INTRODUCTION MACHINE LEARNING

EXERCISE 1

Bauhaus-Universität Weimar

TEACHER:
Johannes Kiesel

GROUP:
Group 16

SUBMITTED BY:

Aaron Perez Herrera
Cesar Fernando Gamba Tiusaba
Chun Ting Lin
Olubunmi Emmanuel Ogunleye

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

a. Define the terms "supervised learning", "unsupervised learning", and "reinforcement learning".

Supervised Learning: Type of learning that uses labeled data to train a model, with the target of being able to differentiate or map new inputs that you give it to the model and is going to the obtain correct outputs based on the training data.

Unsupervised Learning: Type of learning in which the model is trained with an unlabeled dataset, so the model must identify pattern in the data (like clustering) without the guidance of the user.

Reinforcement Learning: the model learns to make decisions by taking actions in an environment, the model receives feedback from the user in form of "Penalties" or "rewards" based on the actions of the model. The model adjusts its behavior to maximize the correct predictions ("rewards").

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.*

1. **Sentiment analysis**: **Supervised Learning**
   This task involves labeled data (attributes relevant with positive or negative sentiment) to train a model.

2. **Data compression**: **Unsupervised Learning**
   This task identifies patterns in the data without predefined labels, aiming to reduce the size of data while retaining essential information and for this is good find patterns of text.

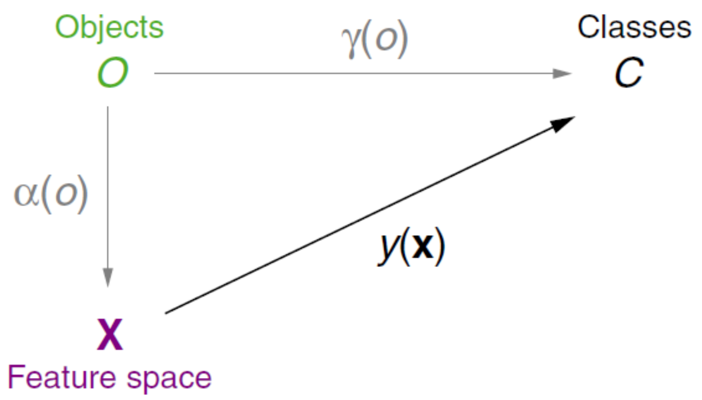
3. **Self-driving cars**: **Reinforcement Learning / Supervised Learning**
   The car learns to navigate and make driving decisions based on feedback from its environment and user feedback to maximize safety and efficiency. So, it improves in search of better result.

4. **Personalized content recommendation**: **Supervised Learning**
   This task often uses labeled data (user preferences) to train a model that predicts what content a user is likely to enjoy.

5. **Spam filtering**: **Supervised Learning**
   This task relies on labeled emails (spam or not spam) to train a model to classify incoming messages.

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**
"O" - Which represent the Objects or Data sets.

b. **A table with the column's "size", "weight", and "color", as well as one row for each mushroom, and the respective measurements in the cells**
"X" - Which represents or describes the feature space or attributes of each mushroom.

c. **A human mushroom expert who can tell whether any mushroom you show them is poisonous or edible**
"ɣ(O)" - Is known as the ideal classifier capable of discerning if a mushroom is edible.

d. **A device that measures size, weight and color of a mushroom**
"α(o)" - Is known as model formation function that helps to construct the feature space of the attributes of the mushrooms.

e. **The set {Poisonous, Edible}**
"C" is known as the Class or Performance Metric.

f. **The machine learning system that you are trying to build.**
"y(x)" - Is known as the model function which is used for learning edible and poisonous mushrooms.

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 must 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.
Check in the Folder "Answers AI or Human texts". You can find there the 4 answers of each team member.

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?
The role fulfilled is the "Ideal Classifier" "ɣ(O)". Because we provide a tag and classification to a dataset. (Provide to each "O" and "C").

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.
Check the python file "extractfeatures.py". there you will find our answer for this point.

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 A1, . . . ,Ap and the related concepts C = {0, 1}. For the attributes A1, . . . ,Ap there are in each case m1, . . . ,mp values, e.g. ai,1, . . . , ai,mi for Ai. The hypothesis space contains the conjunctions of restrictions for the attributes: “A1 has value a1,j1 and . . . and Ap has value ap,jp ”. A question mark in a hypothesis denotes a wildcard for the respective attribute domain. The hypothesis space does also contain the empty hypothesis ⊥, which assigns all examples to the concept 0.

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

$n_p = m_1 \times m_2 \times \ldots \times m_p$

b. Determine the number |Hp| of different hypotheses.

$A'_i = A_i + 1 = m_i + 1$

$H_p = A'_1 \times A'_2 \times \ldots \times A'_p = m_1 + 1 \times m_2 + 1 \times \ldots \times m_p + 1 \qquad (add\ "?" \rightarrow +1)$

$H_p = \left(m_1 + 1 \times m_2 + 1 \times \ldots \times m_p + 1\right) + 1$

c. How will the above answers change, if an additional attribute Ap+1 with mp+1 values are added? Derive a recursion formula.

$n_{(p+1)} = n_p \times m_{p+1}$

$H_p = H_p \times A'_{p+1} = H_p \times (m_{p+1} + 1)$

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. (Monday, yes, ?, ?).

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

$H_{S0} = \{(\perp, \perp, \perp, \perp)\}$

$H_{S1} = \{("Monday", "no", "easygoing", "evening")\}$

$H_{S2} = \{("Monday", "no", "easygoing", "evening")\}$

$H_{S3} = \{("Monday", "no", "easygoing", "evening")\}$

$H_{S4} = \{("Monday", "no", "easygoing", ?)\}$

$H_F = H_{S4} = \{("Monday", "no", "easygoing", ?)\}$

b. Apply the Candidate-Elimination algorithm for the example sequence 1, 2, 3, 4, and identify the boundary sets HS and HG.

$H_{S0} = \{(\perp, \perp, \perp, \perp)\}$
$H_{S1} = \{("Monday", "no", "easygoing", "evening")\}$
$H_{S2} = \{("Monday", "no", "easygoing", "evening")\}$
$H_{S3} = \{("Monday", "no", "easygoing", "evening")\}$
$H_{S4} = \{("Monday", "no", "easygoing", ?)\}$

$L.V.S \rightarrow No\ extra\ Hypotheses$

$H_{G4} = \{(\text{"Monday"}, ?, \text{"easygoing"}, ?), (?, \text{"no"}, \text{"easygoing"}, ?)\}$
$H_{G3} = \{(\text{"Monday"}, ?, \text{"easygoing"}, ?), (?, \text{"no"}, \text{"easygoing"}, ?), (?, ?, "easygoing", "evening")\}$
$H_{G2} = \{(?, ?, "easygoing", ?)\}$
$H_{G1} = \{(?, ?, ?, ?)\}$
$H_{G0} = \{(?, ?, ?, ?)\}$

$H_D = \{(\text{"Monday"}, \text{"no"}, \text{"easygoing"}, ?), (\text{"Monday"}, ?, \text{"easygoing"}, ?), (?, \text{"no"}, \text{"easygoing"}, ?)\}$

c. What is the version space HD for this example?
The Space Hypotheses for the "Candidate-Elimination" Algorithms is:
$H_D = \{(\text{"Monday"}, \text{"no"}, \text{"easygoing"}, ?), (\text{"Monday"}, ?, \text{"easygoing"}, ?), (?, \text{"no"}, \text{"easygoing"}, ?)\}$

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

a. Can a version space HD contain hypotheses that are neither in the set HS nor in the set HG? If so, how?
Yes. It can contain different hypotheses that are $H_D = H_P + H_G + LVS$. Basically, is the union of General hypothesis, Specific Hypotheses and the space of intersection between them (LVS). This LVS are hypotheses that are general enough to cover positive examples, but specific enough to exclude the negative ones.

b. For any two hypotheses y1(), y2(), y1() ≠ y2(), from the set HS of a version space HD holds (check all that apply):

$(y_2() \geq_g y_1()) \lor (y_1() \geq_g y_2())$

$(y_2() \geq_g y_1()) \land (y_1() \geq_g y_2())$

$(y_2() \not\geq_g y_1()) \lor (y_1() \not\geq_g y_2())$ CORRECT

$(y_2() \not\geq_g y_1()) \land (y_1() \not\geq_g y_2())$ CORRECT

c. Which of the two algorithms Find-S and Candidate-Elimination has a stronger inductive bias? Explain your answer.
The algorithm "Candidate Elimination" have a stronger BIAS. This is because it consists of a set of hypotheses, considering both "positive" and "negative" examples to refine the boundaries. On the other hand, the algorithm "Find-S" find a single specific hypothesis, being less robust for defining boundaries.