1 Sobre os trabalhos práticos

1.1 Objetivo

O objetivo do trabalho prático é de realizar um trabalho de casa sobre a análise e implementação de um tema abordado nas aulas com uma parte de programação e de experiência numérica.

- Serão criados grupos de 3-5 alunos;
- Devem produzir um relatorio de aproximadamente 10 páginas e um conjunto de slides para apresentar publicamente o vosso trabalho.
- A classificação deste trabalho será: 50% do relatório, 50% da apresentação oral (marcada para a segunda semana de janeiro).
- Cada grupo deve escolher uma proposta na lista apresentada adiante. Um tema não pode ser escolhido mais do que duas vezes.

1.2 Procedimentos

1. Formação dos grupos, onde cada grupo deve designar um representante .
2. Dia 24/11/2023, a partir de 19h00, cada representante de grupo envia um email para `rmp@math.uminho.pt` com a lista ordenada de preferências do trabalho a realizar (EX: T2 T1 T4 T3 - que significa que primeira preferência é T2, depois T1, depois T4 e depois T3). Vou atribuir os projetos consoante a hora de chegada dos emails (os emails que chegarem antes das 19:00 não serão considerados). Neste e-mail deverá vir nomeada a constituição do grupo, assim como o seu representante.
3. Se um tema já foi escolhido duas vezes, o grupo deve escolher uma outra proposta. O representante do grupo receberá um e-mail a avisar do sucedido.
4. Na quarta feira dia 29/11 será publicada lista de grupos/trabalhos na BB. O representante de cada grupo deverá contactar o tutor.
5. Entregar por email ao tutor respetivo o relatório e os slides em pdf (atá terça feira 2/01/2024 às 21h).
6. As apresentações vão decorrer na aula de 4/01/2024.

Os temas, detalhados no resto do documento, são o seguintes.
- *T1 Algoritmo de tipo Lloyd – figuras* 5×5
- *T2 Algoritmo de tipo Lloyd – Dados reais* $\mathbb{R}^3$
- *T3 Aplicação do PCA em reconhecimento de dígitos*
- *T4 Eigenfaces: reconhecimento de faces.*

2 Algoritmo de LLoyd

2.1 Algoritmo tipo Lloyd – figuras 5×5 - (Tutor: Rui Pereira)

Pretende-se que os alunos implementem um algoritmo do tipo Lloyd para um conjunto de eventos que são figuras com 25 pixels binários a importar do ficheiro de texto **BD1.txt** em anexo. Considere K=3 no algoritmo de Lloyd.

b21	b22	b23	b24	b25
b16	b17	b18	b19	b20
b11	b12	b13	b14	b15
b6	b7	b8	b9	b10
b1	b2	b3	b4	b5

A métrica de dissemelhança a usar é para dois eventos $x, y \in \{0,1\}^{25}$:

$$d(x,y) = \frac{1}{25}\sum_{i=1}^{25}(d_i(x,y))$$

onde,

$$d_i(x,y) = \begin{cases} 1, & \text{se} \quad x_i \neq y_i, \\ 0, & \text{se} \quad x_i = y_i. \end{cases}$$

O representante de cada cluster a usar (do tipo medóide) deverá escolher o elemento do cluster que minimiza a função custo, de **m** ser representante de **C**,

$$E(m;C) = \sum_{x \in C} d(m,x)$$

.
Ou seja deverá escrever função que dados elementos dum cluster calcule o custo de cada um desses elementos serem o representante do cluster, e escolher para representante aquele que faça o custo ser menor. Analisar os resultados dos clusters obtidos.

Poderão estudar outras métricas.

De seguida pretende-se que os alunos implementem um classificador, por forma a que, caso surjam novos eventos depois obter a partição do domínio usando o algoritmo de LLoyd, seja possível saber a que classe (cluster) deveria pertencer sem voltar a usar o algoritmo de clusterização.
São fornecidos dois conjuntos de dados nos ficheiros **teste.txt** e **teste2.txt** que agora em cada linha tem o valor dos 25 pixels da figura e o vigésimo-sexto tem o valor da sua classe.

Pretende-se que os alunos analisem a performance do classificador, tendo como base uma tabela de confusão.

Exemplos de fontes bibliográficas :

[1] Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert, *A Survey of Binary Similarity and Distance Measures*, systemics, cybernetics and informatics, vol. 8(1) (2010)
[2] O. Simeoni, *A Brief Introduction to Machine Learning for Engineers*,
Online at `https://arxiv.org/pdf/1709.02840.pdf`, 2018.
[3] Christopher De Sa, Notes of Cornell University on Introduction to Machine Learning *Lecture 4*, Online at
`https://www.cs.cornell.edu/courses/cs4780/2022sp/corereferences`, 2022.

2.2 Algoritmo tipo Lloyd – Dados reais $\mathbb{R}^3$ (Tutor: Rui Pereira)

Pretende-se que os alunos implementem o algoritmo do tipo Lloyd para um conjunto de dados 3D a importar do ficheiro de texto **BD.txt** em anexo.

A métrica a usar é a Euclidiana (poderão também experimentar outras métricas e comentar os resultados). Será útil representar os dados graficamente para ter uma noção empírica do nº de clusters a considerar. Sabe-se que os 20 primeiros eventos são de uma classe, os 20 seguintes de uma segunda classe, e os terceiros 20 duma terceira classe. Analise os resultados obtidos na clusterização obtida. Comente os resultados.

Para estimar o valor de K (número de clusters), e neste caso confirmar que é 3, é sugerido que use o Elbow method.

É dada uma segunda base de dados **BD2.txt**. Aplique novamente o programa para a nova base de dados. Sabe-se que os 20 primeiros eventos são de uma classe, os 20 seguintes de uma segunda classe, e os terceiros 20 duma terceira classe. Analise os resultados obtidos na clusterização obtida. Comente os resultados.

De seguida pretende-se que os alunos implementem um classificador, por forma a caso surjam novos eventos do mesmo tipo, seja possível saber a que classe (cluster) deveria pertencer. São fornecidos dois novos ficheiros de dados **teste.txt** e **teste2.txt** com o mesmo formato de **BD.txt**, mas, com uma coluna extra que tem a classificação do evento. Pretende-se que os alunos analisem a performance do classificador (por exemplo usando tabelas de confusão).

NOTA: o ficheiro **teste.txt** deve ser usado com **BD1.txt** e o ficheiro **teste2.txt** deverá ser usado com **BD2.txt**.
Outras bases de dados poderão ser usadas para testar o programa desenvolvido.

Exemplos de fontes bibliográficas :

[1] O. Simeoni, *A Brief Introduction to Machine Learning for Engineers*,
Online at `https://arxiv.org/pdf/1709.02840.pdf`, 2018.
[2] Christopher De Sa, Notes of Cornell University on Introduction to Machine Learning *Lecture 4*, Online at
`https://www.cs.cornell.edu/courses/cs4780/2022sp/corereferences`, 2022.

3 Análise de Componentes Principais

3.1 Aplicação do PCA em reconhecimento de dígitos (Tutor: Pedro Patrício)

O *dataset* MNIST contém os dígitos $0, \ldots, 9$ manuscritos. Use o PCA por forma a que se possam reconhecer dígitos obtidos. Divida a bases de dados em treino/teste, ou então assuma que a escrita foi obtida de forma externa, como, por exemplo, através de um dígito que um utilizador escreveu. Neste caso, tenha particular cuidado na dimensão da foto e na cor do fundo.

1. Defina as componentes principais que considera relevantes.
2. Implemente a distância euclidiana e a distância de Mahalanobis para usar no espaço das projecções.
3. Teste o seu modelo para as escolhas que efectuou.

[1] Yann LeCun, Corinna Cortes, Christopher J.C. Burges, "THE MNIST DATABASE of handwritten digits" , `http://yann.lecun.com/exdb/mnist/` Acedido em 8 dezembro 2020.

3.2 Eigenfaces: reconhecimento de faces (Tutor: Pedro Patrício)

O PCA pode ser usado como método de compressão de imagens e aplicado ao reconhecimento de faces, tal como apresentado em [1].
Usando como base o que foi leccionado nas aulas,

1. Obtenha um conjunto de faces de uma base de dados (como [2]; cf. com [3] e [4], caso opte por outra base de dados; pode, em alternativa, criar uma base de dados com os elementos do grupo, ou da forma que achar conveniente – tenha o cuidado com o fundo das fotografias e centrar as faces).

2. No caso de usar uma base de dados de referência, separe em elementos de treino e de teste.

3. Defina as componentes principais que considera relevantes.

4. Implemente a distância euclidiana e a distância de Mahalanobis para usar no espaço das projecções.

5. Teste o seu modelo para as escolhas que efectuou.

[1] M. Turk, A. Pentland, Eigenfaces for Recognition, Journal of Cognitive Neurosicence, Vol. 3, No. 1, 1991, pp. 71-86, `http://www.face-rec.org/algorithms/PCA/jcn.pdf`
[2] *The Yale Face Database*, `http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html` normalizadas em `http://vismod.media.mit.edu/vismod/classes/mas622-00/datasets/`
[3] Ralph Gross, Face Databases, in S.Li and A.Jain, (ed). *Handbook of Face Recognition.* Springer-Verlag, 2005,
`http://ri.cmu.edu/pub_files/pub4/gross_ralph_2005_1/gross_ralph_2005_1.pdf`
[4] `http://www.face-rec.org/databases/`