Reunião 26/07/2024

Tema: modelos de aprendizado de máquina não supervisionados (foco C-Means). Inclui métricas para avaliação de desempenho e visualização de regras de agrupamento. Comparação com resultados obtidos em trabalhos anteriores.

Foco da semana: elaborar um tutorial com um panorama geral sobre modelos não supervisionados para agrupamento, com foco no C-Means. Estudar também o K-Means e o Mean-Shift (ambos disponíveis na ScikitKLearn). Explorar métricas de desempenho. Ler o artigo (no drive) que utiliza árvores de decisão para explicação das regras de agrupamento e explorar outras ferramentas, inclusive gráficas, para apresentação das regras aplicadas pelos modelos no processo de agrupamento. Lembrar de já ir documentando as referências dos artigos e outros materiais utilizados na pesquisa.

Algumas referências:

[C-Means Clustering Explained | Built In](https://builtin.com/data-science/c-means#:~:text=What%20Is%20C%2DMeans%20Clustering,for%20being%20in%20that%20cluster.)

[Understanding Fuzzy C Means Clustering (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2024/05/understanding-fuzzy-c-means-clustering/)

[Fuzzy C-Means Clustering (FCM) Algorithm | by Aman Gupta | Geek Culture | Medium](https://medium.com/geekculture/fuzzy-c-means-clustering-fcm-algorithm-in-machine-learning-c2e51e586fff)

===================================

Reunião 10/08/2024

Tarefas:

1. Revisar o recorte do projeto de tese sobre a metodologia aplicada, os parâmetros utilizados nos modelos de agrupamento e as métricas para avaliação dos modelos. Observe que eu usei essas métricas pq estamos tratando de modelos não supervisionados - nos seus testes com as bases Iris e Titanic, você deve usar métricas para modelos supervisionados. Depois que decidirmos sobre quais modelos vamos aplicar nos dados de saúde é que vamos aplicar as métricas que eu usei;
2. Olhar o notebook que eu gerei (está no drive) e repetir para os modelos K-Means, Mean Shift, C-means e DBSCAN (Não vamos usar por agora o GMM) comparando os resultados sobre as bases de dados Iris e Titanic;
3. Montar uma apresentação explicando cada modelo, aplicações, vantagens e desvantagens, hiperparâmetros com possíveis valores, valores utilizados no exemplo com justificativas e resultados alcançados;

Tarefa extra (se der tempo): aplicar os modelos sobre o .csv de saúde e analisar os resultados alcançados considerando as métricas de modelos não supervisionados. Explicar os grupos gerados.

Reunião 23/08/2024

Tarefas: testar os algoritmos de agrupamento sobre os arquivos .csv (ureia.csv e creatinina.csv). Passos sugeridos:

Passo1: completar cada .csv com os dias faltantes no período de 1 até 7 (coluna day) e eliminar as linhas com day > 7.

Passo2: juntar em um único .csv considerando a igualdade das colunas subject\_id e day.

|  |  | Perfil Hepático | |
| --- | --- | --- | --- |
|  |  | Ureia.csv | Creatinina.csv |
| subject\_id | Dia | UreiaMax | CreatininaMax |
| P1 | 1 | 30 | 45 |
| P1 | 2 | 40 | NULL |
| P1 | 3 | NULL | 60 |
| P1 | 4 | NULL | NULL |
| P1 | 5 | NULL | NULL |
| P1 | 6 | 50 | 70 |
| P1 | 7 | 30 | 20 |

Passo3: transpor as linhas para colunas com o objetivo de ter 1 paciente por linha.

| subject\_id | U1 | U2 | U3 | U4 | U5 | U6 | U7 | C1 | C2 | C3 | C4 | C5 | C6 | C7 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| P1 | 30 | 40 | NULL | NULL | NULL | 50 | 30 | 45 | NULL | 60 | NULL | NULL | 70 | 20 |

=======

Reunião 06set

Tarefas - preparar Relatório parcial:

1. Levantar trabalhos atuais de algoritmos de agrupamento aplicados ETL, tratamento/transformação de séries de dados, redução de dimensionalidade em trabalhos na área de saúde;
2. RefTeórico - artigos/livro Modelos de ML não supervisionados para agrupamento de dados e métricas;
3. RefTeórico - artigos/livro Modelos utilizados com conceito, matemática por trás do modelo, vantagens e desvantagens, aplicação - justificar pelo tipo do dado que estamos trabalhando (estruturados, numéricos, contínuos, com falta de dados, séries temporais de observações) - justificar a escolha do modelo - dbscan, cmeans, kmeans, meanshift;
4. Materiais e métodos:
   1. MIMIC IV, Scklearn, colab, Python + bibliotecas utilizadas;
   2. versões e bibliotecas dos modelos implementados;
   3. Metodologia: dados extraídos + aplicação dos modelos + comparação com resultados dos trabalhos de matheus e leo.

====

Reunião 18set

Passo1: com o .csv de ureia e creatinina 7 dias ( e posteriormente de 15dias), fazer as seguintes simulações:

1. Substituir as células com 99999 ou NULL com o último valor registrado. Quando não houver registro no primeiro dia, eliminar as linhas. Em ambos os casos, guardar em um arquivo separado o subject\_id (do que teve dado imputado e do que foi excluído);
2. Gerar grupos por variável:
   1. MeanShift: num grupos e qualidade dos grupos
   2. KMeans: n=número de grupos sugerido pelo MeanShift
   3. CMeans: n=número de grupos sugerido pelo MeanShift
   4. DBScan: n=número de grupos sugerido pelo MeanShift
   5. KMeansSeries: n=número de grupos sugerido pelo MeanShift
3. Analisar os resultados considerando as métricas, o número de grupos e a qualidade dos grupos (gráfico pairplot)
4. A partir do dataset gerado pelo modelo com melhor desempenho, explicar os grupos utilizando árvores de decisão.
5. No final vc vai comparar os grupos que vc está gerando com os que eu gerei.

Fico devendo: csvs com 15 dias. referência de algoritmo de tendencia

============

Reunião 24/09

Passo 1: Criar um dataset com os dados de ureia e creatinina (pacientes que, após o tratamento de dados (limpeza e preenchimento) possuem séries de Ureia e Creatinina) e aplicar os modelos de agrupamento. Registrar resultados.

Passo 2: Explicar os grupos gerados utilizando árvore de decisão a partir do melhor modelo (escolhido a partir das métricas alcançadas e análise visual do pairplot)

Passo 3: Aplicar o algoritmo de tendência para os dados de Ureia e gerar um dataset saída com o subject\_id e a tendência (veja quais as informações que o MannKendall fornece - pelo que lembro ele te dá pelo menos 4 resultados). Faça o mesmo com Creatinina.

No final vamos comparar os resultados de um modelo de classificação (Random Forest) considerando a entrada original (com proporções), a entrada agrupada com cada variável (para isso vamos aplicar o melhor modelo em cada variável de exames), a entrada transformada com a tendência. Falamos melhor sobre essas comparações na próxima reunião.

============

Reunião 07/out

Para Ureia:

1. Separar o dataset filtrando apenas os pacientes com dados de ureia utilizados como entrada para os modelos de agrupamento (+- 3450) - subject\_id, ureia\_d1, urei\_d2…, ureia\_d15
2. Para os mesmos pacientes do item 1) gerar um dataset com o resultado do agrupamento utilizando o KMeans - subject\_id, grupo
3. Para os mesmos pacientes do item 1, aplicar o Mannkendall na série e gerar um dataset com subject\_id, trend, p, z, Tau, s, var\_S, slope, intercept => ~~aplicar o (MeanShift?), KMeans sobre esse dataset => gerar um dataset de saída com subject\_id, grupo~~