Relatório Semanal 1

Caius C. Souza

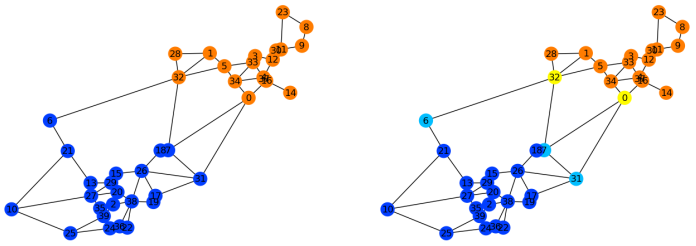
23/04/2024

1 Objetivos

- Construir um grafo de Gabriel a partir de um dataframe.
- Estudar e aplicar técnicas de remoção de ruídos.
- Construir um classificador para aplicar o grafo

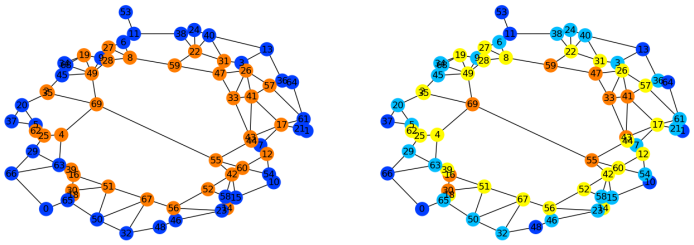
2 Construção do grafo

O Grafo de Gabriel foi implementado em Python por meio de uma classe cujos métodos, além de calcularem o grafo, retornam sua representação por meio de matriz de adjacências e calcula os centros do grafo. O cálculo de tais centros, i.e. nós pertencentes à margem de separação das classes, é importante pois são peças fundamentais para classificadores que utilizam o grafo.



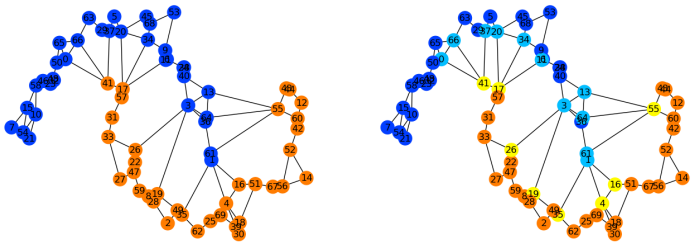
(a) Sem centros (b) Com centros

Figure 1: Grafo de Gabriel de duas gaussianas com a presença ou não de centros representados.



(a) Sem centros (b) Com centros

Figure 2: Grafo de Gabriel para o problema de círculos com a presença ou não de centros representados.



(a) Sem centros (b) Com centros

Figure 3: Grafo de Gabriel para o problema de duas luas com a presença ou não de centros representados.

3 Remoção de ruídos - Wilson editing

Um dos problemas relacionados não só ao Grafo de Gabriel, mas a grande maioria de algoritmos de reconhecimento de padrões são a presença de ruídos, overlap e outliers nos dados. Um bom desempenho de um modelo, principalmente no que se refere a sua capacidade de generalização, depende grandemente do conjunto de dados fornecidos para seu treinamento. Dados muito ruidosos, neste sentido, podem levar o modelo ao overfitting, fazendo com que seu desempenho e capacidade de predição sejam restritos ao conjunto de treinamento. Dados menos ruidosos, por outro lado, podem gerar superfícies de separação mais simples e eficazes. Nessa lógica, foi testado um método de redução de ruídos baseado no algoritmo kNN: o Wilson editing. Tal método, em um problema de classificação binária onde $y_i \in \{-1, 1\}$ funciona da seguinte maneira:

Algoritmo da Edição de Wilson:

(A) Para cada amostra V_i pertencente ao dataset original V:

  (A.1) Encontre seus k vizinhos mais próximos

  (A.2) Realize o somatório dos rótulos dos k vizinhos e aplique a função sinal ao resultado

  (A.3) Se V_i for igual ao resultado, aloque-o a um dataset auxiliar V'

(B) O dataset auxiliar V' é resultante da remoção de ruídos do dataset original V

Para mostrar o resultado da suavização, um grafo de Gabriel foi construído para um problema de duas gaussianas bidimensionais balanceadas com considerável desvio-padrão. Os dados foram suavizados utilizando o método de edição de Wilson tomando-se um k=10.

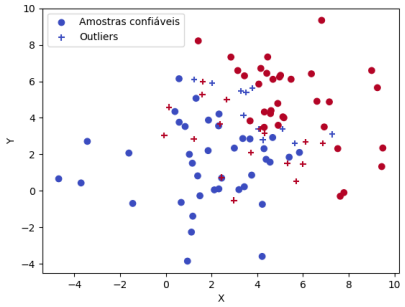
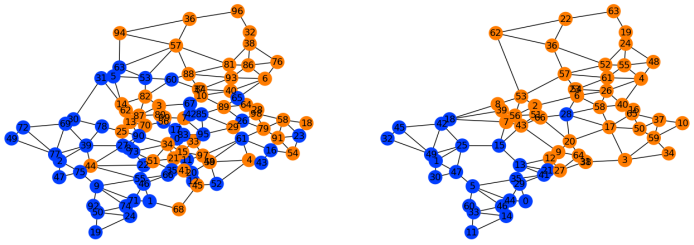


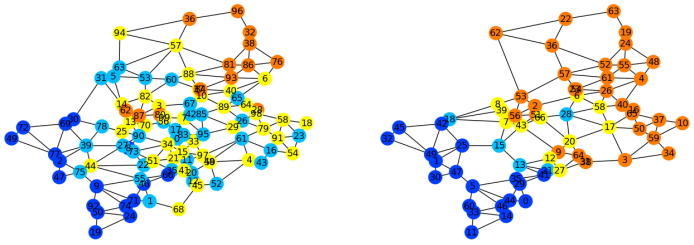
Figure 4: Edição de Wilson realizada em gaussianas



(a) Modelo sem suavização (b) Modelo suavizado com k=10

Figure 5: Representação gráfica comparativa entre modelo suavizado por Wilson editing e não suavizado.

Os centros pré e pós tratamentos podem ser visualizados a seguir.



(a) Modelos com centros, sem suavização

(b) Modelo com centros, suavizado com k=10

Figure 6: Representação gráfica comparativa entre modelo suavizado por Wilson editing e não suavizado, com centros.

4 Modelos de classificação

Para testar o desempenho do grafo de Gabriel, o modelo foi testado em uma rede neural do tipo RBF centrada na margem de separação que realiza classificação binária.

4.1 Gaussianas

A base de dados consiste em duas gaussianas bidimensionais centradas nos pontos (2, 2) e (4, 4). As duas gaussianas possuem um alto desvio-padrão de 3. Cada gaussiana conta com 150 amostras.

4.1.1 Grafo de Gabriel e redução de ruídos

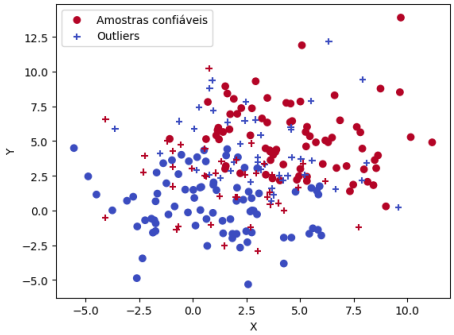
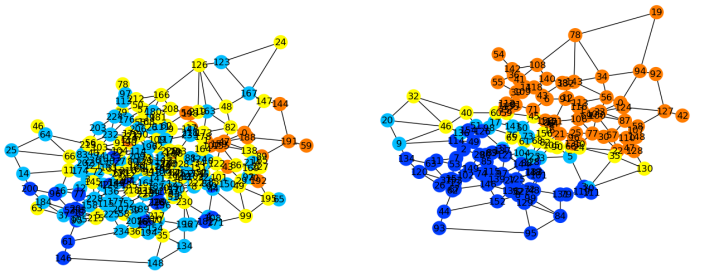


Figure 7: Dados e outliers identificados com base na Edição de Wilson



(a) Pré-remoção

(b) Pós-remoção, suavizado com k=10

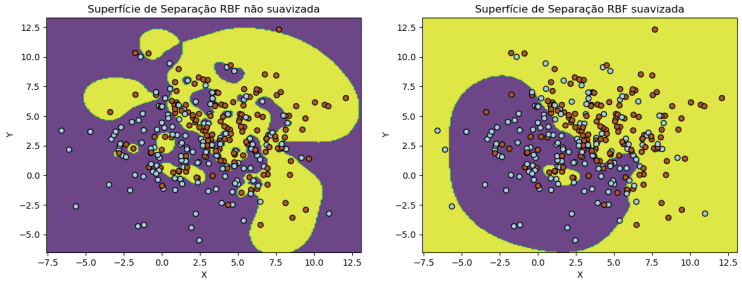
Figure 8: Representação gráfica comparativa entre modelo suavizado por Wilson editing e não suavizado

4.1.2 Desempenho do modelo

O desempenho da rede GGRBF (Gabriel Graph centered Radial Basis Function) com sigma = 1 está descrito a seguir:

Table 1: Desempenho Gaussianas (Média ± σ)

Modelo	Acurácia	AUC
GGRBF não suavizada	56,67 ± 0,00 %	0,56 ± 0,00
GGRBF suavizada	65,00 ± 0,00 %	0,64 ± 0,00



(a) Superfície do modelo sem suavização

(b) Superfície do modelo suavizado

Figure 9: Diferença entre superfícies de separação

4.2 Duas luas

O dataset consiste em duas meias luas em um plano contando com 100 amostraas de cada classe e um ruído=0.5.

4.2.1 Grafo de Gabriel e redução de ruídos

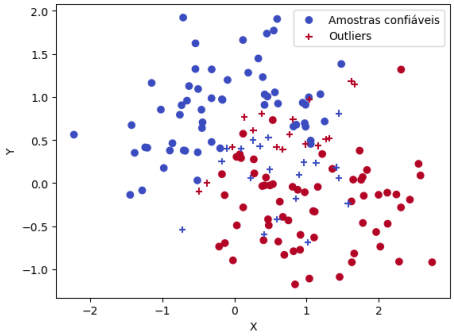
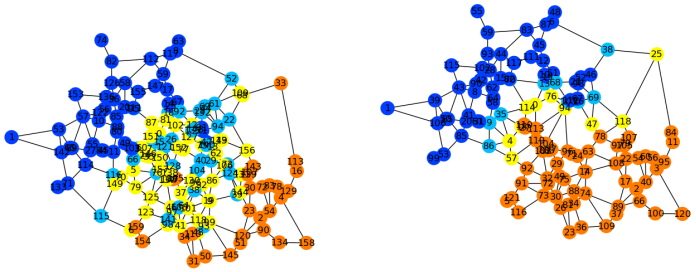


Figure 10: Dados e outliers identificados com base na Edição de Wilson



(a) Pré-remoção

(b) Pós-remoção, suavizado com k=10

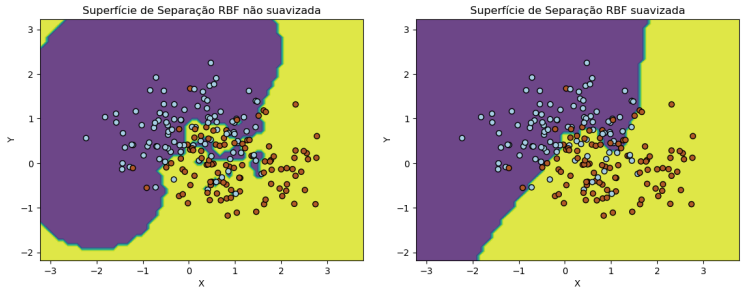
Figure 11: Representação gráfica comparativa entre modelo suavizado por Wilson editing e não suavizado

4.2.2 Desempenho do modelo

O desempenho da rede GGRBF (Gabriel Graph centered Radial Basis Function) com sigma = 1 está descrito a seguir:

Table 2: Desempenho Gaussianas (Média ± σ)

Modelo	Acurácia	AUC
GGRBF não suavizada	75,00 ± 0,00 %	0,75 ± 0,00
GGRBF suavizada	77,50 ± 0,00 %	0,77 ± 0,00



(a) Superfície do modelo sem suavização (b) Superfície do modelo suavizado

Figure 12: Diferença entre superfícies de separação

4.3 Breast Cancer

Após isso, verifica-se o desempenho do modelo no conhecido dataset Breast Cancer.

4.3.1 Desempenho

Tem-se a seguinte curva de desempenho do modelo para cada valor de k, onde k=0 representa o modelo sem a edição de Wilson.

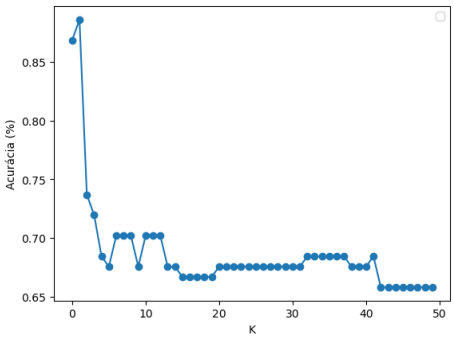


Figure 13: Desempenho do modelo baseado no valor de K.

Graficamente, nota-se que o valor de K ótimo para o modelo é 1.

5 Referências

- Dataset Editing Techniques: A Comparative Study