Exercício Reconhecimento de Padrões
Mistura de Gaussianas

Lucas Ribeiro da Silva - 2022055564
Universidade Federal de Minas Gerais
Belo Horizonte - Minas Gerais - Brasil

lucasrsilvak@ufmg.br

1 Introdução

Nesse relatório consta a realização e implementação da estimação de deverá usar a misturas de gaussianas para alimentar um classificador de Bayes na resolução de um problema de classificação.

2 Base de Dados

Os dados foram carregados do ml.bench com 400 amostras, 1 ciclo e um desvio-padrão de 0.05.

3 Validação Cruzada e Testes de Gaussianas

No próximo passo, foi efetuada a separação em 10 folds e utilizada o número de k de gaussianas que proporcionou maior acurácia.

Fold	2	3	4	5
Fold 1	80%	95%	100%	97.5%
Fold 2	75%	100%	97.5%	100%
Fold 3	70%	100%	100%	100%
Fold 4	65%	97.5%	97.5%	100%
Fold 5	47.5%	97.5%	100%	100%
Fold 6	77.5%	97.5%	97.5%	100%
Fold 7	72.5%	82.5%	97.5%	100%
Fold 8	72.5%	95%	97.5%	100%
Fold 9	82.5%	97.5%	97.5%	100%
Fold 10	77.5%	87.5%	100%	100%
Acurácia média	72%	95%	98.5%	99.75%
Desvio padrão	9.99%	5.65%	1.29%	0.79%

Tabela 1: Resultados de Acurácia e Desvio Padrão por Fold/k

Em seguida, definiu se k como 5, por considerar-se que a acurácia já estava satisfatória. Os clusters foram dividios como mostra a imagem.

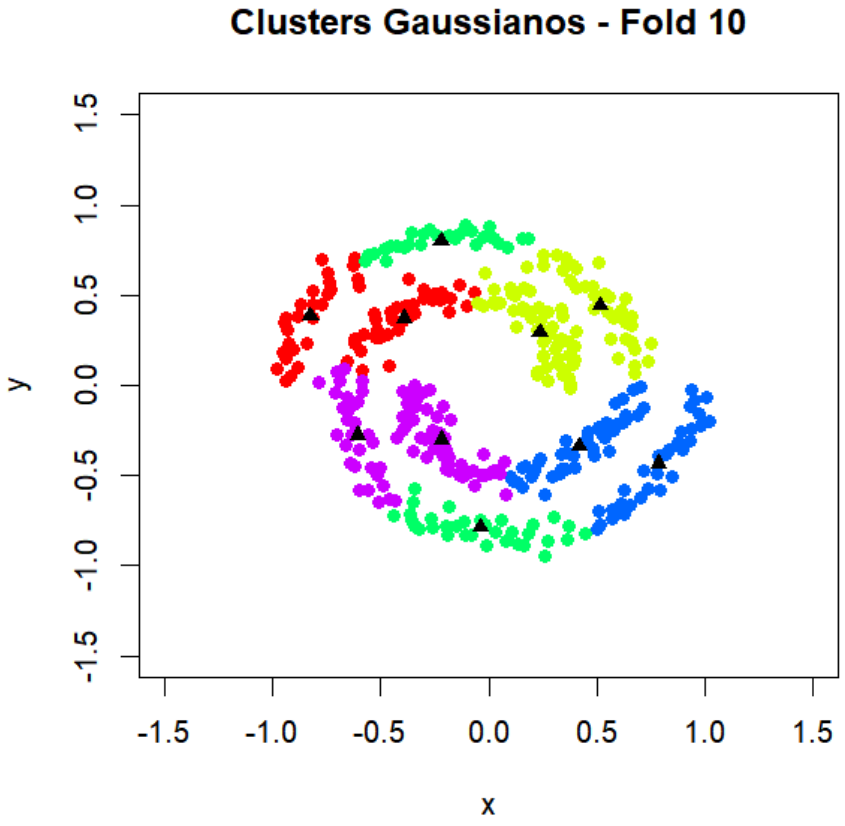

Figura 1: Clusters

4 Plot do Melhor Fold de Treinamento

Em seguida, foi escolhido um dos folds com 100% de acurácia para plotar a classificação do modelo de classificação bayesiana, a fronteira de divisão segue-se como no gráfico abaixo.

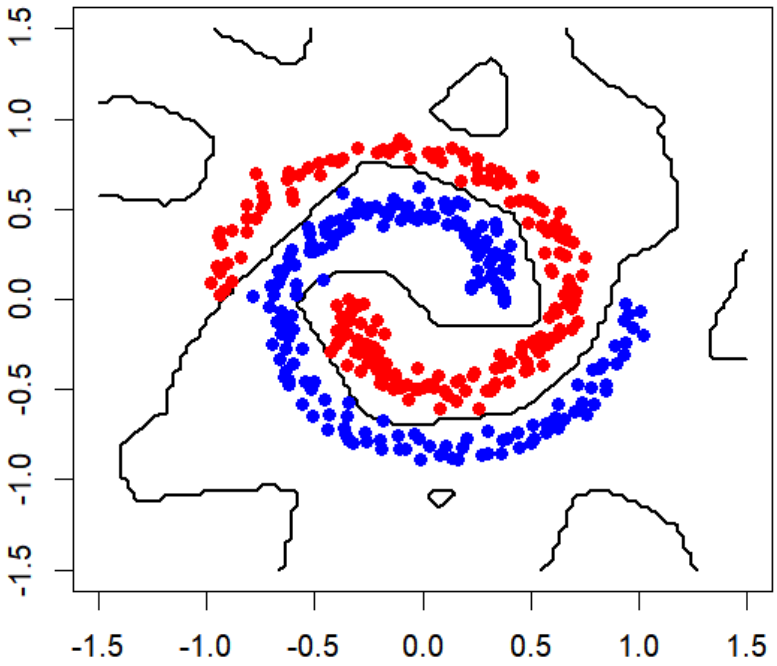

Figura 2: Classificador Bayesiano em Mistura de Gaussianas

5 Conclusões

A mistura de Gaussianas é um método que permite a aplicação de funções de densidade de probabilidade em funções que não apresentam, de ínicio, uma distribuição normal. Isso pode ser realizado amontoando os dados em clusters e calculando uma distribuição normal para cada um dos clusters, tornando a verossimilhança um somatório de gaussianas. O k é o hiperparâmetro a ser definido nestes casos, com um k baixo sendo insatisfatório em dividir o problema das espirais, enquanto um k alto pode ser ótimo, mas computacionalmente mais caro.