Classificação com Misturas de Gaussianas

Savio Lopes Rabelo

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Campus Fortaleza – CE – Brasil

savio.rabelo@ppgcc.ifce.edu.br

Resumo. *Este relatório descreve a implementação do classificador Bayesiano como a utilização de Mistura de Gaussianas, da classe de métodos semi-paramétricos. A metodologia utilizada para a implementação é constituída por duas fases: treinamento e teste, com cada conjunto sendo composto por 80% e 20% das bases de dados, respectivamente. Foram usadas quatro bases de dados disponíveis online no repositório UCI Machine Learning. Os resultados são bastante satisfatórios, chegando em taxas de acerto em 100% em algumas bases.*

1. Introdução

O modelo de Mistura Gaussiana (do inglês, *Gaussian Mixture Models*, GMM) é um método de estimação de densidade semi-paramétrica, que funde o mérito da estimativa de parâmetros e a estimativa não paramétrica, e não limita a forma específica da função de densidade de probabilidade. Além disso, a complexidade do modelo está relacionada apenas com problemas de solução e não tem nada a ver com o tamanho do conjunto de amostras [Fu and Wang 2012].

1.1. Mistura de Gaussianas

Dado uma variável aleatória X de dimensão de uma mistura com K componentes. A função de probabilidade de mistura de Gaussianas pode ser definida por:

$$\Phi(X|\Theta_k) = \sum_{k=1}^{K} \pi_k \phi(X|\theta_k), \tag{1}$$

onde cada θ_i corresponde ao conjunto de parâmetros definidos pela i-ésima componente da mistura, $\pi_i \in [0, 1]$ com $i \in (1, 2, \ldots, K)$ e $\sum_{i=1}^{K} \pi_i = 1$.

O vetor $\Theta_k = (\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k)$ é o conjunto dos parâmetros da mistura. Cada componente $\phi(X|\theta_i)$ da mistura é uma função de densidade de probabilidade Gaussiana definida por:

$$\phi(X = x|\theta_i) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \tag{2}$$

onde $X = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$ é o vetor de médias, Σ é uma matriz $d \times d$, Σ^{-1} é a inversa de Σ, $|\Sigma|$ é a determinante de Σ e $\theta_i = (\mu_i, \Sigma_i)$ representa os parâmetros de uma Gaussiana.

Especificamente, dado um conjunto de dados X com N instâncias x_i e dimensão d, os parâmetros de Θ_k são estimados maximizando o logaritmo da seguinte função de verossimilhança:

$$\begin{aligned} L(\Theta_k|X) &= log \prod_{i=1}^{N} \Phi(x_i|\Theta_k) \\ &= \sum_{i=1}^{N} log \sum_{k=1}^{K} \pi_k \phi(x_i|\theta_k). \end{aligned} \tag{3}$$

1.2. Expectation Maximization

O *Expectation Maximization* (EM) [Dempster et al. 1977] é um algoritmo capaz de encontrar um ótimo local da função de máxima verossimilhança de uma mistura de Gaussianas. O EM possui duas etapas iterativas:

E-Step: Nessa etapa ele calcula a probabilidade de cada ponto pertencer a cada Gaussiana. Além disso, é calculada uma nova estimativa da função de verossimilhança utilizando a Equação 4.

$$P(\pi_k, \theta_k|x_i) = \frac{\pi_k \phi(x_i, \theta_k)}{\sum_{i=1}^{N} \pi_i \phi(s_i, \theta_i)}. \tag{4}$$

M-Step: Nessa etapa, as componentes da mistura são maximizadas através das seguintes Equações:

$$\begin{aligned} \pi_k &= \frac{1}{N} \sum_{i=1}^{N} P(\pi_k, \theta_k|x_i), \\ \mu_k &= \frac{\sum_{i=1}^{N} P(\pi_k, \theta_k|x_i) x_i}{\sum_{i=1}^{N} P(\pi_k, \theta_k|x_i)}, \\ \Sigma_k &= \frac{\sum_{i=1}^{N} P(\pi_k, \theta_k|x_i)(x_i - \mu_k)^T (x_i - \mu_k)}{\sum_{i=1}^{N} P(\pi_k, \theta_k|x_i)}. \end{aligned} \tag{5}$$

2. Metodologia

No primeiro momento foi realizada a separação do conjunto de dados em dois subconjuntos: treinamento e teste. Os valores utilizados para os conjuntos equivalem a 80% do conjunto original para a fase de treinamento e 20% do conjunto original para a fase de teste. Logo depois, os dados foram normalizados para eliminação de redundâncias indesejadas e também foram embaralhados. Por ser um problema com mais de 2 classes, foi exigida uma codificação diferente, Um-versus-Todos (do inglês *One-v-All*, OvA).

Além disso, também foi utilizada a busca em grade com validação cruzada *k-fold*. A busca em grade é uma busca com o objetivo de encontrar os melhores parâmetros. Já o método de validação cruzada *k-fold* consiste em dividir o conjunto total de dados em *k* subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste e os *k-1* restantes são utilizados para estimação dos parâmetros e calcula-se a acurácia do modelo. Este processo é realizado *k* vezes alternando de forma circular o subconjunto de teste. A Figura 1 mostra o esquema realizado pelo *k-fold*. Ao final das *k* iterações calcula-se a acurácia sobre os erros encontrados, obtendo assim uma medida mais confiável sobre a capacidade do modelo de representar o processo gerador dos dados.

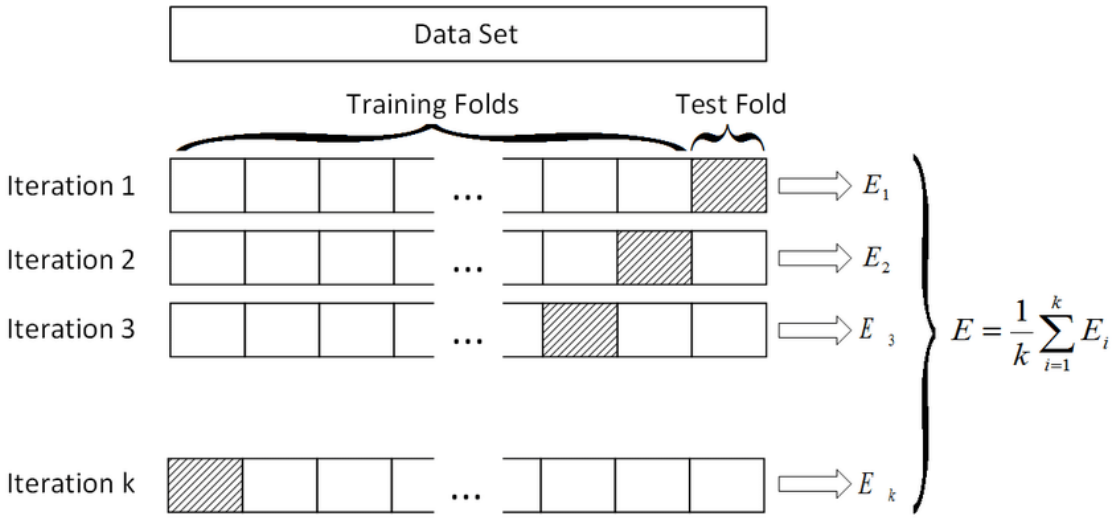

Figura 1. Método k-fold

Para a avaliação dos resultados alcançados na classificação, foram utilizados as seguintes métricas: a precisão ou valor preditivo positivo, taxa de sensibilidade ou taxa positiva verdadeira, especificidade ou taxa real negativa e acurácia. As Equações são apresentadas a seguir:

$$Precisao = \frac{VP}{VP + FP}, \tag{6}$$

$$Sensibilidade = \frac{VP}{VP + VN}, \tag{7}$$

$$Especificidade = \frac{VN}{N} = \frac{VN}{FP + VN}, \tag{8}$$

$$Acuracia = \frac{VP + VN}{P + N}, \tag{9}$$

onde P e N é o número de padrões de cada classe. VP é o verdadeiro positivo. VN é o verdadeiro negativo. FP é o falso positivo e FN é o falso negativo.

3. Conjuntos de Dados

. Para análise comparativas neste estudo, foram usados quatro conjuntos de dados: *Iris Flower Data Set*, *Vertebral Column Data Set*, *Dermatology Data Set* e *Breast Cancer Wisconsin Data Set*; todos disponíveis online no repositório *UCI Machine Learning* [Lichman 2013].

O banco de dados da Íris[1] é o conjunto mais conhecido que se encontra na literatura de reconhecimento de padrões. O conjunto de dados contém 3 classes de 50 instâncias cada, onde cada classe se refere a um tipo de planta de íris. Uma classe é linearmente separável das outras 2 classes.

Informações dos atributos:

1. Tamanho da sépala em cm
2. Largura da sépala em cm
3. Tamanho da pétala em cm
4. Largura da pétala em cm
5. Classe:
    (a) Iris Setosa
    (b) Iris Versicolour
    (c) Iris Virginica

Já o conjunto de dados da Coluna Vertebral[2] contém seis valores para características biomecânicas usadas para classificar pacientes ortopedistas em 3 classes (normal, hérnia de disco ou espondilolistese) ou 2 classes (normal ou anormal). Foi utilizado nessa prática o conjunto com três classes.

Informações dos atributos:

1. Incidência pélvica
2. Inclinação pélvica
3. Ângulo de lordose lombar
4. Inclinação sacra
5. Raio pélvico
6. Grau de espondilolistese
7. Classe:
    (a) Hérnia de Disco (DH)
    (b) Espondilolistese (SL)
    (c) Normal (NO)
    (d) Anormal (AB)

O banco de dados de Dermatologia[3] é constituído de 34 atributos. Esse banco é parte de um estudo que aponta o tipo de Eryhemato-Squamous Disease, uma doença de pele.

Informações dos atributos (valores de 0 a 3, exceto quando indicado):

1. Eritema
2. Escala

[1]Disponível em `https://archive.ics.uci.edu/ml/datasets/iris`
[2]Disponível em `http://archive.ics.uci.edu/ml/datasets/vertebral+column`
[3]Disponível em `http://archive.ics.uci.edu/ml/datasets/dermatology`

3. Fronteiras Definidas
4. Coceira
5. Fenômeno Koebner
6. Pápulas Poligonais
7. Pápulas Foliculares
8. Envolvimento da Mucosa Oral
9. Envolvimento no Joelho e no Cotovelo
10. Envolvimento do Couro Cabeludo
11. Histórico Familiar (0 ou 1)
12. Atributos Histopatológicos

    ⋮

33. Atributos Histopatológicos
34. Idade (Classe de 1 a 6)

E o banco de dados de Câncer de Mama[4] é constituído de 10 atributos. Informações dos atributos:

1. Número do código de amostra (número de identificação)
2. Clump Espessura (1 - 10)
3. Uniformidade do tamanho da célula (1 - 10)
4. Uniformidade da forma da Célula (1 - 10)
5. Adesão Marginal (1 a 10)
6. Tamanho Único de Células Epiteliais (1 - 10)
7. Núcleos Nus (1 - 10)
8. Cromatina Branda (1 a 10)
9. Nucleoli Normal (1 - 10)
10. Mitoses (1 - 10)
11. Classe (2 para benigno, 4 para maligno)

4. Simulações Computacionais

Para realizar os experimentos, foi utilizado um computador com a seguinte configuração: processador Intel(R) Core(TM) i7-6500U a 2.5 GHz com 8 GB de RAM e executando Windowns 10. Além disso, foi utilizado a linguagem de programação MATLAB. Todos os testes foram feitos com 50 realizações em cada base. O número de Gaussianas (K) variou de $[1\ 10]$, com $\Delta K = 1$.

A Tabela 1 mostra os resultados do classificador Bayesiano como a utilização de Mistura de Gaussianas em todas as bases de dados, levando em consideração as métricas já mencionadas.

[4]Disponível em `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\%28Original\%29`

Métricas (%)	Bases de Dados					
	Íris	Coluna (3C)	Coluna (2C)	Dermatologia	Câncer	Artificial
Acurácia	97,87	79,61	100,00	96,77	100,00	100,00
Taxa Mínima	90,00	69,35	100,00	89,25	100,00	100,00
Taxa Máxima	100,00	93,55	100,00	100,00	100,00	100,00
Desvio Padrão	02,84	05,36	00,00	03.97	00,00	00,00
Sensibilidade	97,95	78,93	100,00	93,95	100,00	100,00
Especificidade	98,93	90,29	100,00	97,58	100,00	100,00
Precisão	97,86	77,07	100,00	90.26	100,00	100,00

Tabela 1. Resultados do classificador Bayesiano como a utilização de Mistura de Gaussianas.

Na Figura 2 é apresentada a superfície de decisão com o classificador Bayesiano como a utilização de Mistura de Gaussianas.

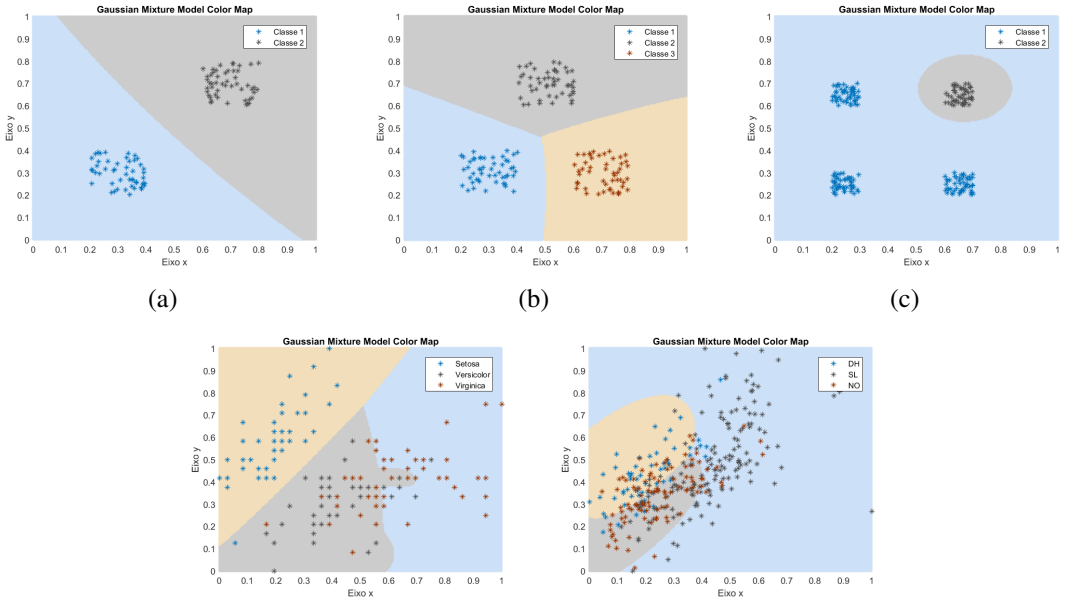

Figura 2. Superfície de decisão do classificador Bayesiano como a utilização de Mistura de Gaussianas. (a) Base artificial com 2 classes. (b) Base artificial com 3 classes. (c) Base artificial AND. (d) Íris com dois primeiros atributos. (e) Coluna com dois primeiros atributos.

A Figura 3 apresenta uma matriz de confusão de cada base de dados. Essa matriz é a matriz que ficou mais perto da acurácia.

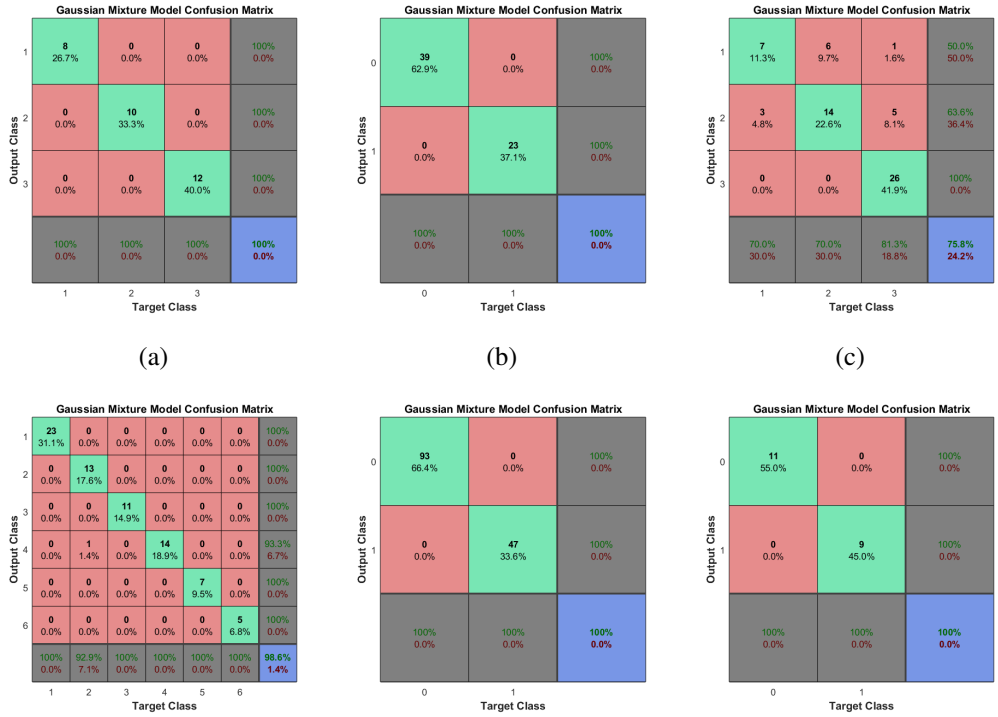

Figura 3. Matriz de Confusão para o classificador Naive Bayes. (a) Íris. (b) Coluna (2C). (c) Coluna (3C). (d) Dermatologia. (e) Câncer. (f) Artificial (2C).

5. Resultados

Misturas de Gaussianas são uma ferramenta muito poderosa e são amplamente utilizados em diversas tarefas que envolvem classificação e agrupamento de dados.

O que foi observado é que este tipo de solução não alterou de forma significante o desempenho de um classificador Bayesiano padrão. Uma das desvantagens é ter que estimar qual o melhor número de componentes da mistura. Embora a busca em linha com validação cruzada seja um método de avaliação bem definido e robusto, a falta de conhecimento prévio acerca do problema a ser solucionado é um fator que pesa na escolha do intervalo de busca do tamanho. Portanto, o uso de Mistura de Gaussianas para estimação de uma *Probability Density Function* (PDF) apresenta-se apenas como mais uma técnica útil que tem a seu favor uma flexibilização do classificador em termos de adaptação a problemas diversos.

Referências

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Fu, Z. and Wang, L. (2012). Color image segmentation using gaussian mixture model and em algorithm. In *International Conference on Multimedia and Signal Processing*, pages 61–66. Springer.

Lichman, M. (2013). UCI machine learning repository. `http://archive.ics.uci.edu/ml`. Acesso em março de 2019.