Instituto Infnet	**Avaliação**	Nota: Visto do Professor:

MIT em Inteligência Artificial, Machine Learning e Deep Learning	
Nome	Mateus Teixeira Ramos da Silva
Link do repositório	https://github.com/GitMateusTeixeira/ml_clustering/tree/main/infnet_clustering_pd
Módulo	Algoritmos de Inteligência Artificial para Clusterização
Prazo	20.11.2024

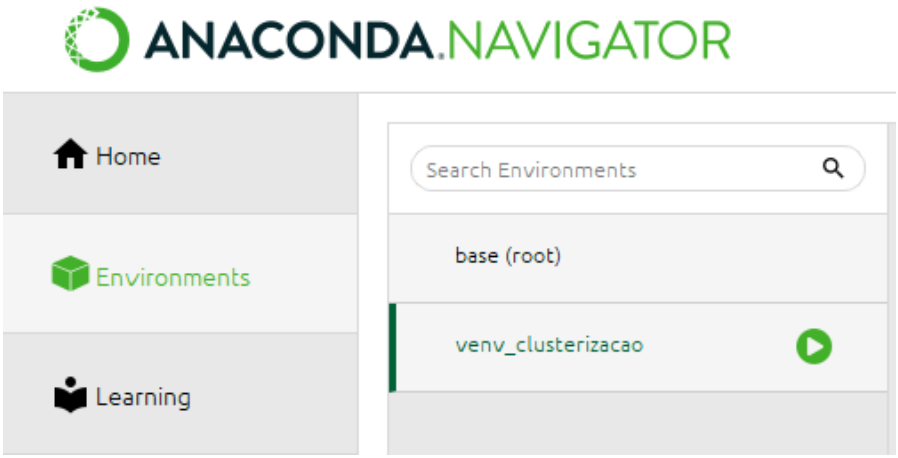
Parte 1. Infraestrutura

1. Você está rodando em Python 3.9+

R: Requerimento atendido no ponto “2.1. Versão do Python e Ambiente Virtual” do arquivo.ipynb.

2. Você está usando um ambiente virtual: Virtualenv ou Anaconda

R: Requerimento atendido no ponto “2.1. Versão do Python e Ambiente Virtual” do arquivo.ipynb.



3. Todas as bibliotecas usadas nesse exercício estão instaladas em um ambiente virtual específico

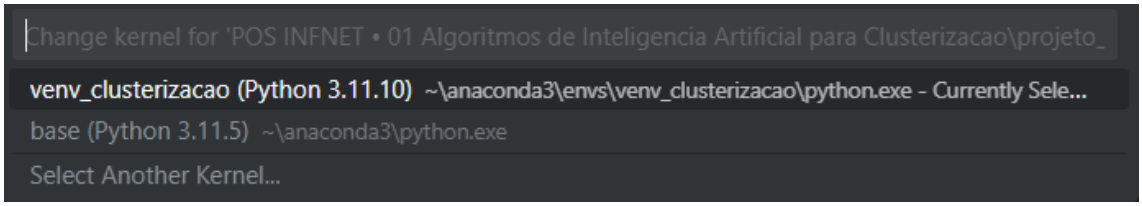
R: Requisito atendido no arquivo 'requirements.txt', presente no repositório (https://github.com/GitMateusTeixeira/ml_clustering/tree/main/infnet_clustering_pd).

4. Gere um arquivo de requerimentos (requirements.txt) com os pacotes necessários. É necessário se certificar que a versão do pacote está disponibilizada.

R: Requisito atendido no arquivo 'requirements.txt', presente no repositório (https://github.com/GitMateusTeixeira/ml_clustering/tree/main/infnet_clustering_pd).

5. Tire um printscreen do ambiente que será usado rodando em sua máquina.

R: Arquivo no repositório (https://github.com/GitMateusTeixeira/ml_clustering/tree/main/infnet_clustering_pd)



6. Disponibilize os códigos gerados, assim como os artefatos acessórios (requirements.txt) e instruções em um repositório GIT público. (se isso não for feito, o diretório com esses arquivos deverá ser enviado compactado no moodle).

R: Link do repositório no GitHub: https://github.com/GitMateusTeixeira/ml_clustering/tree/main/infnet_clustering_pd

Parte 2. Escolha de base de dados

1. Baixe os dados disponibilizados na plataforma Kaggle sobre dados sócio-econômicos e de saúde que determinam o índice de desenvolvimento de um país. Esses dados estão disponibilizados através do link: https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data.

R: Requerimento atendido no arquivo '.\data\country-data.csv' no repositório do GitHub e no ponto "1.2. Importar os dados" do arquivo pd_clusterizacao.ipynb.

2. Quantos países existem no dataset?

R: Respondido no ponto "3.2. Quantos países existem no dataset?" do arquivo pd_clusterizacao.ipynb.

3. Mostre através de gráficos a faixa dinâmica das variáveis que serão usadas nas tarefas de clusterização. Analise os resultados mostrados. O que deve ser feito com os dados antes da etapa de clusterização?

R: Respondido no ponto "3.3. Gráficos sobre a faixa dinâmica das variáveis" do arquivo pd_clusterizacao.ipynb.

Parte 3. Clusterização

2. Para os resultados do K-Médias:

a. Interprete cada um dos clusters obtidos citando:

i. Qual a distribuição das dimensões em cada grupo

R: Respondido no ponto "4.3. Análise gráfica do desenvolvimento econômico" do arquivo pd_clusterizacao.ipynb.

ii. O país, de acordo com o algoritmo, melhor representa o seu agrupamento. Justifique

R: Respondido no ponto "4.3.2. Países que mais representam a clusterização da análise econômica" do arquivo pd_clusterizacao.ipynb.

3. Para os resultados de Clusterização Hierárquica, apresente o dendrograma e interprete os resultados.

R: Respondido no ponto “6.4. Análise do Dendrograma” do arquivo pd_clusterizacao.ipynb.

4. Compare os dois resultados, aponte as semelhanças e diferenças e interprete.

R: Em ambas as análises, pode-se perceber que o dataset, por ser multidimensional (possuindo nove colunas numéricas) necessita da aplicação de técnica de redução de dimensionalidade para uma melhor leitura.

No entanto, mesmo isolando apenas duas colunas do dataset (a exemplo, a coluna de renda per capta e expectativa de vida), verificou-se que os dados se encontram muito próximos, denotando diferenças singelas nos três clusters.

Tanto a técnica do K-Means, quanto a Clusterização Hierárquica se propõem a agrupar dados de forma não supervisionada.

Uma das grandes diferenças entre os dois métodos é que ao passo em que o K-Means necessita do hiperparâmetro do número de clusters, a Clusterização Hierárquica dispensa esse hiperparâmetro (embora seja possível determinar um número específico), demonstrando, através do dendrograma a possível melhor quantidade de clusters para aquele dataset.

Na Clusterização hierárquica, ainda, os dados dos agrupados nos níveis mais baixos vão sendo reagrupados em grupos cada vez maiores nos níveis acima, ao passo em que no K-Means, os dados migram de um grupo para o outro, conforme o deslocamento do centroide, até o momento da convergência.

Parte 4. Escolha dos algoritmos

1. Escreva em tópicos as etapas do algoritmo de K-médias até sua convergência.

R: Embora seja um modelo de aprendizado não supervisionado, é necessário atribuir um hiperparâmetro para que o K-means inicie suas etapas. No caso, esse hiperparâmetro será o número de "clusters" (ou grupos) dos quais os dados se dividirão.

Com esse hiperparâmetro, o algoritmo seguirá as seguintes etapas até sua convergência:

(i) Num primeiro momento, o algoritmo irá sortear, aleatoriamente, as posições iniciais dos centroides;
(ii) A partir daí, ele irá agrupar os dados mais próximos das posições iniciais dos centroides, calculando a média da distância entre os dados e o centroide, utilizando a distância euclidiana para tanto;
(iii) Após isso, o algoritmo vai deslocar a posição dos centroides para o meio dos dados agrupados, novamente calculando a distância euclidiana;
(iv) Nisso, um novo agrupamento irá ocorrer, com os dados mais próximos das novas localizações dos centroides, de modo que os dados poderão migrar de um cluster para o outro;
(v) Os passos 'iii' e 'iv' se repetirão até que o deslocamento dos centroides e dos agrupamentos novos sejam nulos ou minimos, o que chamamos de convergência.

2. O algoritmo de K-médias converge até encontrar os centróides que melhor descrevem os clusters encontrados (até o deslocamento entre as interações dos centróides ser mínimo). Lembrando que o centróide é o baricentro do cluster em questão e não representa, em via de regra, um dado existente na base. Refaça o algoritmo apresentado na questão 1 a fim de garantir que o cluster seja representado pelo dado mais próximo ao seu baricentro em todas as iterações do algoritmo.

Obs: nesse novo algoritmo, o dado escolhido será chamado medóide.

R: Respondido no ponto "5.3. Análise por gráfico" do arquivo pd_clusterizacao.ipynb.

3. O algoritmo de K-médias é sensível a outliers nos dados. Explique.

R: Como explicado na pergunta 1 desta Parte (“4. Escolha dos algoritmos”), o algoritmo do K-Means calcula a posição dos centroides utilizando a média das distâncias extremas dos dados (através da distância euclidiana). Desse modo, se houver algum “outlier” presente (ou seja, algum dado, cujo valor seja exorbitante em relação aos outros), esses dados podem afetar substancialmente a centralização dos centroides porque serão levados em conta na etapa de deslocamento do centroide, fazendo com que sua localização fique substancialmente alterada no momento da convergência.

4. Por que o algoritmo de DBScan é mais robusto à presença de outliers?

R: O DBSCAN é mais robusto a outliers em razão da característica própria do método de agrupamento, no qual ele identifica apenas os dados que se encontram dentro dos hiperparâmetros de “eps” (raio de busca) e o “MinPts” (número de pontos de dados mínimos que o raio de busca daquele dado precisa encontrar).

Se os dados atenderem ambos os parâmetros, serão chamados de ‘core points’, se os dados estiverem dentro do raio de algum vizinho, mas ele, por si só não atender o número minimo de pontos, será chamado de “border point” e os dados que não atenderem nenhum desses parâmetros (possíveis outliers) serão “noises” e ficarão de fora do agrupamento. Assim, os outliers não afetam a implementação do DBSCAN.