 6th, 2024, 11:59 pm CET, solutions to the following exercises must be submitted as one zip-file named `ML24-ex1-group<your-group-number>.zip` via Moodle:
1, 2, 3, 5, and 6.

Exercise 1 : Machine Learning (general) (1.5+1.5=3 Points)

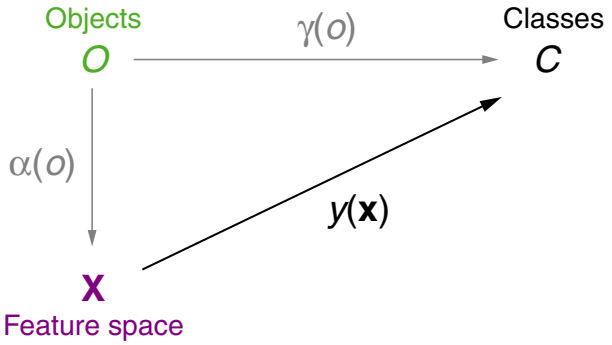
(a) Define the terms “supervised learning”, “unsupervised learning”, and “reinforcement learning”.

(b) Determine the learning paradigm (supervised, unsupervised, reinforcement) for the following tasks.
Note: If more than one learning paradigm is possible, select one and provide a brief (1 sentence) explanation.

- (b1) Sentiment analysis (determine if a text has positive or negative sentiment)
- (b2) Data compression
- (b3) Self-driving cars
- (b4) Personalized content recommendation
- (b5) Spam filtering
- (b6) Sorting fruits in a basket by type

Exercise 2 : Specification of Learning Tasks (3 Points)

The following picture from the lecture slides describes the relationship between Real World and Model World, when it comes to the specification of learning tasks.



Assume you are building a machine learning system that predicts whether a given mushroom is poisonous or edible. For the following list, decide which symbol from the picture most closely matches the given list item:

(a) A pile of Mushrooms.

(b) A table with the columns “size”, “weight”, and “color”, as well as one row for each mushroom, and the respective measurements in the cells.

(c) A human mushroom expert who can tell whether any mushroom you show them is poisonous or edible.

(d) A device that measures size, weight and color of a mushroom.

(e) The set $\{\mathsf{Poisonous}, \mathsf{Edible}\}$

(f) The machine learning system that you are trying to build.

Exercise 3 : P Data Annotation and Feature Engineering (3+1+0=4 Points)

Throughout the programming labs, we will work on the task of text regression: given a text, predict whether it was written by a human or an AI. The dataset comprises multiple text genres, such as news articles, Wikipedia intro texts, or fanfiction.

The main purpose of this exercise is to get familiar with the dataset that we will be using throughout the labs. Each group member has to read 5 texts and label each of them as written by a human or a large language model (LLM).

(a) Annotate your part of the dataset. To get the annotation data, use the "[annotation]" button on Moodle schedule. Click on "Guidelines" and read them carefully before annotating.

After annotating, click "save" to download your annotations. Add the downloaded json files to the zip file that you submit to Moodle.

Note: To get points for this exercise, all group members have to submit their annotations.

(b) Look back to the exercise 2. Which symbol in the picture corresponds to the role that **you** are playing? Which symbol corresponds to the functions that you implement in this exercise?

(c) Implement a program to extract the features from texts, such as the number of sentences and the average word length. Come up with 3 additional features that you think might be useful for this task. The program should take a text as an input and output a CSV file with columns corresponding to the five features and rows corresponding to texts.

Exercise 4 : Rule-Based Learning (0 Points)

The examples of a training set for a classification problem are described by the values of the attributes $A_1, \ldots, A_p$ and the related concepts $C = \{0, 1\}$. For the attributes $A_1, \ldots, A_p$ there are in each case $m_1, \ldots, m_p$ values, e.g. $a_{i,1}, \ldots, a_{i,m_i}$ for A_i. The hypothesis space contains the conjunctions of restrictions for the attributes: "A_1 has value a_{1,j_1} and ... and A_p has value a_{p,j_p}". A question mark in a hypothesis denotes a wildcard for the respective attribute domain. The hypothesis space does also contain the empty hypothesis $\perp$, which assigns all examples to the concept 0.

(a) Determine the number $n(p)$ of all possible examples for this problem.

(b) Determine the number $|H_p|$ of different hypotheses.

(c) How will the above answers change, if an additional attribute A_{p+1} with m_{p+1} values is added? Derive a recursion formula.

Exercise 5 : Rule-Based Learning (Practice) (2+4+1=7 Points)

Given is the following training set D, which you have obtained as co-driver by observing your friend:

	Weekday	Mother-in-the-car	Mood	Time of day	run-a-red-light
1	Monday	no	easygoing	evening	yes
2	Monday	no	annoyed	evening	no
3	Saturday	yes	easygoing	lunchtime	no
4	Monday	no	easygoing	morning	yes

Let the set H contain hypotheses that are built from a conjunction of restrictions for attribute-value combinations; e. g. $(\textit{Monday}, \textit{yes}, ?, ?)$.

(a) Apply the Find-S algorithm for the example sequence 1, 2, 3, 4.

(b) Apply the Candidate-Elimination algorithm for the example sequence 1, 2, 3, 4, and identify the boundary sets H_S and H_G.

(c) What is the version space H_D for this example?

Exercise 6 : Rule-Based Learning (Background) (1+1+1=3 Points)

(a) Can a version space H_D contain hypotheses that are neither in the set H_S nor in the set H_G? If so, how?

(b) For any two hypotheses $y_1(), y_2(), y_1() \neq y_2()$, from the set H_S of a version space H_D holds (check all that apply):

- ☐ $(y_2() \geq_g y_1()) \vee (y_1() \geq_g y_2())$
- ☐ $(y_2() \geq_g y_1()) \wedge (y_1() \geq_g y_2())$
- ☐ $(y_2() \not\geq_g y_1()) \vee (y_1() \not\geq_g y_2())$
- ☐ $(y_2() \not\geq_g y_1()) \wedge (y_1() \not\geq_g y_2())$

(c) Which of the two algorithms Find-S and Candidate-Elimination has a stronger inductive bias? Explain your answer.