UNIVERSIDADE FEDERAL DE MINAS GERAIS
Graduação em Engenharia de Sistemas
Introdução ao Reconhecimento de Padrões - Aulas 9 e 10
Nikolas Dias Magalhães Fantoni - 2018019400

Mistura de Gaussianas

Exercício 1

No exercício proposto, foram gerados dados em formato de espiral pertencentes a duas classes distintas, conforme mostra a figura 1.

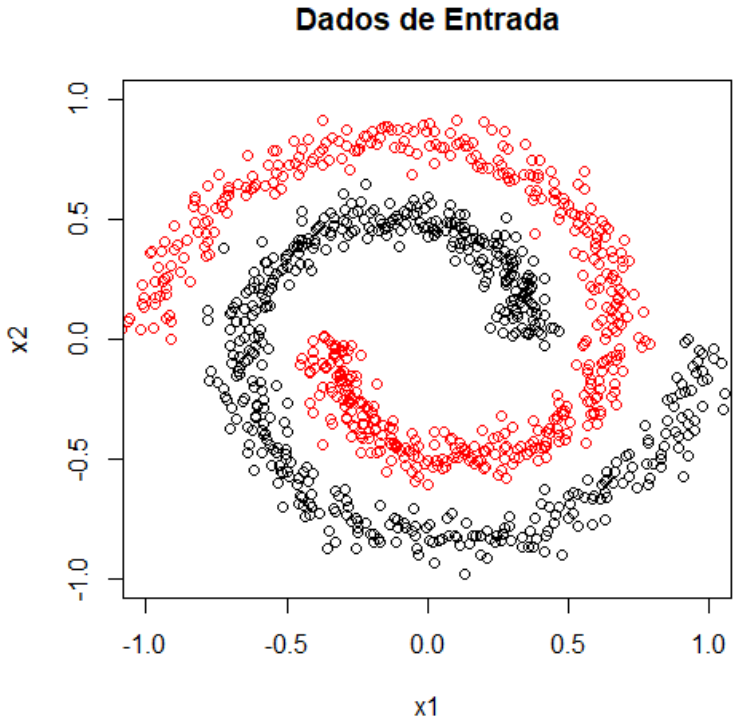

Figura 1: Dados de entrada para o problema.

Os dados foram divididos em 10 *folds* para a técnica de validação cruzada, sendo que a cada iteração, um grupo funcionava como dado de teste e os demais como dados de entrada.

Foi então decidido em quantos *clusters* seriam divididos os dados utilizando o algoritmo *k-means*. Para isso, o algoritmo foi rodado em *looping* com k sendo incrementado em uma unidade a cada iteração. O critério de parada foi de que a média dos valores classes de cada amostra dentro de cada *cluster* não podia ter uma diferença maior que 1% do valor das classes exatas. Por exemplo, um *cluster* que originalmente era da classe 1 deve ter a média da classificação das classes dos dados entre $0.99 < \mu < 1.01$. Com isso, o algoritmo retornou k agrupamentos para cada *fold*, um deles com a divisão e centros mostrados na figura 2.

Assim, o classificador Bayesiano foi aplicado nos k *clusters* para cada *fold* e a acurácia do classificador foi salva. Os resultados estão mostrados na tabela abaixo. O desvio padrão foi de $\sigma = 0,69\%$. Para o gráfico da superfície de separação das classes, foi escolhido o último *fold* devido à facilidade da análise

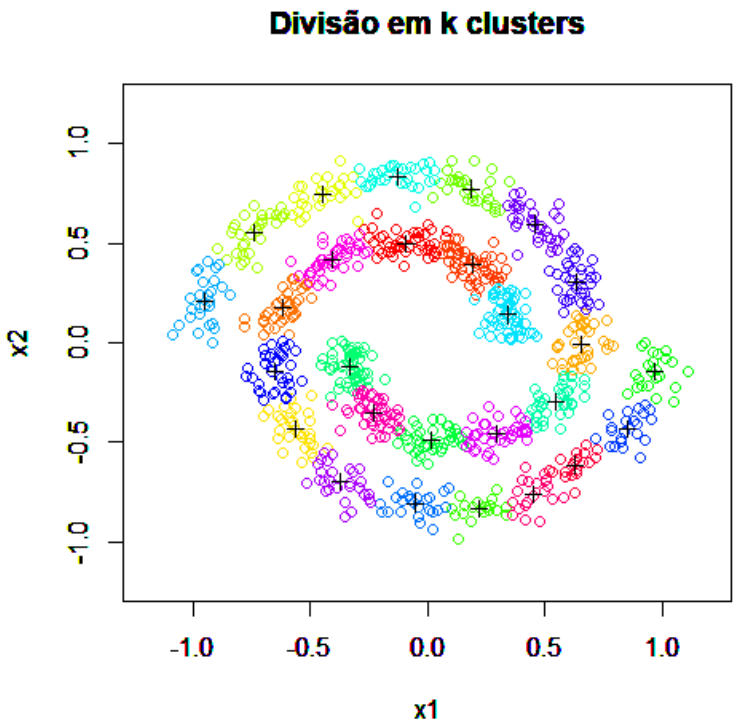

Figura 2: Dados de entrada divididos em k classes.

e o fato de possuir acurácia máxima do problema. Para este *fold*, o valor de k foi de 27 *clusters*. A superfície de separação é mostrada na figura 3.

	Acurácia
	98%
	99%
	100%
	100%
	100%
	100%
	100%
	99%
	100%
	100%
Média:	99,6%

Tabela 1: Matriz de acurácias dos folds.

Uma superfície preenchida é mostrada na figura 4, encontrada usando a função *image2D()*.

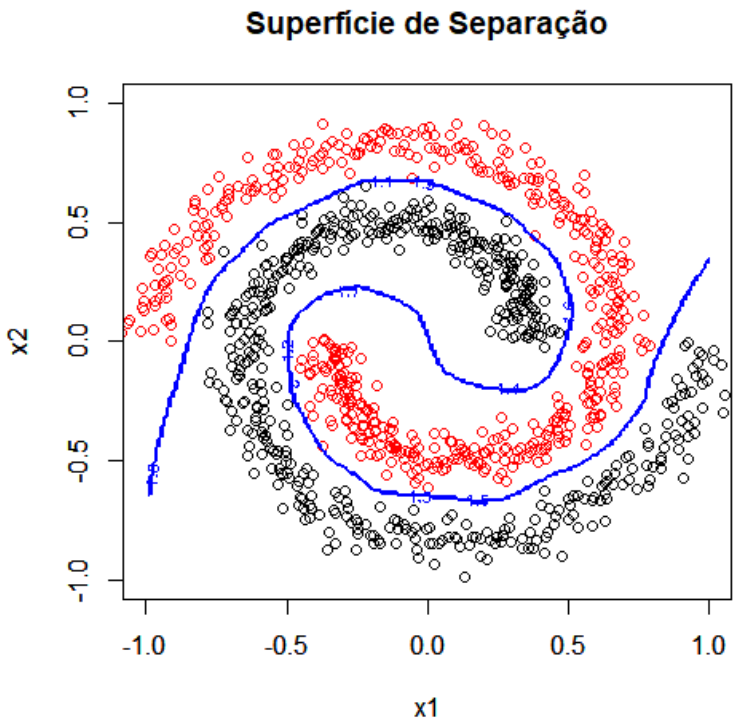

Figura 3: Superfície de separação.

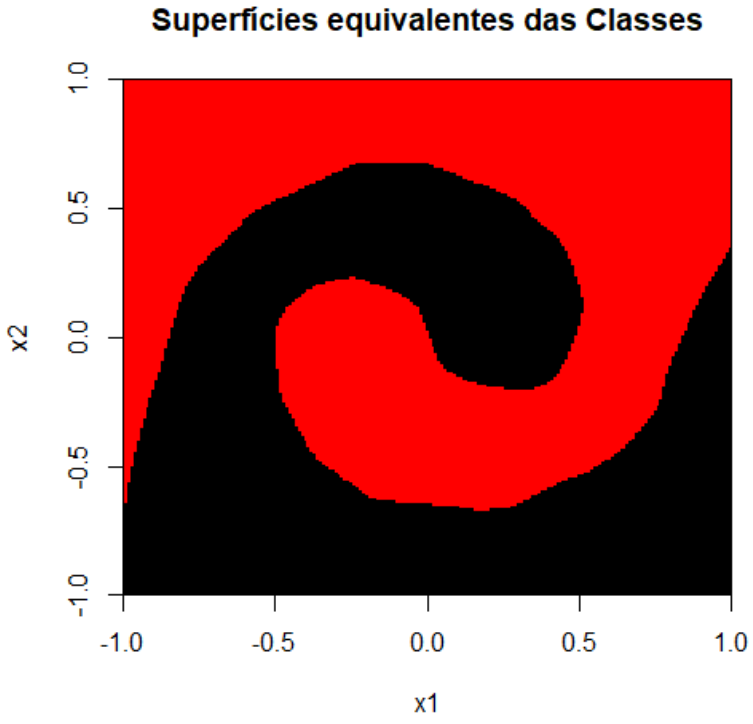

Figura 4: Superfície de separação preenchida.

Exercício 2

Na segunda parte deste trabalho, foi requirida a aplicação da metodologia do exercício 1 em uma base de dados denominada *BreastCancer*, contida no pacote *mlbench* do R. Como este problema possui um vetor de 9 variáveis na entrada, não é possível demonstrar a superfície de separação utilizando o espaço incluindo todas as variáveis. Logo, apenas a tabela das acurácias será mostrada abaixo. O desvio padrão obtido foi de $\sigma = 5,51\%$. Devido à base ser desbalanceada para uma das classes e a predominância de variáveis inteiras que podem assumir poucos valores contidas na base, os resultados são considerados satisfatórios.

	Acurácia
	72,06%
	82,61%
	75,36%
	76,47%
	85,29%
	86,96%
	86,76%
	77,94%
	86,76%
	77,94%
Média:	80,82%

Tabela 2: Acurácias obtidas no treinamento dos 10 *folds* da base *BreastCancer*.