1 Sobre os microprojetos

1.1 Objetivo

O objetivo do microprojeto 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 5-6 alunos;
- Devem produzir um relatorio de 6 páginas e um conjunto de slides para apresentar publicamente o vosso trabalho.
- A classificação deste trabalho será: 50% do guião, 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. Até dia 4/12, formação dos grupos, onde cada grupo deve designar um representante que me enviará um email com a lista dos membros do grupo.
2. Dia 5/12, a partir de 14h00, cada representante de grupo envia-me um email com o nome do projeto que gostaria realizar. Vou atribuir os projetos consoante a hora de chegada dos emails (excluindo os emails que chegarem mais cedo).
3. Se um tema já foi escolhido duas vezes, o grupo deve escolher uma outra proposta.

Os temas, detalhados no resto do documento, são o seguintes.
- *T1 Dissemelhanças com dados binários*
- *T2 Medidas de semelhanças para dados nominativos/categóricos*
- *T3 Clustering hierárquico simples*
- *T4 K-mode e K-medoid clustering*
- *T5 Impuridades e função de ganho para as árvores de decisão*
- *T6 Método de pré-poda para otimizar a construção de uma árvore de decisão*
- *T7 Aplicação do PCA em gestão de portefólios financeiros*
- *T8 Eigenfaces: reconhecimento de faces.*

2 Semelhanças e Dissemelhanças

2.1 Dissemelhanças com dados binarios

Propõe-se neste trabalho um estudo das diferentes dissemelhanças usadas com dados binários, em particular nas áreas onde se trata de classificações de perfil na base de respostas. Como aplicação, usaremos a database SCADI "self-care problems classification based on ICF-CY" sobre a classificação de crianças com deficiências motoras.
Trabalhos a realizar

1. levantamento das diferentes semelhanças e dissemelhanças com dados binarios;
2. propriedades e classificação;
3. implementação das métricas mais relevantes;
4. aplicação com a base de dados SCADI

Os documentos de referências para a construções de métricas binárias são:

[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] Brian Stacey, *A Standardized Treatment of Binary Similarity Measures with an Introduction to k-Vector Percentage Normalized Similarity*, Online at `https://mpra`, MPRA Paper No. 72882.

2.2 Medidas de semelhanças para dados nominativos/categóricos

Em áreas tais como ambiente, história, os dados têm unicamente um carácter categorial logo a noção de semelhança deve tomar em conta o contexto de utilização destes dados. Propomos um trabalho de estudo deste tipo de métrica com aplicação a identificação de cogumelos venenosos (ficheiro fornecido).
Trabalhos a realizar

1. levantamento das diferentes semelhanças com dados nominativos
2. implementação das métricas mais relevantes;
3. aplicação a deteção de congumelos venenosos

Os documentos de referências para o estudo de métricas categoriais são:
[1] Tiago R.L. dos Santos, Luis E. Zárate, *Categorical data clustering: What similarity measure to recommend?*,Expert Systems with Applications 42 (2015) 1247–1260
[2] Shyam Boriah, Varun Chandola, Vipin Kumar, *Similarity Measures for Categorical Data: A Comparative Evaluation*, Conference: Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA (April 2008) DOI: 10.1137/1.9781611972788.22

3 Clustering

3.1 Clustering hierárquico simples

O clustering hierárquico tem a vantagem de não definir "a priori" o número de clusters que podemos deduzir depois do análise da curva associada aos níveis de fusão entre clusters. Propõe-se de estudar e implementar um método de clustering hierárquico usando diferentes técnicas de linkage. Entra também neste estudo a questão da definição do número de clusters. Usamos a base de dados `cars.csv` para as experiências numéricas.

1. descrever o método de clustering hierárquico aglomerativo
2. implementá-lo com diferentes tipos de linkage
3. determinar e implementar uma técnica de identificação do número de clusters
4. aplicação a base de dados `cars.csv`

Os documentos de referências para o estudo de clustering hierárquicos são:
[1] Daniel Müllner, *Modern hierarchical, agglomerative clustering algorithms*, arXiv:1109.2378 (2011)
[2] F. Murtagh, *A Survey of Recent Advances in Hierarchical Clustering Algorithms*, the computer journal, vol. 26, no. 4 (1983)

3.2 K-mode e K-medoid clustering

O método K-mean não pode utilizar dados binários ou de categoria. A extensão natural é de utilizar outros tipos de represententes, nomeadamente o medoid ou o mode. Pretendemos implementar estas duas técnicas usando diversas métricas associadas ao tipo de dados. Usamos a base de dados Dishonest Internet users Dataset.

1. descrição dos métodos K-medoid e K-mode
2. implementação (definir a métricas)
3. aplicação aos dados Dishonest Internet users Dataset

Os documentos de referências para K-medoid e K-mode são:
[1] Archna Kumari, Pramod S. Nair, Sheetal Kumrawat, *An Enhanced K-Medoid Clustering Algorithm*, International Journal on Recent and Innovation Trends in Computing and Communication Volume: 4 Issue: 6
[2] Zhexue Huang, *A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining*, Research Issues on Data Mining and Knowledge Discovery (1997)

4 Árvore de decisão

4.1 Impuridades e função de ganho para as árvores de decisão

A função de ganho (calculada com a impuridade) é a chave na construção da árvore de decisão sendo que determina a função para otimizar localmente com respeito aos atributos. A combinação de vários ingredientes (impuridades, número de valores dos atributos, número de dados em consideração) permite de melhorar a qualidade desta função. Propõe-se estudar um conjunto de novos ganhos e avaliar a sua eficácia em decisão rápida e robústa.

1. levantamento de diferentes funções de ganho que se encontram na literatura
2. implementação e construção da árvore de decisão num código de tipo Hunt
3. avaliação da eficácia
4. aplicação com a base de dados Wine Quality.

Os documentos de referências para as função de ganhos são [1] Fernando Berzal, Juan-Carlos Cubero, Fernando Cuenca, María J. Martín-Bautista, *On the quest for easy-to-understand splitting rules*, Data & Knowledge Engineering 44 (2003) 31–48
[2] Wei Liu, Sanjay Chawla, David A. Cieslak, Nitesh V. Chawla, *A Robust Decision Tree Algorithm for Imbalanced Data Sets*, Proceedings of the 2010 SIAM International Conference on Data Mining, 766-777

4.2 Método de pré-poda para otimizar a construção de uma árvore de decisão

A construção de uma árvore de decisão pode conduzir a uma ramificação excessiva reduzindo a sua eficácia e aumentando o risco de "overfitting". As técnicas de pré-poda ("pre-pruning") servem para cortar previamente os ramos que não são relevantes e assim reduzir o peso global da arvore.

1. levantamento das técnicas de pre-pruning
2. implementação num código de tipo Hunt
3. avaliação de eficácia e do over-fiting
4. utilização da base de dados Pima Indians Diabetes Dataset

Os documentos de referências para o pré-poda são:
[1] Lior Rokach and Oded Maimon, *Top-Down Induction of Decision Trees Classifiers -A Survey*, ieee transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 35, no. 4, november 2005
[2] Johannes Furnkranz, Gerhard Widmer, *Incremental Reduced Error Pruning*, Machine Learning Proceedings 1994 Proceedings of the Eleventh International Conference, Rutgers University, New Brunswick, NJ, July 10–13, Pages 70-77

5 Análise de Componentes Principais

5.1 Aplicação do PCA em gestão de portefólios financeiros

Na área das finanças, o *Machine Learning* tem vindo a assumir-se como ferramenta para a tomada de decisões, e que tem sido explorado pelas denominadas *FinTech*.
Em [1], foi utilizado o PCA por forma a optimizar a carteira de títulos por forma a obter os melhores retornos financeiros. O PCA foi aplicado em 2 grupos distintos de títulos da American Index DJI, Dow Jones Industrial Average, utilizando como base de dados a *Yahoo Finance*, [2], e a junção dos dois grupos.
Pretende-se que

1. Obtenha os dados referenciados em [1], usando [2], e que constam no índice *DJI*.
2. Aplique o PCA.
3. Compare o que obteve com os resultados apresentados em [1].

[1] Giorgia Pasini (2017). Principal Component Analysis for Stock Portfolio Management. *International Journal of Pure and Applied Mathematics*. Volume 115 No. 1 2017, 153-167. `https://ijpam.eu/contents/2017-115-1/12/12.pdf`
[2] *Yahoo Finance*, `https://finance.yahoo.com/`

5.2 Eigenfaces: reconhecimento de faces

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).
2. 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/