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

Conhecimento e Raciocínio 2020/2021

Trabalho Prático

Para a realização do Trabalho Prático propõem-se 2 temas. Mais abaixo encontra-se a descrição detalhada de cada um deles e no Moodle será disponibilizado o material complementar necessário.

No Moodle encontra-se um referendo para que possam escolher o tema que pretendem desenvolver. Apenas um dos alunos de cada grupo de trabalho deve selecionar o tema pretendido.

- Os grupos de trabalho são de 2 alunos;
- A **data única** de **entrega** do trabalho é até às 23.59 do dia **27 de junho de 2021;**
- Devem ser entregues no Moodle o código e todos os ficheiros necessários para a execução e teste do trabalho, bem como o **pdf do relatório**;
- As defesas serão nos dias **28, 29 e 30 de Junho**. Cada grupo terá de fazer a inscrição de 1 (**só 1**) dos seus elementos no Moodle, nos slots que para isso serão oportunamente disponibilizados;
- As defesas/dúvidas do tema Neural Networks serão com a Prof. Anabela Simões e do tema CBR e Inferência Difusa com o Prof. Viriato M. Marques;
- As defesas serão de forma remota, salvo indicação em contrário por ordem de instâncias superiores;
- A defesa do trabalho é obrigatória e com a presença de todos os membros do grupo;
- O trabalho prático tem a cotação de 10 valores (numa escala de 0 a 20).

TEMA 1 – REDES NEURONAIS

Neste tema pretende-se que os estudantes aprofundem os seus conhecimentos sobre redes neuronais. O objetivo consiste na implementação e teste de diferentes arquiteturas de redes neuronais feedforward para classificar corretamente 10 carateres gregos:

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

No Moodle são fornecidos os ficheiros de imagens a preto e branco separadas por três pastas diferentes que devem ser usadas nas tarefas descritas de seguida.

NOTA: As imagens encontram-se no tamanho 3024 x 3024 pixeis, que em alguns computadores poderá levar a tempos e velocidades de treino muito elevadas. Caso seja necessário redimensioná-las use as funções da toolbox de *image processing* do Matlab. Explique no relatório todo o pré-processamento feito às imagens.

Para este trabalho sugere-se a seguinte abordagem:

a) [20%]. Usando as funções de manipulação de imagem do Matlab converta as imagens fornecidas em matrizes binárias. Se achar necessário faça um tratamento prévio às imagens, como redimensionamento, ou outro que achar relevante.
Comece por uma rede neuronal de uma camada com 10 neurónios. Use a rede para treinar o reconhecimento dos caracteres da Pasta_1. Nesta pasta encontra-se uma imagem de cada caracter. Use todos os exemplos no treino. Teste outras arquiteturas (topologias), funções de ativação e de treino, registe os valores de desempenho das diferentes parametrizações e compare os resultados obtidos.

b) [20%]. Usando o modelo base implementado na alínea a) faça as alterações necessárias para implementar e testar várias topologias e parametrizações de RN de forma a obter um bom desempenho para a classificação dos caracteres fornecidos na pasta Pasta_2. Nesta pasta encontram-se 10 imagens de cada caracter.

- Comece por usar uma segmentação do dataset de 70%, 15%, 15% para treino, validação e teste.
- Observe a matriz de confusão, erros de treino e teste.
- Explore e compare várias arquiteturas da rede (número de camadas/nº de neurónios).
- Teste diferentes funções de treino/ativação, diferentes segmentações na divisão dos exemplos. Registe os resultados para as várias redes neuronais que testar. Sugere-se a adaptação do ficheiro Excel dado nas aulas práticas, para registar resultados e obter as conclusões.
- Grave a(s) rede(s) neuronal(ais) com melhor(es) desempenho(s).

c) [25%]. Utilize agora as imagens da Pasta_3. Nesta pasta encontram-se 4 imagens de cada caracter, imagens que não foram usadas no treino anterior.
Para esta tarefa use a melhor rede obtida em b)

- Sem treinar a rede verifique se a classificação dada pela RN é correta. Apresente os resultados obtidos.
- Agora volte a treinar a rede só com os exemplos da Pasta_3. Teste a rede separadamente para as imagens da Pasta_1, Pasta_2 e Pasta_3. Compare e registe os resultados obtidos em cada caso.
- Volte a treinar a rede com todas as imagens fornecidas (Pasta1 + Pasta_2 + Pasta_3). Teste a rede para as imagens da Pasta_1, Pasta_2 e Pasta_3 em separado. Compare e registe os resultados obtidos.

d) [15%]. Desenhe manualmente alguns caracteres gregos que apresentem semelhanças com os exemplos usados no treino da rede. Transcreva os desenhos para matrizes binárias. Desenvolva um pequeno programa para ler um ficheiro correspondente a uma destas imagens e aplicá-lo à melhor rede neuronal obtida em c). Quais os resultados?

e) [20%]. Desenvolva uma aplicação gráfica em Matlab que permita ao utilizador fazer as tarefas desenvolvidas anteriormente de forma fácil e intuitiva:

- Configurar a topologia da rede neuronal
- Escolher funções de treino / ativação
- Treinar a rede neuronal
- Gravar uma rede neuronal previamente treinada
- Carregar uma rede neuronal previamente treinada e aplicá-la a um dataset
- Desenhar uma nova letra, ou carregar um ficheiro de imagem onde esta já se encontre desenhada. Aplicar uma rede neuronal para classificar a letra desenhada
- Visualizar os resultados da classificação
- Geração/gravação de ficheiros de resultados se achar relevante e necessário

f) Elabore um relatório do trabalho realizado. Uma má qualidade do relatório pode descontar até 50% na classificação total obtida nos pontos anteriores.

TEMA 2 – CBR e INFERÊNCIA DIFUSA

Neste tema pretende-se que os estudantes experimentem e compreendam a filosofia subjacente ao ciclo CBR (Case-based Reasoning) e a sua possível interligação com outros modelos tais como a inferência de Mamdani para adaptação de soluções. O objetivo consiste na implementação de um sistema CBR para avaliação de automóveis usados, em função das suas características.

É fornecido um ficheiro Excel, que se descreve no final deste enunciado, e que servirá como ponto de partida para a implementação da biblioteca de casos inicial. Esse ficheiro tem 26 atributos dos quais o último, o 26º, é o valor do automóvel, tendo assim o papel de solução de cada caso, alvo ou target: ***é este valor que se pretende prever***.

NOTA IMPORTANTE: O ficheiro Excel tem alguns valores desconhecidos. Opte por eliminar esses records ou preenchê-los manualmente se achar que consegue uma estimativa razoável dos valores em falta.

1. Implemente um sistema baseado no paradigma CBR, destinado a avaliar um automóvel caracterizado pelos atributos 1 a 25 do dataset abaixo descrito:

a) (20%) Implemente a fase de **Retrieve**, que deve incluir:
   - A leitura da biblioteca de casos;
   - A recolha da descrição de um caso novo em interface texto ou formulário (à sua escolha);
   - A definição de medidas adequadas à determinação da semelhança global entre o novo caso e os casos passados, atendendo a que existem atributos numéricos, booleanos e nominais;
   - A possibilidade de filtragem opcional da biblioteca por marca do automóvel, se possível;
   - A apresentação dos N casos mais semelhantes ordenados por ordem decrescente de semelhança
   - O desenvolvimento de uma pequena aplicação de administração que permita definir:
     - Este valor N de número de casos semelhantes a apresentar na solução;
     - Introduzir, alterar ou eliminar um caso manualmente na biblioteca de casos;
     - Definição de medidas adequadas à determinação da semelhança entre atributos nominais e/ou ordinais;
     - Atribuir ponderações a cada um dos 25 atributos;

b) (30%) Implemente a fase de **Reuse**, com a seguinte estrutura:
   - Se o novo caso for considerado "suficientemente próximo" de uma anterior, o preço do automóvel conhecido pode ser usado como avaliação do novo automóvel;
   - Caso contrário, o preço do automóvel antigo tem de ser adaptado. Para realizar essa adaptação deve atender ao significado e implicações de cada um dos atributos que considerar relevantes. Por exemplo, um Mercedes tem tendência a valer mais que um Fiat; um carro com menos quilómetros que outro deve valer mais; um carro mais valho deve valer menos, etc. Implemente esta função de adaptação da forma que quiser, mas atendendo a todos os fatores que, naturalmente, influenciam o valor do automóvel.

c) (10%) Implemente a fase de **Revise**: como sabe, nesta fase dá-se a intervenção de algum tipo de "professor externo" que vai confirmar ou corrigir a solução proposta pelo sistema. Quando poderá ocorrer esta confirmação / correção? Em que consistirá?
d) (10%) Implemente a fase de **Retain**: nesta fase são retidos os casos novos que contêm alguma lição, algo de novo. Embora todos os query cases devam ser retidos num "ficheiro histórico" de casos (*raw cases*) a retenção na biblioteca será apenas de casos selecionados. Sugere-se a seguinte implementação:
    a. Inclusão, na aplicação de administração referida em a), de uma opção "Atualização da Biblioteca de Casos"
    b. Nesta opção o sistema deverá mostrar os casos novos contidos na ficheiro histórico e que estejam em condições de ser retidos, dando a opção ao administrador de os reter ou não na biblioteca de casos.

2. (30%) Reimplemente a fase de **Reuse**, realizando a adaptação do preço, quando necessária, da seguinte forma:
    a. Para cada atributo numérico suscetível de ser *fuzificado*, defina termos linguísticos e funções de pertença adequadas;
    b. Crie regras de inferência destinadas a adaptar o preço tais como, por exemplo: "se o carro tem muito menos anos que o do caso passado, então deve valer muito mais" (isto implica, naturalmente, definir também termos linguísticos e funções de pertença para a conclusão das regras: "muito menos valor, menos valor … muito mais valor …")
    c. Realize a adaptação do preço através da Inferência de Mamdani (em vez do sistema implementado na versão 1.)

3. Elabore um relatório do trabalho realizado. Uma má qualidade do relatório pode descontar até 50% na classificação total obtida nos pontos anteriores.

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