UNIVERSIDADE FEDERAL DE MINAS GERAIS

ELT135 - Introdução ao Reconhecimento de Padrões

Exercício 10 - 22/09/2019

Espaço de Verossimilhanças

Hernane Braga Pereira - 2014112627

1. Introdução

Este relatório tem como objetivo demonstrar o espaço das verossimilhanças, geradas à na classificação de problemas não linearmente separáveis.

2. Problema espiral

Para este exercício foi utilizado o problema *Espiral* com 1000 amostras, que foi gerado usando o pacote *mlbench* do R, e foi pedido que o problema fosse classificado. Como as classes 1 (vermelha) e classe 2 (preta) não são linearmente separáveis, utilizou-se o método das misturas de gaussianas para linearizar o problema, e então utilizar o classificador bayesiano para separação.

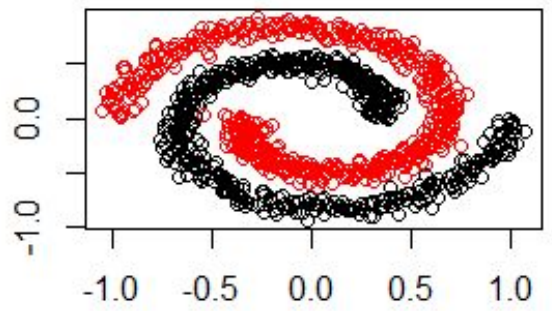

Figura 1. Problema *Espiral* utilizado

Para gerar as gaussianas, foi utilizado o método *K-means* para gerar 30 clusters, de forma que nenhum cluster estivesse em duas classes.

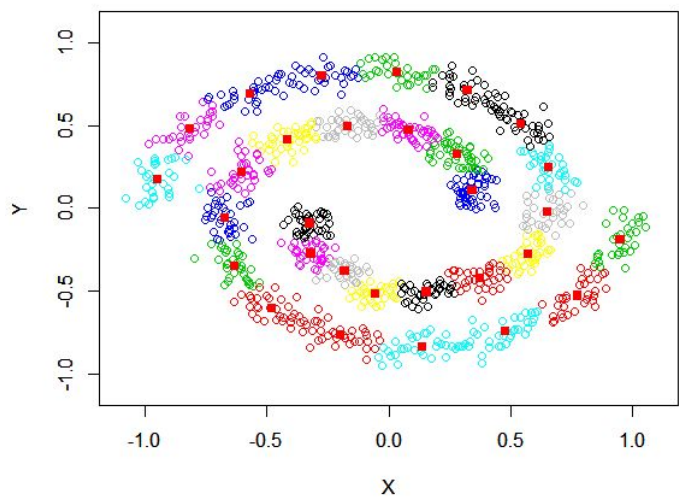

Figura 2. Clusters criados (k=30)

Após os clusters serem criados, os dados foram separados em 90% para treino e 10% para teste. Como estamos assumindo distribuição normal em cada cluster, a equação do classificador bayesiano foi adaptada para este contexto.

$$P(\mathbf{x}|S_1, \cdots, S_p) = \sum_{k=1}^{p} \pi_k \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_k)^{\mathrm{T}} \mathbf{\Sigma}_k^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_k)\right)$$

Equação 1. Probabilidade de um ponto x pertencer à classe desejada, dado que os clusters $S1,\ S2..., S_p$ pertencem à esta classe

Onde μ_k é o vetor de médias do cluster k e $\pi_k = \frac{N_k}{N}$ é o coeficiente da combinação linear onde N_k é o número de amostras do cluster k e N o número total de amostras. Após a execução do modelo o espaço das verossimilhanças encontrado foi:

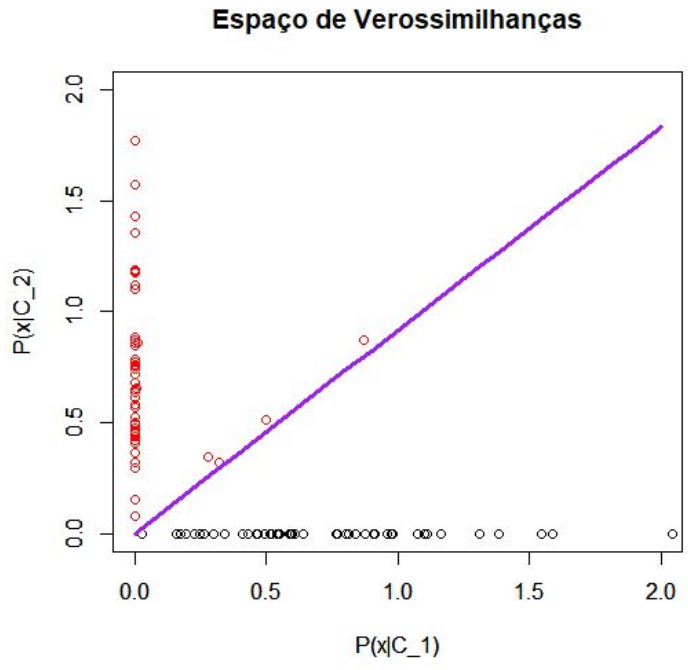

Figura 3. Espaço de Verossimilhanças do problema espiral

Ao analisar os resultados, conclui-se que o método da mistura de gaussianas foi efetivo em separar as duas classes do problema *Espiral*, o que se confirma através do gráfico do espaço da verossimilhança.

3. Problema *breastcancer*

Para este exercício foi pedido que se classifica-se o problema *breastcancer,* do pacote *mlbench* do R, utilizando as mesmas técnicas de separação de clusters e classificação utilizando as misturas de gaussianas, porém em um problema real de detecção de câncer de mama em um dataset com 9 variáveis de entrada. Foram usados 5 clusters na classificação do problema.Abaixo o gráfico do espaço de verossimilhanças encontrado:

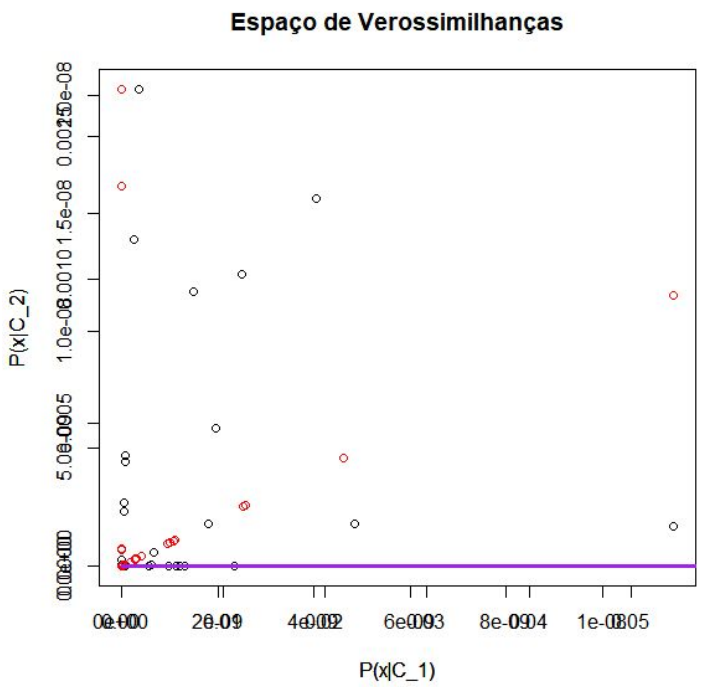

Figura 3. Espaço de Verossimilhanças do problema breastcancer

Ao analisar o gráfico, nota-se que as classes não estão visivelmente separadas no espaço da verossimilhança. Por isso, abaixo estão os gráficos de verossimilhança de cada classe para melhor visualização.

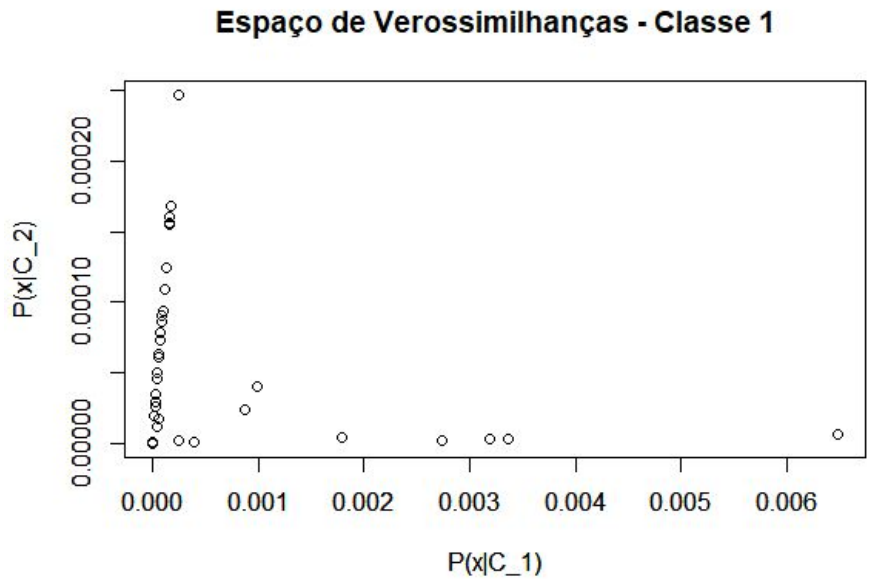

Figura 4. Espaço de Verossimilhanças do problema breastcancer para a classe 1

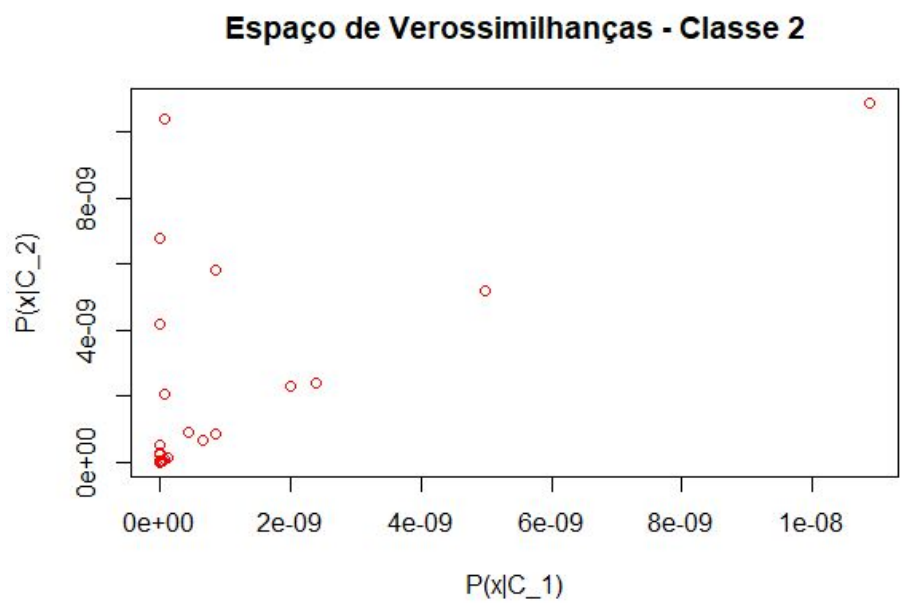

Figura 5. Espaço de Verossimilhanças do problema breastcancer para a classe 2

O classificador obteve 93.4% de acurácia para a classe 1 e 100% para a classe 2. A acurácia total foi de 95% e desvio padrão de 0.24 para a classe 1, 0 para a classe 2.

4. Referências

[1] Clustering, Notas de aula, agosto de 2019.
[2] Misturas de Gaussianas, Notas de aula, setembro de 2019.
[3] Misturas de Gaussianas: qualidade de partições, Notas de aula, setembro de 2019.