

DM-ISEL
MESTRADO EM MATEMÁTICA APLICADA PARA A INDÚSTRIA
Elementos de Aprendizagem Estatística

Docentes: Aleixo, S.; Geraldês, C.; Pinto, I.

Projeto

Neste projeto pretende-se implementar um software básico de *Optical Character Recognition* (OCR).

1. Faça *download* do ficheiro de trabalho a partir do *site*:

<https://www.kaggle.com/datasets/crawford/emnist>.

A seguir, abra os arquivos "*emnist – balanced – train.csv*" e "*emnist – balanced – test.csv*" com um programa de edição de texto e substitua todos os espaços por ", ". Confirme se o separador decimal é ".".

O arquivo contém imagens de caracteres alfanuméricos manuscritos, incluindo letras maiúsculas (Classificações de 10 a 35) e minúsculas (Classificações de 36 a 61), além de dígitos de 0 a 9 (Classificações de 0 a 9). Cada linha do arquivo corresponde a uma imagem única. Na primeira coluna, é descrito o carácter ao qual determinada linha se refere (variável resposta). As colunas restantes correspondem aos píxeis da imagem, onde cada imagem tem uma resolução de $28 \times 28 = 784$ píxeis. O primeiro píxel no canto superior esquerdo da imagem corresponde à segunda coluna do arquivo, e assim por diante (o píxel 28 corresponde à coluna 29 e o primeiro píxel da segunda linha da imagem corresponderá à coluna 30).

Para visualizar um carácter específico, pode-se utilizar a seguinte função, levando em consideração o objeto que corresponde ao arquivo já carregado e a linha do arquivo onde se encontra um exemplo do carácter desejado.

```
> library(plot.matrix)
> plotfigure <- function(row,dataset){
+   X = NULL
+   if(!is.null(nrow(dataset))){
+
+     X = data.frame(matrix(dataset[row,2:785],nrow=28))
```

```

+
+   }else{
+
+       X = data.frame(matrix(dataset[2:785],nrow=28))
+   }
+   m1 = data.matrix(X)
+   plot(m1, cex=0.5)
+ }
>

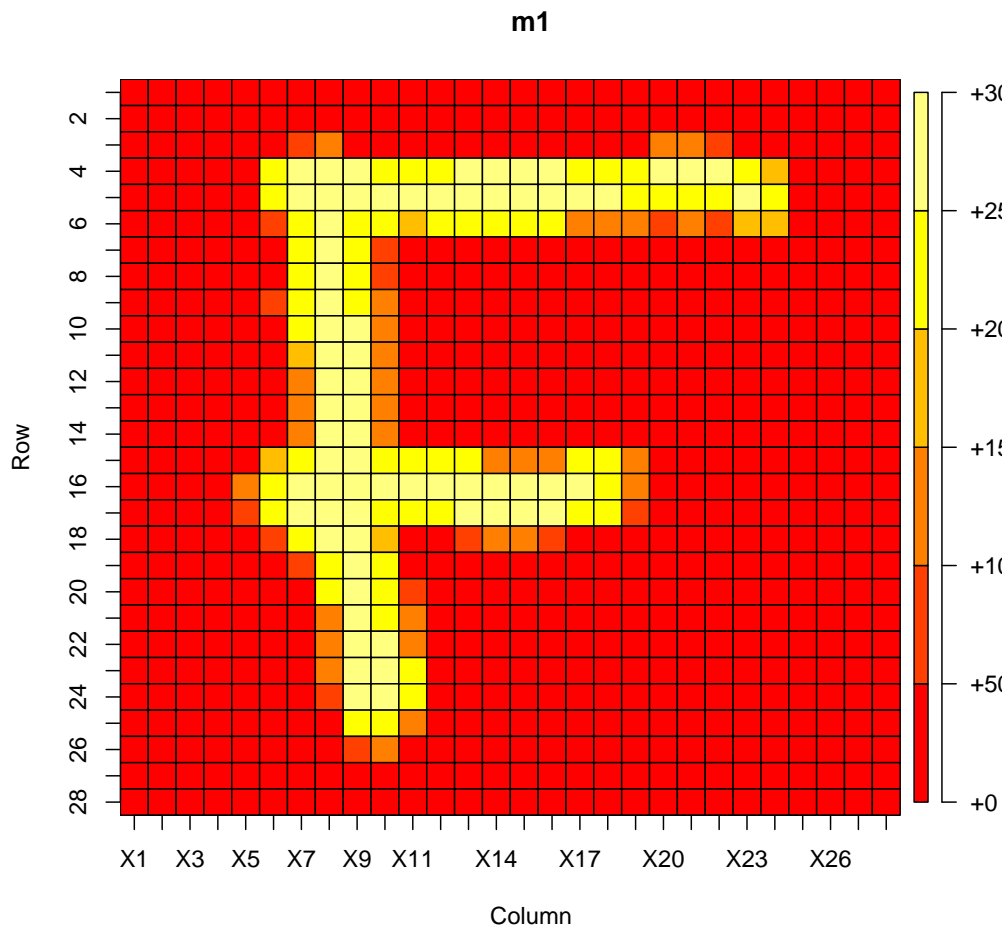
```

A seguir pode ver-se um exemplo da imagem do caracter "F" que se encontra na linha 13 do ficheiro e corresponde ao código 15 na primeira coluna.

```

> #Read the zip codes dataset.
> #Each line corresponds to a handwritten figure.
> #The first column shows the corresponding symbol.
> #The next 256 (16x16) columns correspond to the orange color of each píxel in the figure
>
> dataset <- read.csv("bd2/emnist-balanced-train.csv",sep=" ",header = FALSE)
> #Function that plots the figure
> #corresponding to a specific dataset row
>
> plotfigure(13,dataset)
>

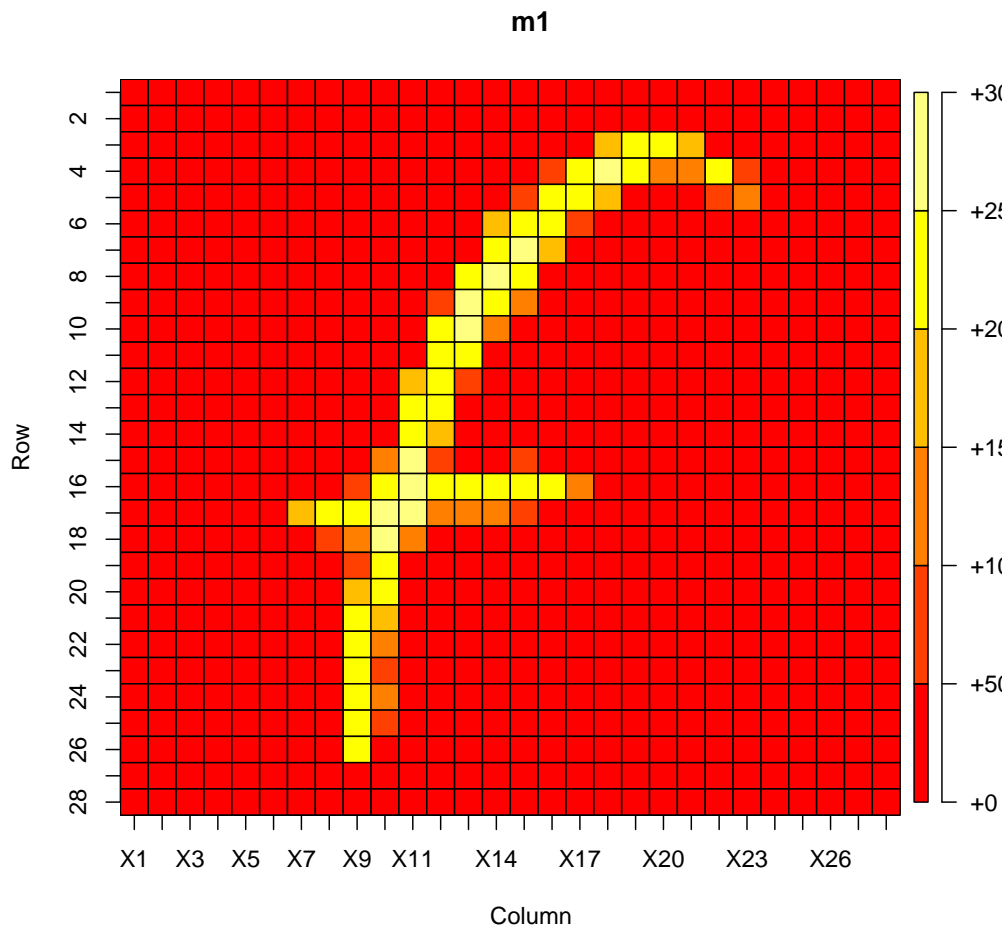
```



Como se pode verificar, ao aplicar a função *"plotfigure"*, é devolvida uma imagem cujas cores variam em tons de laranja, sendo que o amarelo corresponde ao valor 1 e o vermelho a -1 .

De seguida apresenta-se um outro exemplo do mesmo dígito manuscrito, que está na linha 12 do ficheiro.

```
> plotfigure(4,dataset)
```



Os dois caracteres atribuídos a cada grupo de trabalho são os seguintes:

- Grupo I - A e C
- Grupo II - D e F
- Grupo III - O e Z

Dos dois dígitos que foram atribuídos ao seu grupo, escolha um deles e redefina as categorias da variável resposta de forma a que esta seja binária. O objetivo posterior será o de discriminar o dígito eleito dos demais (onde se encontra também o dígito não eleito). Responda então às seguintes questões:

1. Aplique uma árvore de classificação/decisão e efetue todos os procedimentos de forma a obter o modelo mais parcimonioso. Faça um estudo de estatística descritiva a partir de dois píxeis: um correspondente ao nó principal e outro a um dos nós internos à escolha (utilize, entre outros métodos, uma caixa de bigodes para comparar as duas distribuições). Interprete os resultados.
2. Determine os nós mais importantes (até a um máximo de dez) de modo a obter os píxeis mais relevantes que permitam classificar corretamente, no mínimo 80% das imagens do dígito eleito na amostra de treino.
3. Faça uma análise exploratória sobre os píxeis que achar mais relevantes (caixas de bigodes, correlação entre píxeis, distâncias)
4. Aplique o método DBSCAN para classificar o carácter eleito em relação ao outro que sobejou.
5. Faça o mesmo exercício da alínea anterior, só que utilizando o Classificador Naive Bayes.
6. Através da aplicação de uma Análise em Componentes Principais (PCA), encontre e compare os píxeis mais importantes de cada um dos dois caracteres que lhe foi atribuído.
7. Proceda à redução da dimensionalidade do seu *dataset* e a seguir utilize um dos métodos abordado nas aulas, à sua escolha, para classificação. Verifique se obtém algum ganho de desempenho na classificação com ou sem PCA.
8. Aproveite a aplicação do método PCA e utilizando as duas componentes mais importantes proceda a uma análise de clusters hierárquica e uma outra utilizando o método *K – Means*. Apresente o respetivo dendograma, o gráfico da silhueta e respetivo *score*. No final, considere dois clusters e compare os resultados obtidos com a classificação observada.

Deverá ser elaborado um relatório no formato de pdf, com a designação do grupo, o número e o nome de aluno dos respetivos elementos. Estruture o relatório de forma a que seja possível identificar as respostas para cada uma das perguntas pedidas.