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

Mantenham as mesmas equipes da tarefa_1_A

Objetivo geral: determinação de dois agrupamentos de componentes conexos sobre o grafo da tarefa1A usando algoritmos BFS ou DFS adaptados.

Linguagem de programação: C (exceto para a exibição de grafo via Python);

Objetivos específicos:

1. Utilizar o grafo como modelo para clustering (agrupador) com base no conceito de componentes conexos;
2. Obter dois agrupamentos finais;
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. Deseja-se avaliar o grau de sucesso na separação de espécies da flor Iris em duas classes.

Requisitos funcionais:

1. Estudo sobre a distribuição de componentes conexos sobre o grafo da tarefa 1A (Para o grafo obtido na tarefa_1A, informe quantos componentes conexos o grafo possui e quais são seus tamanhos – quantidades de vértices);
2. Treinamento do modelo visando a separação em 2 grandes agrupamentos disjuntos;
3. Cálculo do centro de cada agrupamento e, se necessário, a agregação dos agrupamentos (fragmentos) aos dois maiores agrupamentos, o objetivo é obter apenas dois agrupamentos (possivelmente: setosa e não setosa);
4. Exibição do grafo final;
5. Persistência do grafo final pós agrupamento em duas clusters;

Requisitos não funcionais:

1. Programação em C (exceto para a exibição de grafo via Python);
2. O estudo utiliza a distância euclideana normalizada e limiares para construir as arestas do grafo;
3. 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;
4. 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 final dos agrupamentos (agregação de vértices isolados e/ou microgrupos);

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)

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:

Acurácia=(TP+TN)/(TP+ FP+ TN+ FN).

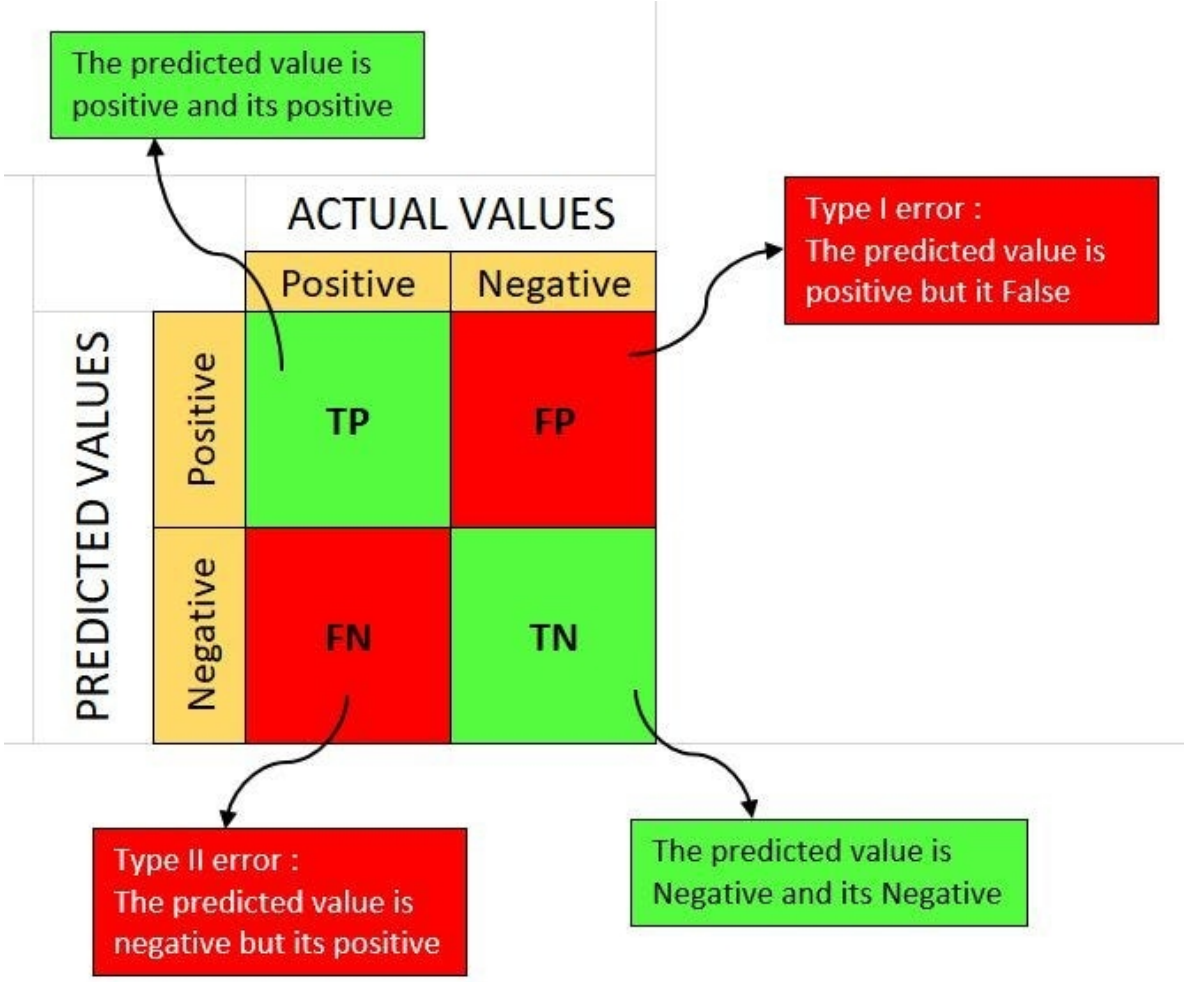

Figura 1: Matriz de confusão Para duas classes. Fonte: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

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

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