Mistura de Gaussianas

ENG121 - Reconhecimento de Padrões

Autor: Matheus Castro Silva

RESUMO

O presente trabalho tem como objetivo introduzir um classificador utilizando o modelo de mistura de gaussianas para a base de dados Wisconsin Breast Cancer.

Palavras Chaves **- algoritmos, mistura de gaussianas, classificação, reconhecimento de padrões, bayes.**

I. INTRODUÇÃO

O presente trabalho busca introduzir um classificador utilizando o modelo de mistura de gaussianas para a base de dados Wisconsin Breast Cancer. E por fim identificar cada uma das amostras coletadas é do tipo maligno ou benigno.

II. DEFINIÇÃO DO PROBLEMA

Dado um conjunto da base de dados Wisconsin Breast Cancer identifique se a amostra é benigna ou maligna com base no classificador gerado utilizando mistura de gaussianas.

III. MODELO DE MISTURA DE GAUSSIANAS

Para o modelo de mistura de gaussianas, considere a equação de Bayes:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

Com esta equação é possível determinar a classe *Ck* de uma amostra *x*, de acordo com a probabilidade a priori de cada classe e da probabilidade a posteriori pertinência de cada amostra à pdf de cada classe. Como a probabilidade a priori das amostras é constante, ela pode ser desconsiderada no classificador, dessa forma o classificador é implementado utilizando a seguinte equação:

$$p(C_k|x) = p(C_k)p(x|C_k)$$

A mistura de gaussianas deve ser utilizadas para definir as probabilidades a posteriori de cada amostra condicional a cada classe, para isto deve ser gerada uma mistura para cada classe utilizando o conjunto de treinamento, podem ser utilizadas as funções desenvolvidas em sala (preferencialmente) ou o pacote *mclust.*

Para o cálculo das probabilidades a priori de cada classe, deve ser determinada a quantidade de ocorrências desta classe no conjunto de treinamento em relação à quantidade total de amostras.

IV. TREINAMENTO E TESTE

A base de dados deve ser dividida em dois conjuntos, um de treinamento e outro de teste, sendo 70% para treinamento e 30% para teste.

- Com o grupo de testes, deve ser utilizada a mistura de gaussianas para determinar um modelo para cada classe e a probabilidade a priori da cada classe.
- O grupo de treinamento deve ser classificado de acordo com os modelos estimados no treinamento. Dado que a $p(x|Ck)$ para cada classe pode ser estimada a partir dos modelos de misturas de gaussianas estimados no treinamento, assim como as probabilidades a priori $p(Ck)$. A classe que apresentar o maior probabilidade $p(Ck|x)$ deve ser considerada como a classe estimada para a amostra.

V. IMPLEMENTAÇÃO

1. Carregar os dados do pacote *mlbench* e substituir os dados faltantes por 0, por exemplo
2. Dividir de forma aleatória os dados em grupos de treinamento de teste de acordo com uma razão pré definida.
3. Utilizar a função de treinamento que estime os modelos de mistura de gaussianas e as probabilidades a priori. Utilizando o pacote *mclust,* a função para cada modelo de mistura de gaussianas é:

   modelo <- densityMclust(dadosTreinamento)

4. Utilizar uma rotina função de teste que verifique pertinência de cada amostra a cada mistura de gaussianas e determine, de acordo com a regra de Bayes, a qual classe cada amostra pertence.

   Prob <- dens(modelName=modelo$modelName, data = dadosTeste, parameters = modelo$parameters)

5. Calcular o erro quadrático médio (MSE) percentual do classificador.
6. Repetir 10 vezes os procedimentos 2 ao 5 e estimar o MSE percentual médio e o desvio padrão do classificador.

VI. RESULTADOS

Utilizando o passo a passo do tópico de implementação V neste trabalho os resultados obtidos, considerando n = 10 foram:

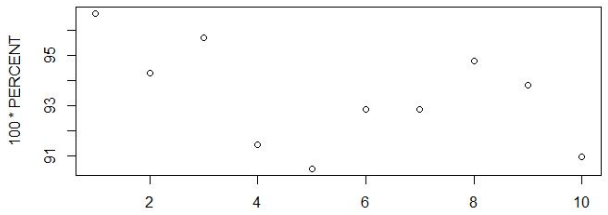

Imagem 1: Ilustração da Porcentagem de acerto x Iteração considerando n = 10.

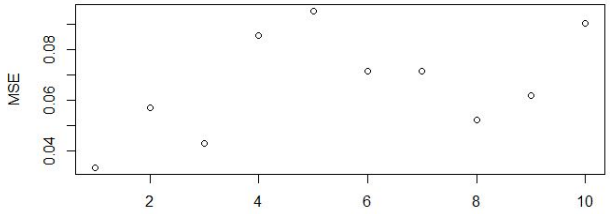

Imagem 2: Ilustração da curva MSE x Iteração considerando n = 10.

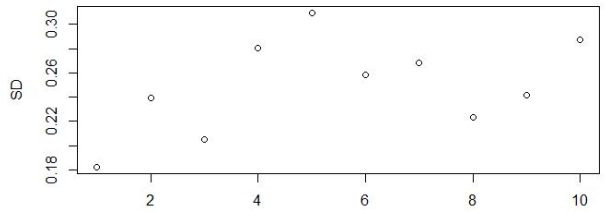

Imagem 3: Ilustração da curva Desvio padrão x Iteração considerando n = 10.

Utilizando também o passo a passo do tópico de implementação V neste trabalho os resultados obtidos, considerando a variação das iterações foram:

Iterações	**% Acerto**	**MSE**	**Desvio Padrão**
n = 1	93.80	0.062	0.247
n = 10	93.38	0.066	0.249
n = 25	93.12	0.068	0.252
n = 50	92.22	0.077	0.260
n = 100	92.77	0.722	0.258

O percentual de acerto perante os testes ficou entre 92.22% à 93.80%, considerando a técnica de mistura de gaussianas um método bastante eficiente porém que ainda tem pontos a melhorar. Também temos que levar em consideração que a base de dados era pequena e quanto maior o número de dados melhor seria a aproximação do nosso modelo neste problema em questão.

VII. CONCLUSÕES

O algoritmo de mistura de gaussianas foi bastante satisfatório para encontrar a separação dos tipos de câncer, mas também foi visto que o algoritmo nem sempre terá 100% de acerto, podendo apresentar resultados piores em alguns momentos mas todos acima de 92% de acerto.

Para obter uma melhora no algoritmo seria interessante aumentar o tamanho da base de dados de entrada para o modelo de treinamento e também outros fatores que este trabalho não irá abordar.

VIII. REFERÊNCIAS

[1] R tutorial, http://www.r-tutor.com/r-introduction, acessado 30/08/2017.

[2] Antônio de Pádua Braga - Notas de Aulas de Redes Neurais Artificiais e de Reconhecimento de Padrões, acessado 22/09/2017.