Module MCS 016
Machine Learning for Cyber-Security & Artificial Intelligence

In this course we will discuss methods and algorithms that allow computer systems to "learn" how to solve a problem based on past experience. Note that machine learning algorithms are not designed to solve a problem that be can be solved in a precise way (through specific algorithms). Instead, machine learning algorithms are developed to identify patterns present in past experiences (where the problem in focus has been successfully solved) and, using these patterns to solve new instances of the same problem.

We will see that past experiences can be represented in many ways, and when they are represented based on databases (one of the most common representations) the methods and algorithms are called "data mining methods". Thus, it can be said that data mining is a form of machine learning application, in which past experiences are represented in the form of (structured) databases.

In this class, exercises will be performed to show an analogy about how human beings "learn" how to solve problems (based on past experience) and the way how machines perform such learning. So, suppose a friend asks your help for choosing which contact lens is right for him to start using. Your first answer should be: ask a doctor as contact lenses cannot be prescribed without proper ophthalmological knowledge and authority.

Suppose, however, that after your colleague has been examined by a specialist doctor, you two try (just for fun) to identify (in a simplistic manner) how would it be possible to identify which contact lens would be suitable for random cases of patients in an ophthalmological clinic. Suppose you can find a database on the internet containing data from medical records of 23 patients. The data is simplified and for each of the 23 patients there is information on the patient identification (ID), age (categorized in terms of ophthalmological: young, pre-presbyopic and prebiotic), the official diagnosis (myopic or hyperopic), the presence of astigmatism (yes or no), the patient's tear production (normal or reduced), and finally, which lens was prescribed (none, gelatinous or rigid).

ID1, young, myopic, no, reduced, none
ID2, young, myopic, no, normal, gelatinous
ID3, young, myopic, sim, reduced, none
ID4, young, myopic, sim, normal, rigid
ID5, young, hipermetrope, no, reduced, gelatinous
ID6, young, hipermetrope, no, normal, gelatinous
ID7, young, hipermetrope, sim, reduced, rigid
ID8, young, hipermetrope, sim, normal, rigid
ID9, pre-presbiópico, myopic, no, reduced, none
ID10, pre-presbiópico, myopic, no, normal, gelatinous
ID11, pre-presbiópico, myopic, sim, reduced, none
ID12, pre-presbiópico, myopic, sim, normal, rigid
ID13, pre-presbiópico, hipermetrope, no, reduced, none
ID14, pre-presbiópico, hipermetrope, no, normal, gelatinous
ID15, pre-presbiópico, hipermetrope, sim, reduced, rigid

ID16, pre-presbiópico, hipermetrope, sim, normal, none
ID17, presbiópico, myopic, no, reduced, none
ID18, presbiópico, myopic, no, normal, none
ID19, presbiópico, myopic, sim, reduced, none
ID20, presbiópico, myopic, sim, normal, rigid
ID21, presbiópico, hipermetrope, no, reduced, none
ID22, presbiópico, hipermetrope, no, normal, gelatinous
ID23, presbiópico, hipermetrope, sim, reduced, none

Based on these data obtained from the internet, and based on your human apprentice intuition, answer:

1) which contact lens would you prescribe for other patients who are not listed in the data you obtained on the internet, and which are described below (justify each of your prescriptions)?
    a) ID31, pre-presbyopic, hyperopic, yes, normal,?
    b) ID32, presbyopic, hyperopic, yes, normal,?
    c) ID33, young, myopic, yes, normal,?
    d) ID34, presbyopic, hyperopic, yes, normal,?

2) which of the attributes (patient identification (ID), age, official diagnosis, presence of astigmatism, patient's tear production) is the most significant to help identify the best contact lens? Why? And which is the least significant? Why?

3) do you think that this database would be sufficient for a doctor could be considered fit to prescribe contact lenses contact? Why? Which characteristics would be most important for that you can consider a "perfect" database to learn to solve a problem (mention the influence of the amount of instances, number of attributes, the type of attributes (real, integer, nominal), etc.)?

4) Do you think it would be possible, in a database like the one shown above, that two patients with the same characteristics different prescriptions? Why? Give examples.

5) imagine that the age attribute was a numeric attribute (defining how old the patient is), how would you go about checking the similarity between the age of two patients? Why?

6) Considering that you are not a doctor, instead of based on the patterns present in past experiences, you would have an algorithm better to prescribe contact lenses? Justify your answer.

7) Some undergraduate courses (the medical course at UFSCar, for example) follow the "case-based learning" process. Assuming that such courses did not require theoretical study (which is not true, because students also need to base decisions in each case experienced, through models and theories present in the literature). What do you think of this supposed method? How many cases does a student would I need to experience to graduate?

8) build an algorithm (in pseudo-code, or in some language programming that you are familiar with) capable of responding to question 1 of this activity.

9) build an algorithm (in pseudo-code, or in some language programming that you are familiar with) capable of responding to question 2 of this activity.