hugomcp@di.ubi.pt, 2020/21

Trabalho Prático 1
OCR Data Set

1. Considere o conjunto de dados disponível na página da disciplina (https://www.di.ubi.pt/~hugomcp/visaoComp/OCRDataSet.zip). Contém uma matriz com 36,000 linhas e 50 colunas, a representar 36 caracteres diferentes ('a'→'z' e '0'→'9'), num total de 20 instâncias por caracter.

Cada instância de cada caracter está representada por uma matriz de 50 linhas x 50 colunas do conjunto original. Em todas as imagens, o fundo (background) está assinalado a branco e a informação de cada caracter aparece a preto.

Exemplos:

- A informação desde a linha 1 até à linha 50 representa o primeiro elemento da classe 'a':

- A informação desde a linha 35951 até à linha 36000 representa a última ocorrência da classe '9':

O primeiro trabalho prático da disciplina consiste em implementar um script (conjunto de scripts) em linguagem *Python*, capazes de:

1. Dividir o conjunto de dados em 2 sub-conjuntos disjuntos, para "Aprendizagem" e "Teste", com base num valor-proporção (para treino) desejado. Por exemplo, ao especificar o valor "0.6" estamos a determinar que o conjunto de aprendizagem fique com 60% dos dados, enquanto o conjunto de teste ficará com os restantes 40%. Note que em cada sub-conjunto, o número de elementos de cada classe deverá ser igual.
2. Uma vez que a reduzida quantidade de dados disponíveis inviabiliza a utilização de técnicas baseadas em Aprendizagem Profunda (*Deep Learning*), será necessário

extrair um conjunto de características (à sua escolha) para cada instância, capazes de discriminar entre os elementos de cada classe. Neste caso a tarefa consiste em criar uma estrutura bidimensional (em forma de tabela), onde cada linha contém as características de cada instância, e o número de colunas será igual ao total de características extraídas.

0.25	0.56	24	56	7.20	2223	727	1.27	2.75	3.87
-0.4	0.88	11	0.01	0.18	28	988	1.212	1.27	9.99

3. Analise/compare a eficácia de diferentes classificadores disponíveis em *Python* (por exemplo *KNN*, árvores de decisão, regressão linear, redes neuronais, *SVMs*...), na discriminação dos caracteres, utilizando as características extraídas na alínea anterior.

   Analise de que forma os resultados variam em função de:

   a) Diferentes formas de normalização dos dados;

   b) Tipo de classificador utilizado;

   c) Proporção de dados utilizados para "Aprendizagem" e "Teste";

   d) Número e tipo de características extraídas.

   Para cada experiência, obtenha/reporte (pelo menos) as seguintes medidas de erro:

   - Matriz de confusão;

   - Precisão (*Accuracy*) global;

   - Score F-1 (por cada classe):