UNIVERSIDADE FEDERAL DE MINAS GERAIS

ELT135 - Introdução ao Reconhecimento de Padrões

Exercício 9 - 15/09/2019

Classificação de problemas não linearmente separáveis utilizando mistura de gaussianas

Hernane Braga Pereira - 2014112627

1. Introdução

Este relatório tem como objetivo exemplificar o uso da técnica de mistura de gaussianas, geradas à partir da separação de clusters através do método k-means, para 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.

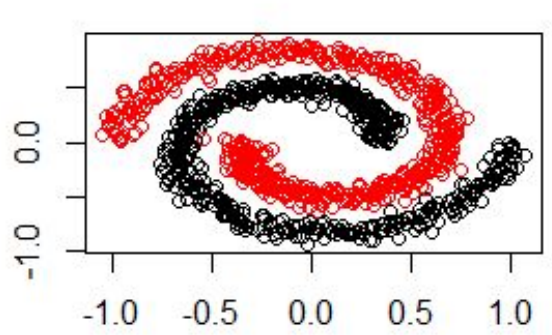

Figura 1. Problema *Espiral* utilizado

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

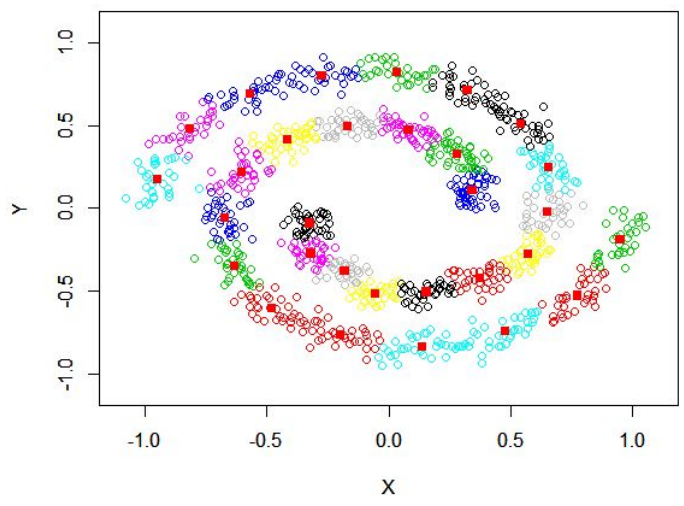

Figura 2. Clusters criados (k=30)

Após os clusters serem criados, os dados foram separados em 10 partições, ou *folds*, para realização da validação cruzada. Para cada iteração, 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.

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

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 os seguintes resultados foram obtidos:

Nº Teste	Taxa de acurácia para a Classe 1	Taxa de acurácia para a Classe 2	Taxa de acurácia Total
1	100%	100%	100%
2	100%	100%	100%
3	100%	100%	100%
4	100%	100%	100%
5	100%	100%	100%
6	100%	100%	100%
7	100%	100%	100%
8	100%	100%	100%
9	100%	100%	100%
10	100%	100%	100%
Média Geral	**100%**	**100%**	**100%**

Quadro 1. Comparação de acurácia para 30 clusters

O desvio padrão encontrado foi de 0.

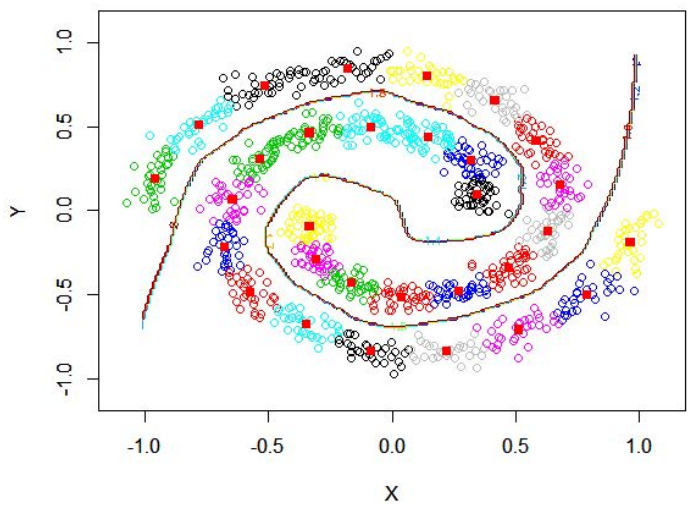

Figura 3. Superfície de contorno da solução encontrada usando 30 clusters

Como os resultados indicaram 100% de acerto em todos as tentativas, repetiu-se o experimento utilizando 16 clusters, ao invés de 30.

Nº Teste	Taxa de acurácia para a Classe 1	Taxa de acurácia para a Classe 2	Taxa de acurácia Total
1	88%	84%	86%
2	96%	76%	86%
3	92%	78%	85%
4	92%	84%	88%
5	88%	86%	87%
6	92%	88%	90%
7	80%	90%	85%
8	98%	84%	91%
9	86%	82%	84%
10	90%	82%	87%
Média Geral	**90%**	**84%**	**87%**

Quadro 2. Comparação de acurácia para 16 clusters

Nº Teste	Desvio padrão para a Classe 1	Desvio padrão para a Classe 2	Desvio padrão Total
1	0.3282607	0.3703280	0.5021167
2	0.1979487	0.4314191	0.4923660
3	0.2740475	0.4184520	0.4975699
4	0.2740475	0.3703280	0.5009083
5	0.3282607	0.3505098	0.5024184
6	0.2740475	0.3282607	0.5021167
7	0.4040610	0.3030458	0.5000000
8	0.1414214	0.3703280	0.4975699
9	0.3505098	0.3880879	0.5021167
10	0.3030458	0.3703280	0.5016136
Média Geral	**0.2875651**	**0.3701087**	**0.4998796**

Quadro 3. Comparação de desvio padrão para 16 clusters

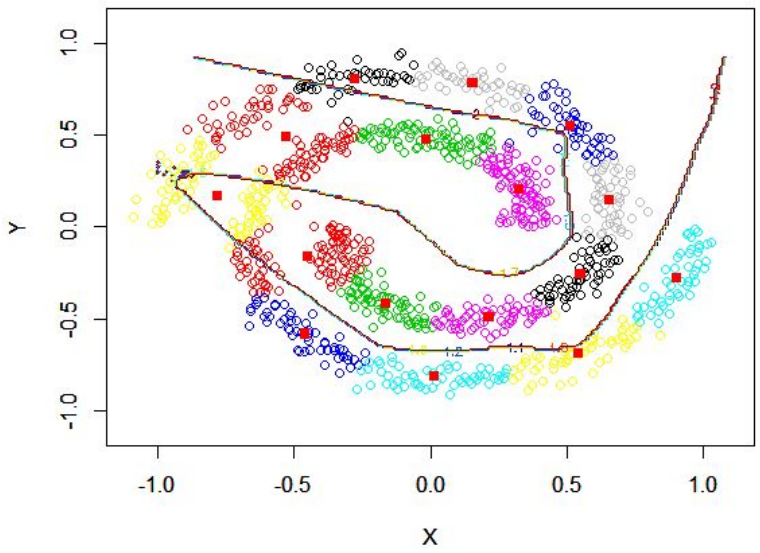

Figura 3. Superfície de contorno da solução encontrada usando 16 clusters

Ao analisar os resultados, conclui-se que o método da mistura de gaussianas foi efetivo em separar as duas classes do problema *Espiral*, porém nota-se que há uma dificuldade em se determinar qual é o número ideal de agrupamentos necessários para a solução do problema. Ao utilizar um número alto de agrupamento, houve maior custo computacional e classificou-se com acurácia de 100%, enquanto que usando quase a metade de clusters (16), houve uma acurácia total média 87%.

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. Abaixo os resultados encontrados usando 5 clusters.

Nº Teste	Taxa de acurácia para a Classe 1	Taxa de acurácia para a Classe 2	Taxa de acurácia Total
1	89%	100%	93%
2	96%	100%	97%
3	83%	100%	89%
4	96%	96%	96%
5	98%	100%	99%
6	91%	100%	94%
7	93%	100%	96%
8	96%	100%	97%

9	96%	100%	97%
10	98%	100%	99%
Média Geral	**93%**	**100%**	**96%**

Quadro 4. Comparação de acurácia para 5 clusters

Nº Teste	Desvio padrão para a Classe 1	Desvio padrão para a Classe 2	Desvio padrão Total
1	0,3146964	0	0,4974786
2	0,2061846	0	0,4866755
3	0,383223	0	0,5017567
4	0,2061846	0,2041241	0,4826171
5	0,147442	0	0,4826171
6	0,2848849	0	0,4934352
7	0,2496374	0	0,4902782
8	0,2061846	0	0,4866755
9	0,2084091	0	0,4881372
10	0,1490712	0	0,4841917
Média Geral	**0,2355918**	**0,0204124**	**0,4893863**

Quadro 5. Comparação de desvio padrão para 5 clusters

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.