Departamento de Matemática, Universidade de Aveiro

Aprendizagem Computacional — Trabalho Computacional

2023 - 2024

---

1 Enquadramento

Neste trabalho pretende-se que os estudantes apliquem os conhecimentos adquiridos na UC de Aprendizagem Computacional em problemas de índole real. São propostos 4 problemas, 2 de regressão e 2 de classificação, que a seguir se descrevem.

Concrete Data - Regressão nível N1

O betão é o material mais importante na engenharia civil. A sua resistência à compressão é uma função altamente não linear da idade e ingredientes. Neste problema pretende-se criar um modelo preditivo para a compressão à resistência do betão em função de variáveis como a quantidade de cimento, escória de alto forno, cinzas volantes, entre outras. Os dados podem ser obtidos aqui e uma das publicações originais é

I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks,"Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998)

Song popularity Data - Regressão nível N2

O ser humano tem uma forte associação com canções e músicas, sendo sugerido que a música pode beneficiar a saúde física e mental de várias maneiras, nomeadamente melhorando o humor ou diminuindo a dor e a ansiedade.
Vários estudos têm sido realizados para compreender as músicas e a sua popularidade. Neste problema pretende-se criar um modelo preditivo da popularidade de músicas com base em determinadas variáveis como a energia, a acústica, a instrumentalidade, a vivacidade, entre outros.
Os dados podem ser obtidos aqui.

Rice Dataset - Classificação nível N1

O arroz é uma das culturas de cereais mais produzidas e consumidas no mundo, passando por algumas etapas de fabricação como a limpeza, a triagem por cor e a classificação. Neste problema pretende-se criar um classificador que discrimine entre duas espécies de arroz (*Commeo* e *Osmancik*) a partir de um conjunto de 7 características morfológicas obtidas para cada grão de arroz.

Os dados podem ser obtidos aqui e uma das publicações originais é

Cinar, I. and Koklu, M. (2019). Classification of Rice Varieties Using Artificial Intelligence Methods. International Journal of Intelligent Systems and Applications in Engineering, vol.7, no.3 (Sep. 2019), pp.188-194. https://doi.org/10.18201/ijisae.2019355381.

Anuran Dataset - Classificação nível N2

Os anuros são uma ordem de animais da classe *Amphibia* que inclui rãs e sapos, apresentando diferentes especificidades no seu coaxar. Este conjunto de dados foi criado segmentando 60 registos de áudio, recolhidos *in situ* em condições reais de ruído, pertencentes a 4 famílias diferentes, 8 géneros e 10 espécies. Cada áudio corresponde a um espécime (um sapo individual) do qual são extraídas várias sílabas. Neste problema pretende-se criar um classificador que discrimine as famílias, os géneros ou as espécies de anuros com base nas características acústicas extraídas dos registos áudio.
Os dados podem ser obtidos aqui e uma das publicações originais é

COLONNA, J. G.; CRISTO, M.; NAKAMURA, E. F. (2014, August). A Distributed Approach for Classifying Anuran Species Based on Their Calls. In Pattern Recognition (ICPR), 2014 22nd International Conference on (pp. 1242-1247). IEEE.

2 Métodos, relatório e prazos

Cada grupo de trabalho deverá escolher apenas um de entre os *datasets* acima propostos, que estão categorizados por grau de dificuldade: **N1** para mais fácil ou **N2** para mais difícil. Devem também escolher pelo menos 3 modelos, de entre os listados na Tabela 1, para aplicar ao *dataset* escolhido.

Tabela 1: Categorização de modelos

Categoria	Modelo
A	Boosting
	SVM
	Redes Neuronais
B	Random Forest
	Bagging
	Árvore de decisão
C	k-NN
	Ridge
	LASSO
D	Regressão clássica
	Regressão Logística
	Naive Bayes

O grau de dificuldade do par (Categoria, Dataset) é refletido numa diferenciação da nota máxima possível de obter no trabalho, de acordo com a Tabela 2.

Tabela 2: Notas máximas obtidas para cada combinação de Categoria (A,B,C,D) e Dataset (N1,N2)

	Categoria	A	B	C	D
Dataset	N1	18	16	14	11
	N2	20	18	16	13

Cada elemento do grupo deverá ficar responsável pela aplicação computacional de (pelo menos) um modelo, sendo objeto de avaliação específica aquando da apresentação. Os algoritmos deverão ser comparados adequadamente quanto à sua performance. Deverão ainda utilizar uma metodologia adequada de seleção de modelos e escolha de parâmetros como por exemplo a validação cruzada, sempre que aplicável.

O relatório será apresentado sob a forma de slides, com uma estrutura a especificar posteriormente. O código produzido deverá ser submetido em conjunto com o relatório e deverá estar perfeitamente funcional.

O trabalho é realizado em grupos de 3 alunos. O prazo para entrega é 31 de maio, na plataforma de *e-learning*.

3 Critérios de avaliação

A avaliação do trabalho será realizada seguindo o conjunto de critérios que a seguir se descrevem. Sugere-se, por isso, que a estrutura da apresentação reflita e facilite a avaliação desses critérios.

1. Apresentação digital (75%)
   (a) Contextualização do problema
   (b) Análise inicial dos dados e pré-processamento
   (c) Procedimento experimental (inclui treino dos modelos e afinação de hiperparâmetros)
   (d) Apresentação e comparação de resultados
   (e) Discussão e conclusões
   (f) Grau de empenho na obtenção das melhores soluções
2. Apresentação Oral e Discussão (20%)
3. Código (R ou Python) bem estruturado e comentado (5%)

A apresentação digital pode seguir as normas estabelecidas na UA e, em particular, para o DMat. Podem encontrar no elearning um ficheiro ZIP com os ficheiros para um template de apresentação.