UNIVERSIDADE FEDERAL DE MINAS GERAIS
Graduação em Engenharia de Sistemas
Introdução ao Reconhecimento de Padrões - Aulas 11 e 12
Nikolas Dias Magalhães Fantoni - 2018019400

Espaço de Verossimilhanças

Exercício 1

No exercício proposto, foram gerados dados em formato de espiral pertencentes às duas classes distintas, conforme mostra a figura 1, obtidos a partir da biblioteca *mlbench* do R.

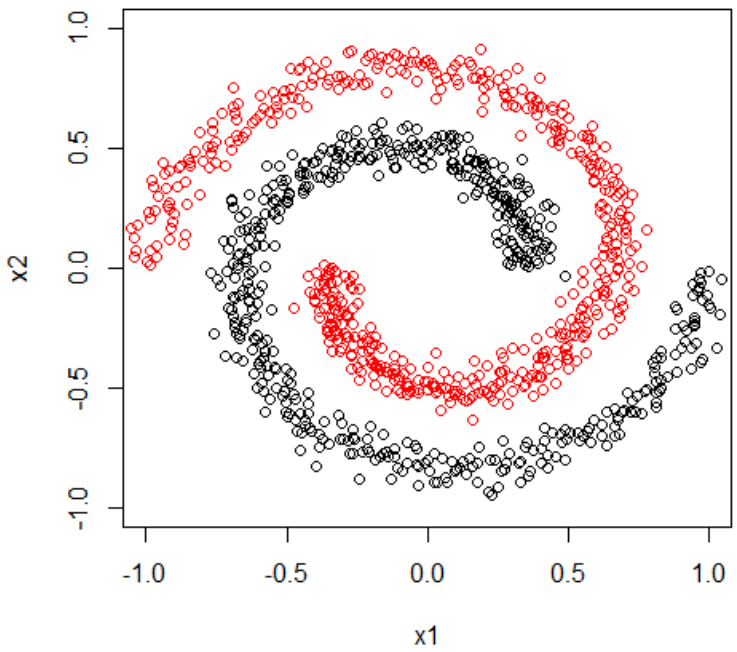

Figura 1: Dados de entrada para o problema.

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.

O resultado mostrado na figura 2 mostra o obtido quando $k = 28$, no segundo *fold* realizado pelo teste, cuja acurácia foi de 100%.

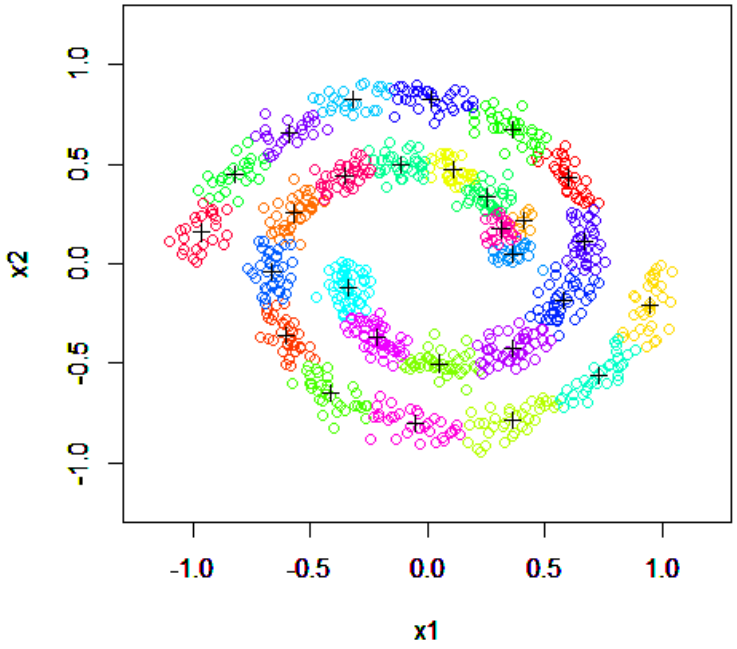

Figura 2: Dados de entrada divididos em k classes.

Assim, o espaço de verossimilhanças foi plotado a partir do valor da PDF encontrada para cada ponto do teste. Este espaço é mostrado na figura 3.

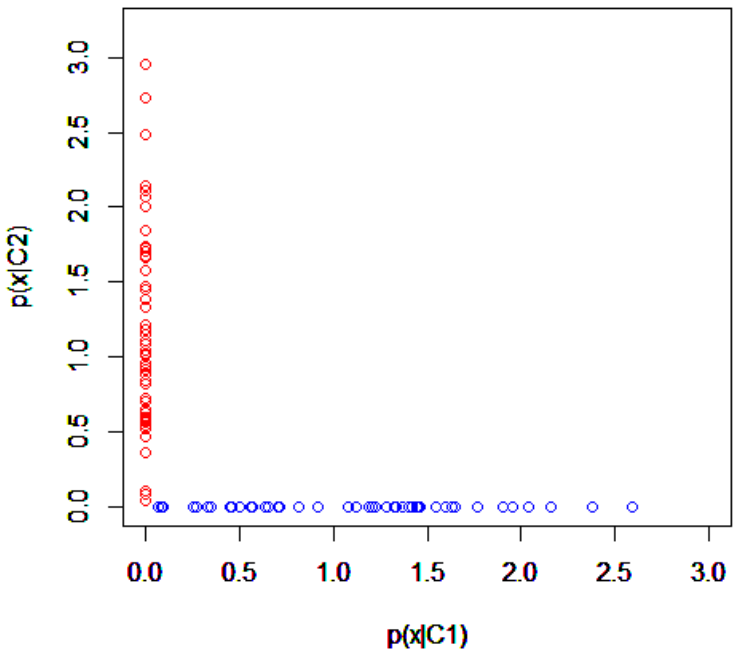

Figura 3: Espaço de verossimilhança do exercício 1.

Por fim, a reta que separa os espaços de verossimilhanças, dada por $y = ax$ em que $a = P(C_1)/P(C_2)$, é mostrada na figura 4. Pode-se notar que a curva é uma reta linear, uma vez que a distribuição de dados é balanceada. Assim, a probabilidade *a priori* da classe 1 equivale à da classe 2, tornando o problema perfeitamente separável linearmente no espaço de verossimilhança.

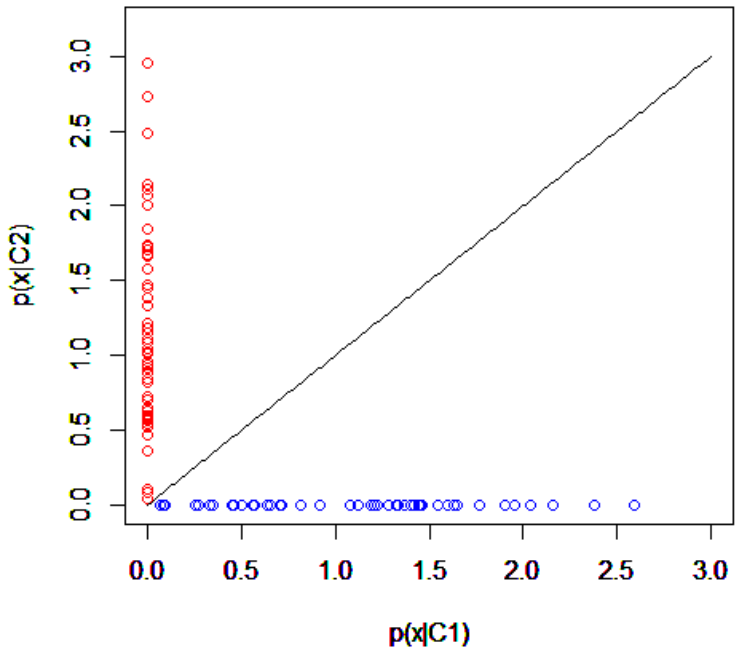

Figura 4: Superfície de separação do espaço de verossimilhança do exercício 1.

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 em um plano de duas dimensões a distribuição dos dados para analisar o problema visualmente.

Ainda assim, foram feitos os mesmos procedimentos do exercício 1 e, com isso, foram obtidos um total de $k = 5$ *folds*. Depois, foi impresso o espaço de verossimilhanças, mostrado na figura 5. Nela já é possível perceber que o eixo x encontra-se em uma ordem de grandeza diferente do eixo y. Isso ocorre pois os dados de entrada são desbalanceados para uma das classes (sendo a majoritária os dados da classe 1). Assim, o classificador bayesiano implementado, fortemente sensível ao desbalanceamento, encontrou dificuldades na classificação dos dados.

Por fim, a curva de separação do espaço de verossimilhança é mostrada na figura 6. Mais uma vez percebe-se a influência do desbalanceamento na separação dos dados. Como a probabilidade *a priori* da classe 1 é maior (cerca de 3 vezes), o gráfico aparece com uma inclinação maior. Assim, percebe-se a forte tendência do algoritmo de classificar os dados como pertencentes da classe 1. Isso explica a acurácia média entre 80% e 85% nos testes, resultado relativamente bom mas que poderia ser melhorado consideravelmente caso fosse aplicado alguma técnica de balanceamento nos dados de entrada, como um *undersampling* ou um *oversampling*.

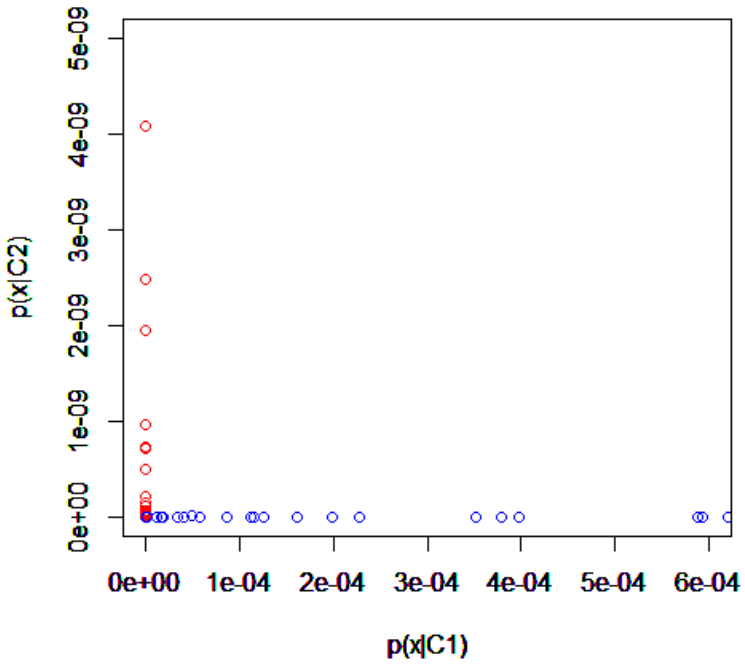

Figura 5: Espaço de verossimilhança do exercício 2.

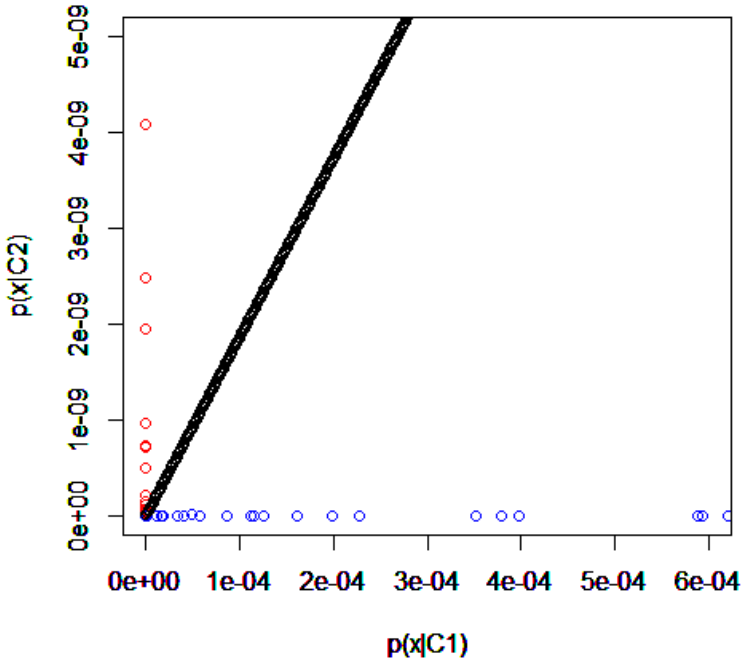

Figura 6: Superfície de separação do espaço de verossimilhança do exercício 2.