Tarefa 1_B ---------- TEG

Em duplas, por favor mantenham as mesmas duplas da tarefa_1_A

Entrega por *upload* via Moodle acompanhada de relatório e demais dados.

Objetivo geral: implementar computacionalmente uma estrutura de dados que represente um grafo e aplicar essa estrutura em um estudo prático.

Objetivos específicos:

1. Utilizar o grafo como modelo para clustering (agrupador);
2. Treinar o modelo com base em algoritmos da teoria de grafos;
3. Avaliar o treinamento do clustering por meio de métricas usuais de Machine Learning;
4. A partir dos resultados obtidos na tarefa 1_A, deseja-se avaliar o grau de sucesso na separação de espécies da flor Iris (classes setosa, virgínica e versicolor) em componentes conexos distintos (clusters).

Requisitos funcionais:

1. Estudo sobre a distribuição de componentes conexos sobre o grafo da tarefa 1A;
2. Grafo com carga primária a partir de um arquivo CSV conforme a tarefa 1A;
3. Treinamento do modelo visando a separação em 3 agrupamentos disjuntos;
4. Cada agrupamento será caracterizado individualmente como grupo-setosa, grupo-virgínica ou grupo-versicolor;
5. O centro geométrico de cada grupo deve ser determinado;
6. Exibição do grafo final com os agrupamentos determinados no treinamento;
7. Determinação das métricas de avaliação e sua análise.

Requisitos não funcionais:

1. Programação em C (exceto para a exibição de grafo via Python);
2. O estudo sobre a distribuição de componentes conexos (clusters) sobre o grafo da tarefa 1A deve ser feito com base em algoritmos de BFS ou DFS adaptados para tal finalidade;
3. O estudo utiliza limiares para analisar a relação: número de componentes conexos (clusters) X tamanho de cada componentes conexo. Utilize histogramas para sintetizar os dados;
4. A determinação do “nome” de cada componente (grupo-setosa, grupo-virgínica ou grupo-versicolor) ocorrerá por meio do predomínio de certa instância no grupo (se predomina setosa, o grupo é setosa).
5. O centro de cada grupo é calculado pela média das coordenadas (largura de pétala, comprimento de pétala, largura de sépala, comprimento de sépala) de todos os seus membros;
6. Podem ser necessários ajustes complementares na determinação dos clusters;

7. As métricas de avaliação: matriz de confusão, TPR, FPR, etc (veja nos fundamentos). A determinação das métricas utiliza as classificações originais das flores;

Fundamentação:

Clustering é uma técnica de Machine Learning de aprendizado não supervisionado, ou seja, durante o treinamento são desconhecidos os rótulos que identificam as classes das instâncias de dados.

O treinamento visa separar “às cegas” o conjunto de dados em um número de três clusters que sabemos existir (vamos usar essa informação a priori).

Uma vez determinados os clusters, cada agrupamento poderá ser identificado com um tipo de flor (conforme descrito nos requisitos)

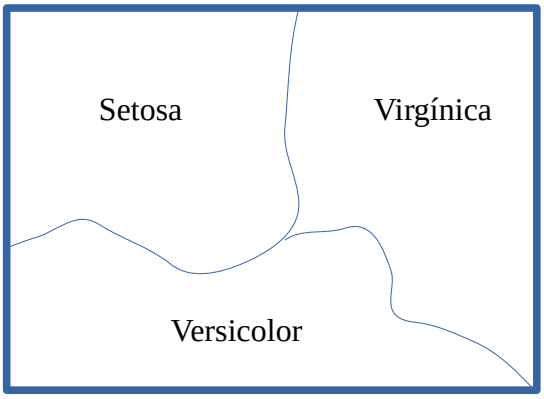

Figura 1: Busca-se um grafo desconexo que seja composto por 3 clusters, ou seja, G(V,E), V=|Vse| U |Vvi| U |Vve|

Ocorre que esse treinamento pode incorrer em erros, sendo necessário avaliar os resultados.

Por exemplo, um certo agrupamento (componente conexo) C2 pode estar agrupando todos os casos de versicolor e possivelmente algumas ocorrências de casos que podem estar erroneamente nesse grupo. Diante desses fatos, deseja-se avaliar a acurácia dos resultados obtidos.

Com as premissas acima descritas é possível construir uma matriz de confusão (Figura 1), identificando os casos TP (true positive), FP (false positive), TN (true negative) e FN (false negative) e realizar a extração das métricas de qualidade da classificação, por exemplo a acurácia:

$$\text{Acurácia}=(TP+TN)/(TP+FP+TN+FN).$$

Outras métricas devem ser levantadas adicionalmente: Recall, Precisin, F1-score (https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9).

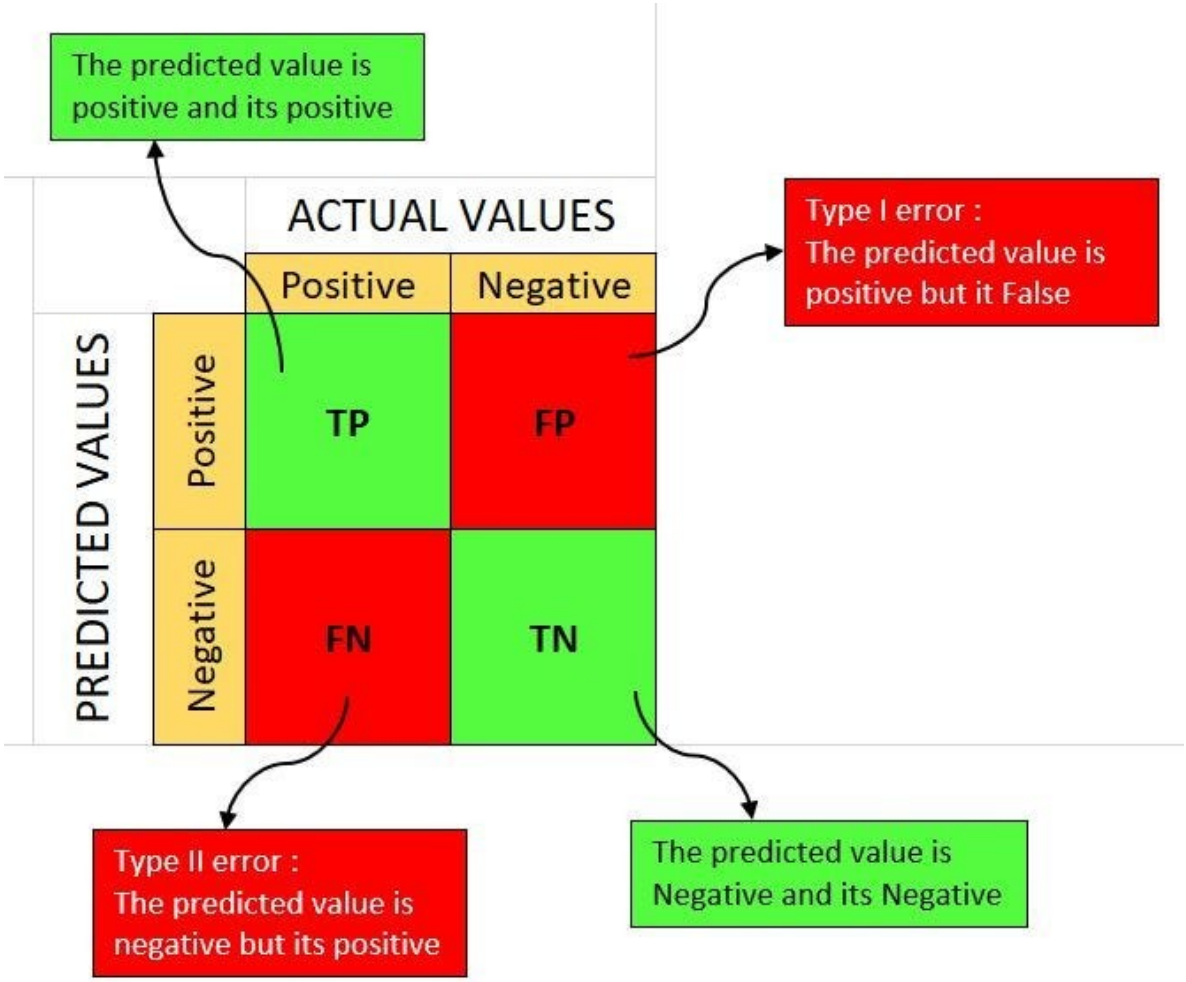

Figura 1: Matriz de confusão Para duas classes. Fonte: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

O exemplo acima é um modelo de classificação com apenas 2 saídas, então obtivemos uma matriz de confusão 2 X 2.

A Figura 2 exibe uma situação para três classes. Nesse caso, como calcular TP, FN, FP e TN?

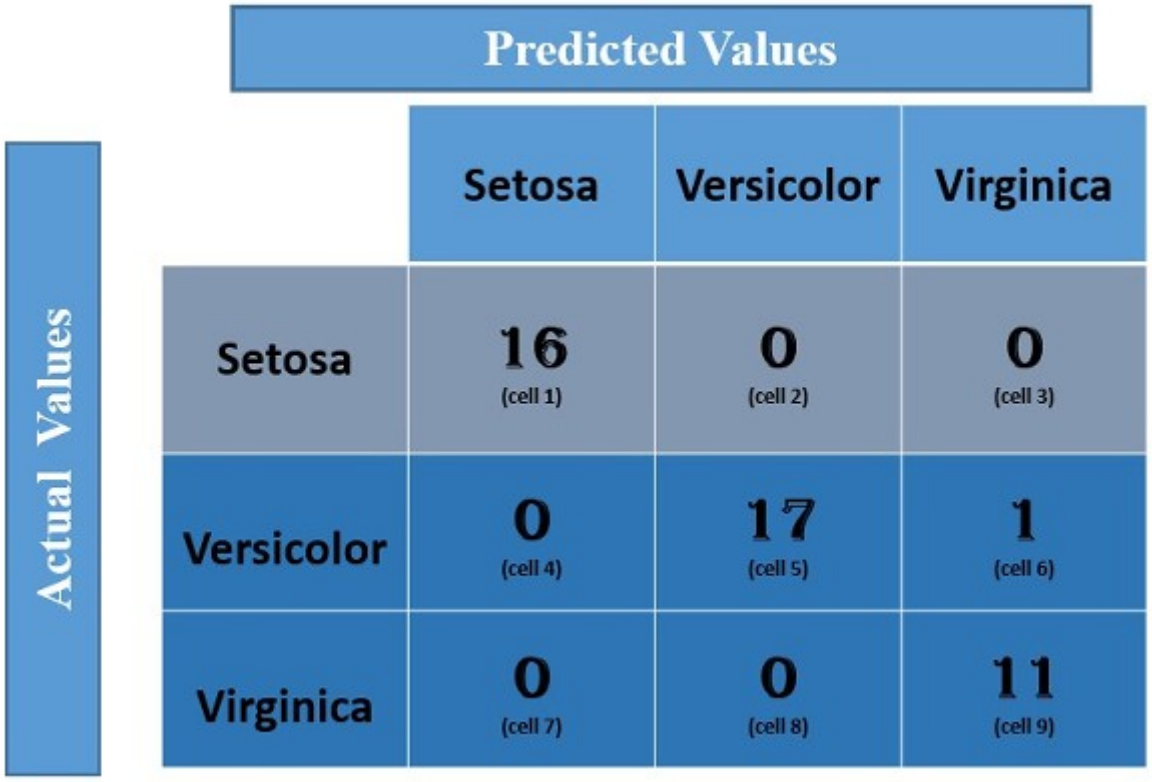

Figura 2: Exemplo de matriz de confusão para três classe. Fonte: https://www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classification/s

Calculando os valores de TP, TN, FP e FN para a classe Setosa usando a matriz na Figura 2, acima:

TP (setosa): O valor real e o valor previsto devem ser iguais. Portanto, em relação à classe Setosa, o valor da célula 1 é o valor TP.

- TP=16

FN (setosa): A soma dos valores das linhas correspondentes, exceto para o valor TP:

- FN = (célula 2 + célula 3) = (0 + 0) = 0

FP (setosa): A soma dos valores da coluna correspondente, exceto para o valor TP:

- FP = (célula 4 + célula 7) = (0 + 0) = 0

TN (setosa): a soma dos valores de todas as colunas e linhas, exceto os valores dessa classe para a qual estamos calculando os valores.

- TN = (célula 5 + célula 6 + célula 8 + célula 9) = 17 + 1 +0 + 11 = 29

Raciocínio similar é aplicado para obter TP, TN, FP, FN para as outras classes.
Exemplo para a classe Versicolor: os valores/métricas são calculados conforme abaixo:

- TP: 17 (célula 5);
- FN: 0 + 1 = 1 (célula 4 + célula 6);
- FP: 0 + 0 = 0 (célula 2 + célula 8) e
- TN: 16 +0 +0 + 11 =27 (célula 1 + célula 3 + célula 7 + célula 9).

Para o modelo utilizado na tarefa corrente:

Supondo que o treinamento do modelo forneça três clusters identificadas como Ve (onde predominam casos da variedade versicolor), Vi (onde predominam casos da variedade virgínica) e Se (para o predomínio das Setosas), conforme a Figura 3.

Analisando: TP, FP, TN e FN para o grupo Ve:

Nesse caso, o modelo prediz que todos os elemento do grupo Ve são do tipo Versicolor (positivos para Ve). No entanto, a realidade demonstrada na base de dados identifica dois elementos (Vi e Se) que não são de fato da classe Ve e estão erroneamente incluídos nesse grupo. Portanto há dois falso-positivos (FP) para Ve.

Adicionalmente, O modelo prediz que os grupos Vi e Se possuem apenas os casos que não são do tipo Ve, ou seja, os casos em Vi e Se são negativos para Ve. No entanto pela base de dados há três casos que são de fato do tipo Ve e estão distribuídos dentro de outro grupo (Vi ou Se).

Esses três casos deveriam ser negativos para Ve (ou seja Vi ou Se), mas não o são, então se diz que esses dois casos são falso-negativos (FN) para Ve.

Os casos TP para Ve são os casos Ve que estão dentro do grupo Ve.
Os casos TN para Ve são aqueles Vi e Se que ocorrem fora de Ve.

Na Figura 3:
FP (Versicolor) = 2, FN (Versicolor) = 3, TP (Versicolor) = 4, e TN (Versicolor) = 7

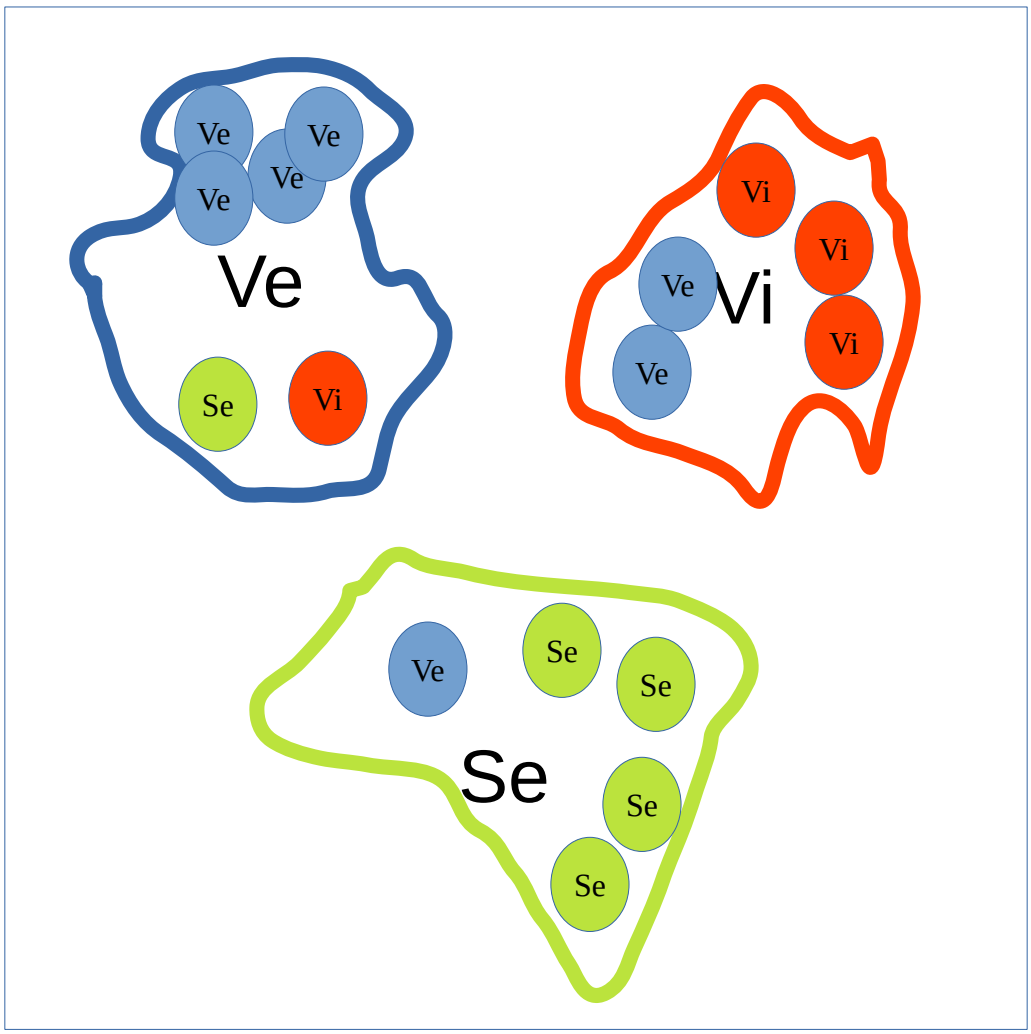

Figura 3: Exemplo de um resultado do treinamento do modelo proposto.