Aprendizagem de Máquina

Aprendizagem não-supervisionada II

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB Departamento de ESTATÍSTICA

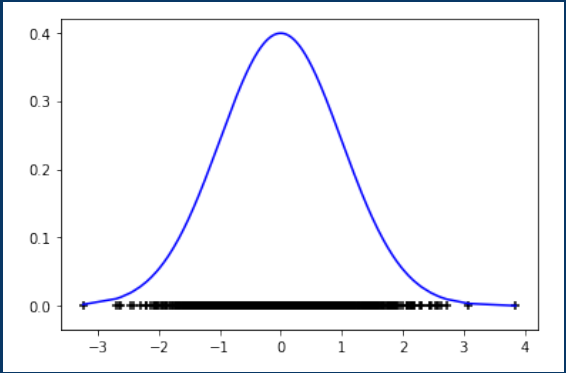
Sumário

Estimação de Densidade

- Frequentemente precisamos entender a distribuição dos nossos dados
- Seja para gerar novas amostras, para calcular probabilidades a posteriori em modelos baseados no Teorema de Bayes ou outras finalidades

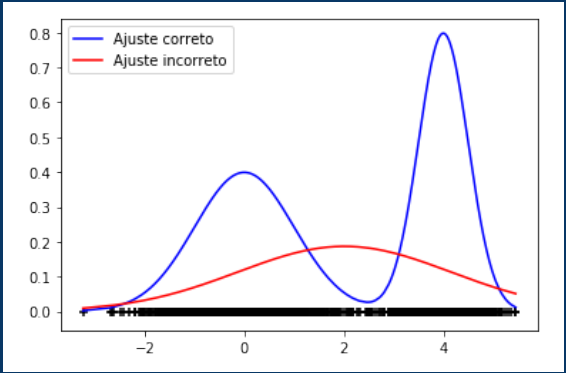
Estimação de Densidade

- No caso univariado, comumente assumimos que os dados seguem uma distribuição Normal ou Gaussiana, aí basta calcular a média e a variância/desvio-padrão a partir dos dados



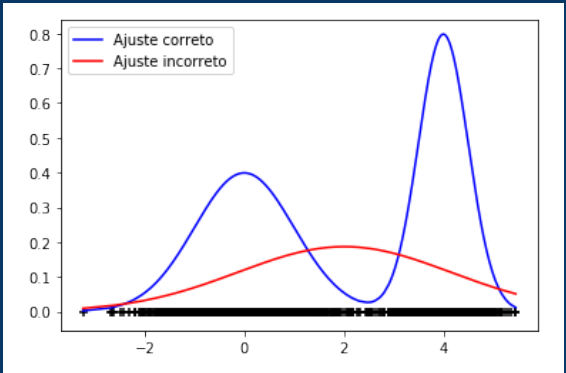
Estimação de Densidade

- No entanto, nem sempre uma Gaussiana é um bom ajuste para os dados



Estimação de Densidade

- No exemplo abaixo, os dados são na verdade modelados por uma mistura de duas Gaussianas diferentes $\mathcal{N}(0, 1)$ e $\mathcal{N}(4, 0.25)$



Expectation Maximization

- Como fazemos então para descobrir qual a mistura correta de Gaussianas para modelar nossos dados?
- Usamos um processo de maximização da verossimilhança

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k^2)$$

- Em que π_k é a probabilidade a priori do componente k da mistura e $\sum_{k=1}^{K} \pi_k = 1$

Expectation Maximization

- O algoritmo Expectation Maximization (EM) é um processo iterativo que encontra os parâmetros da mistura para maximizar a verossimilhança dos dados sob o modelo de mistura de distribuições
- Ele lembra os passos do algoritmo K-médias

Expectation Maximization

- No primeiro passo (passo E), calculamos para toda observação i

$$r_{ik} = \frac{\pi_k \mathcal{N}(x_i \mu_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \sigma_j^2)}$$

- No segundo passo (passo M), atualizamos os parâmetros dos componentes da mistura

$$\pi_k = \frac{\sum_{i=1}^{N} r_{ik}}{N}$$

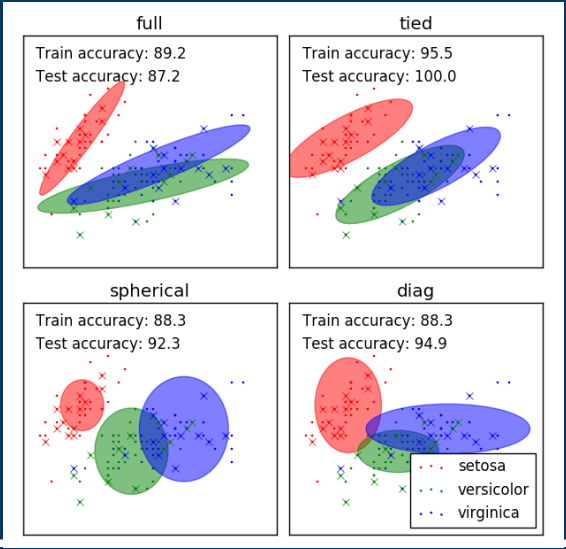
$$\mu_k = \frac{\sum_{i=1}^{N} r_{ik} x_i}{\sum_{i=1}^{N} r_{ik}}$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{N} r_{ik} (x_i - \mu_k)^2}{\sum_{i=1}^{N} r_{ik}}$$

Expectation Maximization

- O processo se repete até convergir para um mínimo local
- O mesmo processo funciona também para normais multivariadas
  - Nesse caso, cada normal é modelada por um vetor de médias e uma matriz de covariâncias
- Assim como o K-médias, é preciso definir um número K de componentes e um ponto de partida para a busca dos parâmetros
- O algoritmo é usado também para a tarefa de agrupamento
  - Diferente do K-médias o algoritmo de mistura de Gaussianas por EM consegue encontrar clusters com formas variadas (elípticas)
  - Além disso, as observações não pertencem a grupos exclusivos

Expectation Maximization



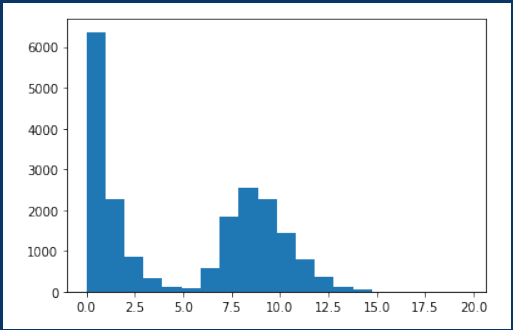
Estimação de Densidade não-Paramétrica

Estimação de Densidade não-Paramétrica

- Na metodologia clássica, faz-se alguma suposição sobre a forma funcional paramétrica dos dados
- Com uma forma paramétrica imposta, tudo que resta é estimar os parâmetros através dos dados (Máxima verossimilhança, por exemplo)
- Muitas vezes, a suposição acerca da forma funcional paramétrica pode ser muito restritiva ou, em alguns casos, inadequada
- Abordagens não-paramétricas permitem que a gente lide com um número maior de situações

Histograma

- Método não-paramétrico mais antigo de estimação de densidades
- Dada uma origem x_0 e um comprimento de intervalo h, definimos os retângulos do histograma como sendo os intervalos $[x_0 + (r-1)h, x_0 + rh)$ para valores inteiros positivos e negativos de r
- Empiricamente a ideia é contar o número de observações que estão contidas em cada intervalo

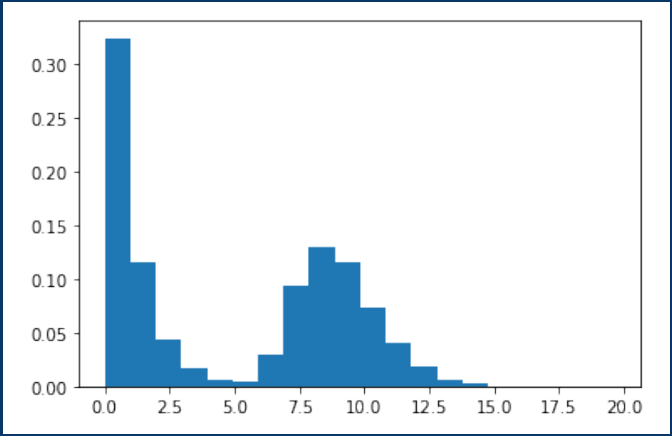


Histograma

- Sem perda de generalidade, seja o intervalo $[-h/2, h/2)$. A probabilidade de uma observação pertencer ao intervalo $[-h/2, h/2)$ é dada por $P(X \in [-h/2, h/2)) = \int_{-h/2}^{h/2} f(x)dx$, onde f é a densidade de X
- Uma aproximação natural para a probabilidade acima é $P(X \in [-h/2, h/2)) \approx \frac{1}{N}\#\{x_i \in [-h/2, h/2)\}$
- Dessa forma, uma estimativa para f seria

$$\hat{f}(x) = \frac{1}{Nh}\#\{x_i \in [-h/2, h/2)\}, \qquad \forall x \in [-h/2, h/2)$$

Histograma



Histograma

- Este estimador não é contínuo e depende fortemente da escolha de h, conhecido como parâmetro de suavização
- Variando o valor de h obtemos diferentes formas de $\hat{f}_h(x)$. Nos extremos, digamos, quando $h \to 0$, temos uma representação muito ruidosa dos dados. Na situação oposta, quando $h \to \infty$, temos uma representação muito suave dos dados

Histograma

- A ideia do histograma serve como base para um estimador de densidades mais geral conhecido como estimador *naive* [Silverman (1986)]. Seja X uma v.a. com densidade f. Então,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

- Para h fixo, podemos estimar $P(x - h < X < x + h)$ pela proporção de observações da amostra pertencentes ao intervalo $(x - h, x + h)$. Desse modo, um estimador natural de f, escolhendo h pequeno, é

$$\hat{f}(x) = \frac{1}{2Nh} \#\{x_i \in (x - h, x + h)\}$$

Histograma

- Para expressar este estimador de uma forma que será útil mais à frente, seja a função peso w:

$$w(z) = \begin{cases} \frac{1}{2} & \text{se } |z| < 1 \\ 0 & \text{caso contrário.} \end{cases} \tag{1}$$

- Então, uma estimativa para f neste caso é dada por

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h} w\left(\frac{x - x_i}{h}\right) \tag{2}$$

- A partir de (1) podemos notar que o estimador (2) é construído colocando-se um retângulo de largura $2h$ e altura $(2Nh)^{-1}$ em cada observação e então somando para obter a estimativa $\hat{f}$

Estimação de Densidades Univariadas pelo Método Kernel

- $\hat{f}$ não é uma função contínua e tem derivada nula em todos os pontos exceto nos pontos de salto $x \pm h$
- O estimador de densidades baseado em uma função kernel é obtido substituindo a função peso w por uma função não-negativa K, denominada função kernel, satisfazendo a condição $\int_{-\infty}^{\infty} K(x)dx = 1$
- Usualmente, mas não sempre, K será uma função densidade de probabilidade simétrica (Por exemplo, a função densidade de probabilidade normal)

Estimação de Densidades Univariadas pelo Método Kernel

- No caso univariado o estimador kernel para uma amostra aleatória $x_1, \ldots, x_N$ retirada de uma distribuição com densidade comum f, pode ser definido como

$$\hat{f}(x; h) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) = \frac{1}{N} \sum_{i=1}^{N} K_h(x - x_i), \tag{3}$$

  onde h é o parâmetro de suavização, positivo e não-aleatório, e K é a função kernel, não-negativa, satisfazendo a condição $\int_{-\infty}^{+\infty} K(x)dx = 1$
- A relação entre K e K_h é dada por $K_h(t) = h^{-1}K(h^{-1}t)$

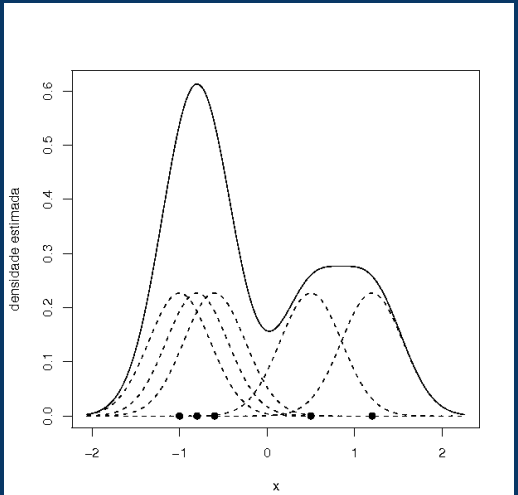
Estimação de Densidades Univariadas pelo Método Kernel

- Em cada ponto, uma função kernel dimensionada K_h com massa de probabilidade n^{-1} é colocada. Estas são então somadas para fornecer a curva composta
- A escolha da função kernel não é crucial para a performance do método, e é mais razoável escolher um kernel que auxilie na eficiência computacional [Silverman (1986), Epanechnikov (1969)]

Tabela: Funções kernel comumente utilizadas com dados univariados

Função kernel	Forma analática, $K(x)$
Retangular	$\frac{1}{2}$ para $\lvert x\rvert < 1$, 0 caso contrário
Triangular	$1 - \lvert x\rvert$ para $\lvert x\rvert < 1$, 0 caso contrário
Biweight	$\frac{15}{16}(1 - x^2)^2$ para $\lvert x\rvert < 1$, 0 caso contrário
Normal	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$
Epanechnikov	$\frac{3}{4}\left(1 - x^2/5\right)/\sqrt{5}$ para $\lvert x\rvert < \sqrt{5}$, 0 caso contrário

Figura: Linha sólida: densidade estimada; Linhas tracejadas: funções kernel individuais. A amostra é composta pelos valores $x_1 = -1.0, x_2 = -0.8, x_3 = -0.6, x_4 = 0.5, x_5 = 1.2$



Estimação de Densidades Multivariadas pelo Método Kernel

- A extensão para dados multivariados é direta, com o estimador de densidades p-dimensional, para uma amostra aleatória $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ retirada de uma densidade comum f, definido por

$$\hat{f}(\mathbf{x}) = \frac{1}{Nh^p} \sum_{i=1}^{N} K\left(\frac{1}{h}(\mathbf{x} - \mathbf{x}_i)\right), \tag{4}$$

  onde $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$ e $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, $i = 1, 2, \ldots, N$
- A função kernel multivariada $K(\mathbf{x})$ é agora uma função definida no espaço p-dimensional, satisfazendo $\int_{\mathbb{R}^p} K(\mathbf{x}) d\mathbf{x} = 1$
- Usualmente K será uma função densidade de probabilidade unimodal radialmente simétrica

Exemplos de funções kernel multivariadas são a distribuição normal padrão multivariada

$$K(\mathbf{x}) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right),$$

e a função kernel Bartlett-Epanechnikov

$$K(\mathbf{x}) = \begin{cases} \frac{(1-\mathbf{x}^T\mathbf{x})(p+2)}{2c_p} & \text{para } |\mathbf{x}| < 1 \\ 0 & \text{caso contrário,} \end{cases}$$

onde

$$c_p = \frac{\pi^{p/2}}{\Gamma((p/2)+1)}$$

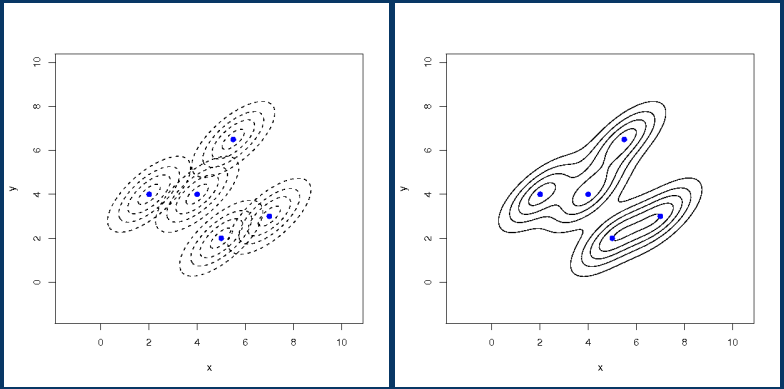
é o volume de uma esfera unitária p-dimensional

- O uso de um único parâmetro de suavização em (4) implica que a função kernel colocada em cada ponto é dimensionada igualmente em todas as direções e isso pode ser inadequado em muitas situações
- Uma forma da estimativa da função de densidade de probabilidade comumente utilizada é a soma do produto de funções kernel (sem, contudo, a implicação de independência entre as variáveis)

$$\hat{f}(\mathbf{x}) = \frac{1}{N}\frac{1}{h_1 \cdots h_p}\sum_{i=1}^{N}\prod_{j=1}^{p} K_j\left(\frac{x_j - x_{ij}}{h_j}\right), \tag{5}$$

onde existem diferentes parâmetros de suavização associados com cada variável. Pode-se assumir algum kernel univariado para os K_j, $j = 1, \ldots, p$. Usualmente, a mesma forma é assumida para todos os K_j.

Figura: Estimativa da densidade bivariada pelo método kernel. Linha sólida: curvas de nível da densidade estimada; Linhas tracejadas: curvas de nível das funções kernel individuais; Amostra: $\mathbf{x}_1 = (7, 3)$, $\mathbf{x}_2 = (2, 4)$, $\mathbf{x}_3 = (4, 4)$, $\mathbf{x}_4 = (5, 2)$, $\mathbf{x}_5 = (5.5, 6.5)$;



Classificação Usando Estimação de Densidade

Classificação Usando Estimação de Densidade

- É possível usar estimadores de densidade para a tarefa de classificação
- Para isso, dado um conjunto de treinamento, estimamos separadamente a densidade $\hat{f}_c$ dos dados que pertencem a cada classe $c = 1, \ldots, C$ e a priori da classe c (π_c)
- $\hat{f}_c$ pode ser estimada de forma paramétrica, por exemplo usando uma mistura de Gaussianas, ou de forma não-paramétrica
- Com isso, podemos usar o Teorema de Bayes:

$$P(Y = c|\mathbf{x}) = \frac{\hat{f}_c(\mathbf{x})\pi_c}{\sum_{g=1}^{C} \hat{f}_g(\mathbf{x})\pi_g}$$

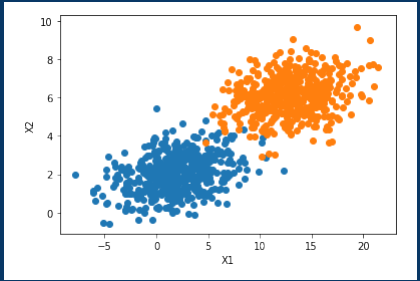
Análise de Componentes Principais

Análise de Componentes Principais

- Dado um conjunto de dados, será que todas as variáveis são necessárias para tomar uma decisão/reconhecer padrões?
- Será que podemos transformar os dados de forma que precisemos de menos variáveis para tomar nossas decisões?

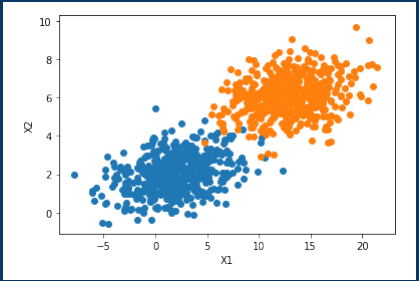
Análise de Componentes Principais

- Neste gráfico, temos dois grupos já demarcados



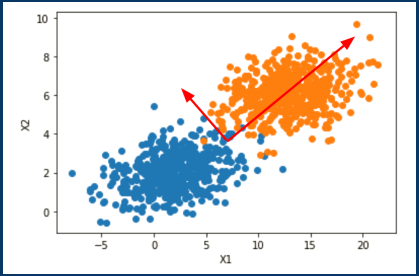
Análise de Componentes Principais

- O grupo laranja se localiza mais à direita e mais acima em relação ao grupo azul



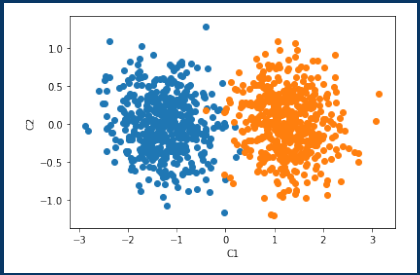
Análise de Componentes Principais

E se a gente pudesse mexer nos eixos do gráfico, de forma que o conjunto de dados não estivesse na diagonal?



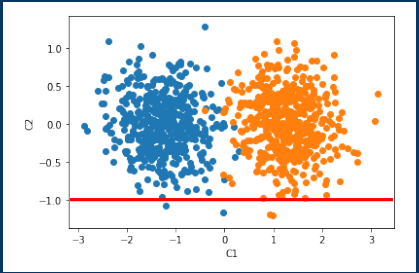
Análise de Componentes Principais

- Os novos eixos são chamados de componentes (C1 e C2)



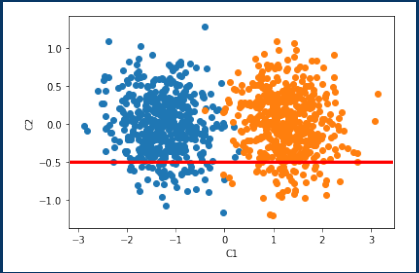
Análise de Componentes Principais

- Olhando para os componentes, os grupos não parecem separáveis de acordo com C2



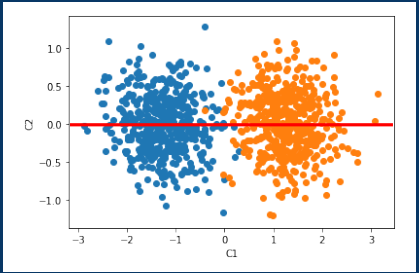
Análise de Componentes Principais

- Olhando para os componentes, os grupos não parecem separáveis de acordo com C2



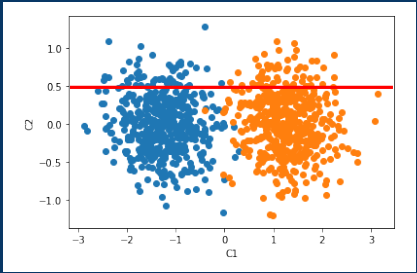
Análise de Componentes Principais

- Olhando para os componentes, os grupos não parecem separáveis de acordo com C2



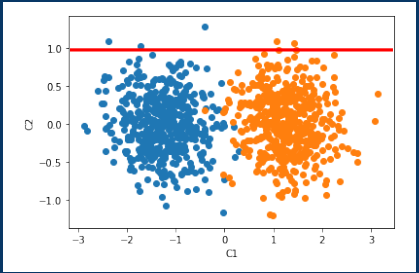
Análise de Componentes Principais

- Olhando para os componentes, os grupos não parecem separáveis de acordo com C2



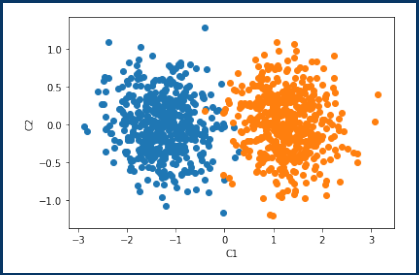
Análise de Componentes Principais

- Olhando para os componentes, os grupos não parecem separáveis de acordo com C2



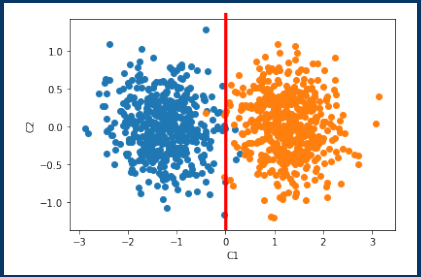
Análise de Componentes Principais

- O componente C1 por outro lado, oferece uma boa separação para os grupos



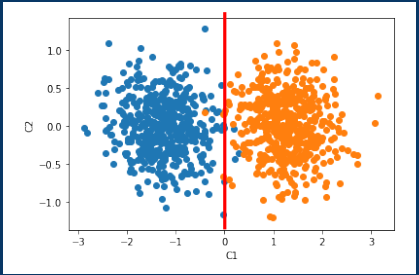
Análise de Componentes Principais

- O componente C1 por outro lado, oferece uma boa separação para os grupos



Análise de Componentes Principais

C1 é chamado de primeiro componente principal e consegue explicar quase toda a variância (97%) do conjunto



Análise de Componentes Principais

- A análise de componentes principais (PCA - Principal Component Analysis) consegue encontrar os novos eixos, chamados de componentes, automaticamente
- A partir daí, podemos usar apenas os componentes que explicam uma fatia maior da variância do nosso conjunto nas nossas análises
- **Observação:** é importante normalizar os dados antes de aplicar o PCA, para que ele capture as variâncias corretamente em seus componentes

Para Terminar

- O uso de classificadores baseados em estimação de densidade (modelos generativos) é o melhor que podemos fazer se o ajuste das densidades for perfeito
- Na prática, isso nunca acontece:
  - Modelos paramétricos podem não se ajustar à densidade real dos dados
  - Modelos não-paramétricos normalmente necessitam de mais dados para aproximar a densidade real
- Os modelos que tentam maximizar a acurácia sem se preocupar com as distribuições dos dados (modelos discriminativos) frequentemente se saem melhor na prática, pois fazem um melhor ajuste da fronteira de decisão

Para Terminar

- Ao usar PCA no seu experimento, ele deve ser aplicado dentro da validação cruzada, ou seja para cada fold de teste:
  1. PCA é ajustado ao conjunto de treinamento
  2. O conjunto transformado é alimentado ao algoritmo de aprendizagem
  3. O PCA ajustado é usado pra transformar o conjunto de teste
  4. O conjunto de teste transformado é alimentado ao algorimo de aprendizagem para avaliação

Sugestão de Atividade

- Use os scripts de geração de conjuntos de dados, ajuste modelos de estimação de densidade (mistura de Gaussianas com diferentes números de componentes e KDE) para as classes e use o Teorema de Bayes para implementar um classificador
- Faça isso em uma validação cruzada
- Use essa estrutura experimental com um conjunto real adicionando o PCA

FIM

Aprendizagem de Máquina

Aprendizagem não-supervisionada II

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB Departamento de ESTATÍSTICA