Machine Learning
Tutorial 1

ENSIA

February 2024

Exercise 1:

1. What is predictive data analytics?
2. What is supervised machine learning?
3. Machine learning is often referred to as an ill-posed problem. What does this mean?
4. The following table (Table 1) lists a dataset from the credit scoring domain. Underneath the table we list two prediction models consistent with this dataset, Model 1 and Model 2.

Table 1: Loan-Salary Data

ID	Occupation	Age	Ratio	Outcome
1	industrial	39	3.40	default
2	industrial	22	4.02	default
3	professional	30	2.7	repay
4	professional	27	3.32	default
5	professional	40	2.04	repay
6	professional	50	6.95	default
7	industrial	27	3.00	repay
8	industrial	33	2.60	repay
9	industrial	30	4.5	default
10	professional	45	2.78	repay

(a) Which of these two models do you think will generalize better to instances not contained in the dataset?

(b) Propose an inductive bias that would enable a machine learning algorithm to make the same preference choice that you made in Part (a).

(c) Do you think that the model that you rejected in Part (a) of this question is overfitting or underfitting the data?

Algorithm 1 Model 1

```
if LOAN-SALARY RATIO > 3.00 then
    OUTCOME = default
else
    OUTCOME = repay
end if
```

Algorithm 2 Model 2

```
if AGE = 50 then
    OUTCOME = default
else if AGE = 39 then
    OUTCOME = default
else if AGE = 30 and OCCUPATION = industrial then
    OUTCOME = default
else if AGE = 27 and OCCUPATION = professional then
    OUTCOME = default
else
    OUTCOME = repay
end if
```

Exercise 2:

1. What is meant by the term inductive bias?

2. How do machine learning algorithms deal with the fact that machine learning is an ill-posed problem?

3. What can go wrong when an inappropriate inductive bias is used?

4. It is often said that 80% of the work done on predictive data analytics projects is done in the Business Understanding, Data Understanding, and Data Preparation phases of CRISP-DM, and just 20% is spent on the Modeling, Evaluation, and Deployment phases. Why do you think this would be the case?

Exercise 3:

1- The following table (Table 2) lists a dataset of five individuals described via a set of stroke risk factors and their probability of suffering a stroke in the next five years. This dataset has been prepared by an analytics team who are developing a model as a decision support tool for doctors. The goal of the model is to classify individuals into groups on the basis of their risk of suffering a stroke STROKE RISK. In this dataset there are three categories of risk: low, medium, and high. All the descriptive features are Boolean, taking two levels: true or false.

(a) How many possible models exist for the scenario described by the features in this dataset?

(b) How many of these potential models would be consistent with this sample of data?

Table 2: Stroke Risk Factors Dataset

ID	High Blood Pressure	Smoker	Diabetes	Heart Disease	Stroke Risk
1	true	false	true	true	high
2	true	true	true	true	high
3	true	false	false	true	medium
4	false	false	false	false	low
5	true	true	true	false	high

2- You are using U.S. census data to build a prediction model. On inspecting the data you notice that the RACE feature has a higher proportion of the category White than you expected. Why do you think this might be?

3- Why might a prediction model that has very high accuracy on a dataset not generalize well after it is deployed?