DEIS - Departamento de Engenharia Informática e Sistemas
ISEC - Instituto Superior de Engenharia de Coimbra

Knowledge and Reasoning 2020/2021
Pratical Assignment

For the accomplishment of the Practical Work, we propose two different themes. Below you will find a detailed description of each one of them and in Moodle the necessary complementary material will be made available. You must chose only one topic to develop the project.

In Moodle there is a referendum where you can choose the topic to develop. Only one of the students in each working group must select the desired topic.

- The working groups are made up of 2 students;
- The date of delivery of the work is until 23.59 of June 27, 2021;
- The code and all files necessary for the execution and testing of the work, as well as the pdf of the report, must be delivered to Moodle;
- Defenses will take place on the 28th, 29th and 30th of June. Each group will have to register 1 (only 1) of its members in Moodle, in the slots that will be made available in due course;
- The defenses / doubts of the Neural Networks theme will be with Prof. Anabela Simões and the CBR and Diffuse Inference with Prof. Viriato M. Marques;
- Defenses will be remotely, unless otherwise indicated by order of higher authorities;
- The defense of work is mandatory and with the presence of all members of the group;
- The practical work is priced at 10 values (on a scale of 0 to 20).

TOPIC 1 – NEURAL NETWORKS

In this theme, it is intended that students deepen their knowledge of neural networks. The goal is to implement and test different feedforward neural network architectures to correctly classify 10 Greek characters:

$$\alpha \ \beta \ \gamma \ \varepsilon \ \eta \ \theta \ \pi \ \rho \ \psi \ \omega$$

In Moodle, black and white image files are provided, separated by three different folders that should be used in the tasks described below..

NOTE: The images are in the size 3024 x 3024 pixels, which in some computers can lead to very high training time. If it is necessary to resize them, use the functions of the Matlab image processing toolbox. Explain in the report all the pre-processing done to the images.

For this work, the following approach is suggested:

a) [20%]. Using Matlab's image manipulation functions convert the provided images into binary arrays. If you find it necessary, make a previous treatment to the images, such as resizing, or any other that you find relevant.
Start with a layered neuronal network of 10 neurons. Use the network to train the recognition of the characters in Folder_1. This folder contains an image of each character. Use all the examples in training. Test other architectures (topologies), activation and training functions, record the performance values of the different parameters and compare the results obtained.

b) [20%]. Using the base model implemented in point a) make the necessary changes to implement and test various topologies and parameterizations of NN in order to obtain a good performance for the classification of the characters provided in the folder Pasta_2. In this folder there are 10 images of each character.

- Start by using a 70%, 15%, 15% dataset segmentation for training, validation and testing.
- Observe the matrix of confusion, training and test errors.
- Explore and compare various network architectures (number of layers / number of neurons).
- Test different training / activation functions, different segmentations in the division of the examples. Record the results for the various neural networks you test. It is suggested to adapt the Excel file given in the practical classes, to record results and obtain the conclusions.
- Save the best performing neuronal network (s)

c) [25%]. Now use the images from Folder_3. In this folder are 4 images of each character, images that were not used in the previous training. For this task use the best network obtained in b)

- Without training the network, check if the classification given by the NN is correct. Present the results obtained.

- Now re-train the network with just the examples in Folder_3. Test the network separately for the images from Folder_1, Folder_2 and Folder_3. Compare and record the results obtained in each case.
- Re-train the network with all the images provided (Folder_1 + Folder_2 + Folder_3). Test the network for the Folder_1, Folder_2 and Folder_3 images separately. Compare and record the results obtained.

d) [15%]. Manually draw some Greek characters that are similar to the examples used in training the net. Transcribe the drawings to binary matrices. Develop a small program to read a file corresponding to one of these images and apply it to the best neural network obtained in c). What are the results?

e) [20%]. Develop a graphical application in Matlab that allows the user to do the tasks previously developed in an easy and intuitive way:

- Configure the topology of the neural network
- Choose training / activation functions
- Train the neural network
- Record a previously trained neuronal network
- Load a previously trained neuronal network and apply it to a dataset
- Draw a new letter, or load an image file where it is already drawn. Apply a neural network to classify the drawn letter
- View the results of the classification
- Generation / recording of result files if deemed relevant and necessary

f) Write a report describing the work done, the results and main conclusions. Poor report quality can discount up to 50% of the total score obtained in the previous points.

TOPIC 2 – CBR and FUZZY LOGIC

In this topic, it is intended that students experience and understand the philosophy underlying the CBR (Case-based Reasoning) cycle and its possible interconnection with other models such as Mamdani's inference for adapting solutions. The objective is to implement a CBR system for the evaluation of used cars, according to their characteristics.

An Excel file is provided, which is described at the end of this statement, and which will serve as a starting point for the implementation of the initial case library. This file has 26 attributes, of which the last, the 26th, is the value of the car, thus having the role of solving each case, target or target: ***it is this value that is intended to be predicted***.

IMPORTANT NOTE: The Excel file has some unknown values. Choose to delete these records or fill them in manually if you think you can get a reasonable estimate of the missing values.

1. Implement a system based on the CBR paradigm, designed to evaluate a car characterized by attributes 1 to 25 of the dataset described below:

a) (20%) Implement the Retrieve phase, which should include:
   - Reading the case library;
   - The collection of the description of a new case in a text interface or form (of your choice);
   - The definition of appropriate measures to determine the global similarity between the new case and the past cases, taking into account that there are numerical, Boolean and nominal attributes;
   - The possibility of optional filtering of the library by car brand, if possible;
   - The presentation of the N most similar cases ordered in descending order of similarity
   - The development of a small administration application that allows defining:
     - This N-value for the number of similar cases to be presented in the solution;
     - Enter, change or delete a case manually in the case library;
     - Definition of appropriate measures to determine the similarity between nominal and / or ordinal attributes;
     - Assign weights to each of the 25 attributes;

b) (30%) Implement the **Reuse** phase, with the following structure:
   - If the new case is considered “close enough” to a previous one, the price of the known car can be used as an assessment of the new car;
   - Otherwise, the price of the old car must be adapted. To make this adaptation, you must take into account the meaning and implications of each of the attributes that you consider relevant. For example, a Mercedes tends to be worth more than a Fiat; a car with fewer kilometers than another must be worth more; a better car should be worth less, etc. Implement this adaptation function however you want, but taking into account all the factors that naturally influence the car's value.

c) (10%) Implement the **Revise** phase: as you know, this phase involves the intervention of some type of "external teacher" who will confirm or correct the solution proposed by the system. When can this confirmation / correction take place? What it will consist of?

d) (10%) Implement the **Retain** phase: in this phase, new cases are retained that contain some lesson, something new. Although all query cases must be retained in a "historical file" of cases (raw cases), the retention in the library will only be of selected cases. The following implementation is suggested:
   - The inclusion, in the administration application referred to in a), of an option "Update of the Case Library"
   - In this option, the system should show the new cases contained in the historical file and that are in a condition to be retained, giving the administrator the option of retaining them or not in the case library.

2. (30%) Reimplement the **Reuse** phase, adjusting the price, when necessary, as follows:
   - For each numeric attribute liable to be fuzzified, define appropriate linguistic terms and membership functions;
   - Create a fuzzy inference rules designed to adapt the price such as, for example: "if the car is a lot less years old than the one in the past case, then it should be worth a lot more" (this naturally implies also defining linguistic terms and belonging functions for the conclusion of the rules: "much less value, less value… much more value…")
   - Adapt the price using Mamdani Inference (instead of the system implemented in version 1.)

3. Write a report of the work done. Poor report quality can discount up to 50% of the total score obtained in the previous points.

Excel file Attribute Description

This data set consists of three types of entities:

1. The specification of an auto in terms of various characteristics
2. Its assigned insurance risk rating
3. Its normalized losses in use as compared to other cars.

The second rating (2.) corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

The third factor (3.) is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year.

NOTE: Several of the attributes in the database could be used as a "class" attribute.

Number of Instances: 205

Number of Attributes: 26 total
   -- 15 continuous
   -- 1 integer
   -- 10 nominal

Attribute Information:

	Attribute:	Attribute Range:
1.	symboling:	-3, -2, -1, 0, 1, 2, 3.
2.	normalized-losses:	continuous from 65 to 256.
3.	make:	alfa-romero, audi, bmw, chevrolet, dodge, honda,isuzu, jaguar, mazda, mercedes-benz, mercury,mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4.	fuel-type:	diesel, gas.
5.	aspiration:	std, turbo.
6.	num-of-doors:	four, two.
7.	body-style:	hardtop, wagon, sedan, hatchback, convertible
8.	drive-wheels:	4wd, fwd, rwd.
9.	engine-location:	front, rear.
10.	wheel-base:	continuous from 86.6 120.9.
11.	length:	continuous from 141.1 to 208.1.
12.	width:	continuous from 60.3 to 72.3.
13.	height:	continuous from 47.8 to 59.8.
14.	curb-weight:	continuous from 1488 to 4066.
15.	engine-type:	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16.	num-of-cylinders:	eight, five, four, six, three, twelve, two.
17.	engine-size:	continuous from 61 to 326.
18.	fuel-system:	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19.	bore:	continuous from 2.54 to 3.94.
20.	stroke:	continuous from 2.07 to 4.17.
21.	compression-ratio:	continuous from 7 to 23.
22.	horsepower:	continuous from 48 to 288.
23.	peak-rpm:	continuous from 4150 to 6600.
24.	city-mpg:	continuous from 13 to 49.
25.	highway-mpg:	continuous from 16 to 54.
26.	price:	continuous from 5118 to 45400.

8. Missing Attribute Values: (denoted by "?")

Attribute #:	Number of instances missing a value:
2.	41
6.	2
19.	4
20.	4
22.	2
23.	2
26.	4