Instituto Federal Sul Riograndense
Curso Superior em Engenharia Elétrica
Aprendizado de Máquina

# Análise em Distribuições Gaussianas

Oscar Schmitt Kremer.

Professor Ms. Lucian Schiavon

Pelotas
2019

# Conteúdo

# 1 Objetivo

Estudo e implementação em MATLab de um discriminante Gaussiano multivariado, estudando conceitos sobre probabilidade como densidade de probabilidade, matriz de covariância, limiar de decisão e distribuições Gaussianas.

# 2 Fundamentação Teórica

Ao modelar um determinado conjunto de dados como uma distribuição normal multivariada define-se como o comportamento da função de densidade de probabilidade o mostrado em (1) para o caso de uma distribuição de probabilidade monovariada.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left( - \frac{(x-\mu)^2}{2\sigma^2} \right) \tag{1}$$

A generalização para o caso multivariado leva à representação (2), onde a matriz $\Sigma$ representa a matriz de covariância entre as variáveis aleatórias presentes, possuindo assim dimensão $k \times k$, dependendo diretamente da dimensão da entrada.

$$f(\mathbf{x}) = \frac{1}{\sqrt{2\pi^k|\Sigma|}} \exp\left( - \frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) \right) \tag{2}$$

Para o caso multivariável a média também assumirá comportamento vetorial, a matriz de covariância por sua vez para ser encontrada necessita que seja levada em consideração a média dos dados de entrada de acordo com a saída gerada pelos mesmos. Para encontrar a matriz de covariância de cada classe é necessário que seja levada em conta a a influência de cada uma das dimensões dos dados de entrada.

$$c_{ij} = \frac{1}{m}\sum_{k=0}^{m}(x^{i,k} - \mu^i)(x^{j,k} - \mu^j) \tag{3}$$

Onde a matriz de covariância formada no final será definida como para cada classe como sendo no formato (4).

$$\Sigma = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \tag{4}$$

No caso de um sistema onde os dados possam ser classificados em múltiplas classes, a matriz de covariância total é definida como a soma das matrizes de covariâncias, tornando assim com que a matriz de covariância para ambos os dados seja a mesma. O limiar de decisão pode ser encontrado por sua vez com a análise do gráfico dos contornos, onde um ponto da reta encontra-se no ponto médio dos centroides das distribuições e a inclinação da reta é definida como sendo $\mathbf{w}$, logo a partir da inclinação e de um ponto encontra-se a

reta que define o limiar de decisão.

$$\mathbf{w} = \begin{bmatrix} -\frac{\mu_1^2 - \mu_2^2}{\sigma_2^2} \\ \frac{\mu_1^1 - \mu_2^1}{\sigma_1^2} \end{bmatrix} \tag{5}$$

# 3 Implementação

Para implementação primeiramente fora utilizado uma função para geração do gráfico tridimensional de cada distribuição gaussiana, os gráficos para cada um dos conjuntos pode ser visualizado na Figura 1. Os contornos da mesma, ou seja, suas curvas de nível.

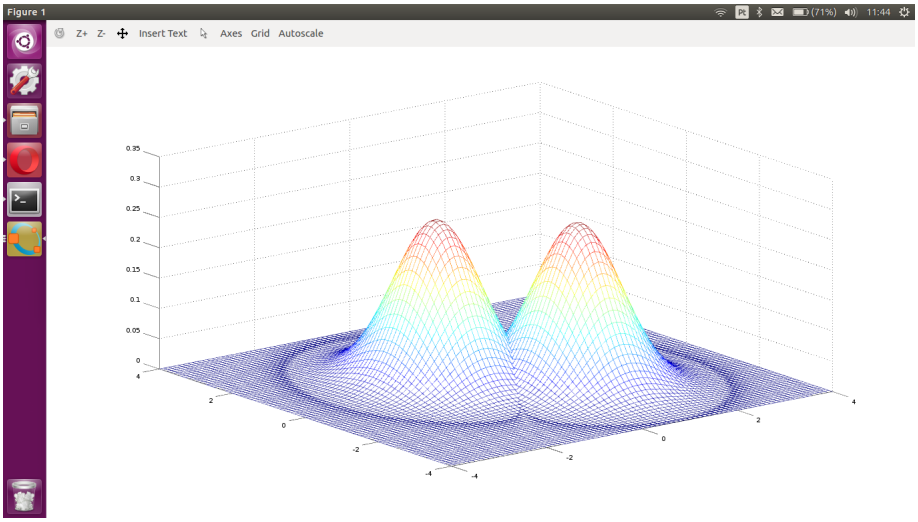


Figura 1: Distribuição em Representação Tridimensional

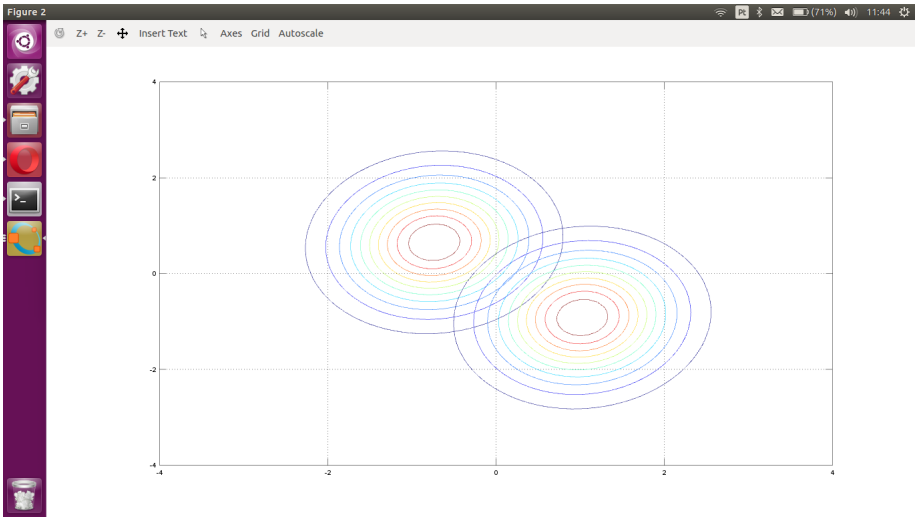


Figura 2: Contornos X-Y

Analisando os dados originais separados pelos valores fornecidos no próprio dataset a separação pode ser visualizada na Figura 3, onde em verde estão representadas os membros da classe cuja saída é 0 e em preto a que possui saída 1. Neste caso tratado os dados para treino e testes são os mesmos, ao contrário da boa prática que recomenda a utilização de no mínimo dos conjuntos para treino e teste ou até mesmo três, treino, teste e validação, porém tratando apenas de um caso didático este fato é relevado.

Na figura abaixo pode ser visualizado o limite de decisão do sistema considerado, o limite de decisão possui sua orientação definida no caso de modo a ser ortogonal à um vetor

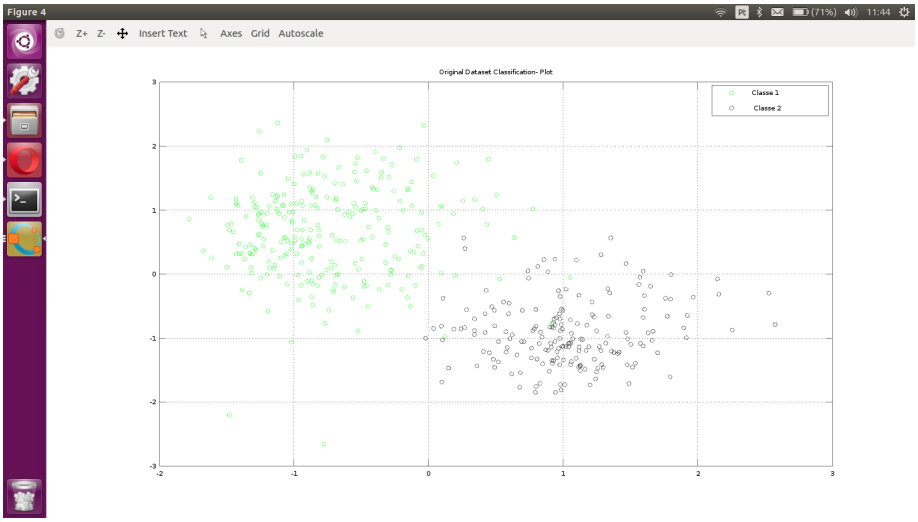


Figura 3: Dados de Treinamento

**v**. Nos resultados observados torna-se claro que a utilização da metodologia apresentada para a geração do limiar de decisão possui limitações, podendo ser ampliado para que não definido de forma linear e sim curvilínea.

$$\mathbf{v} = \Sigma^{-1}(\mu_1 - \mu_2) \tag{6}$$

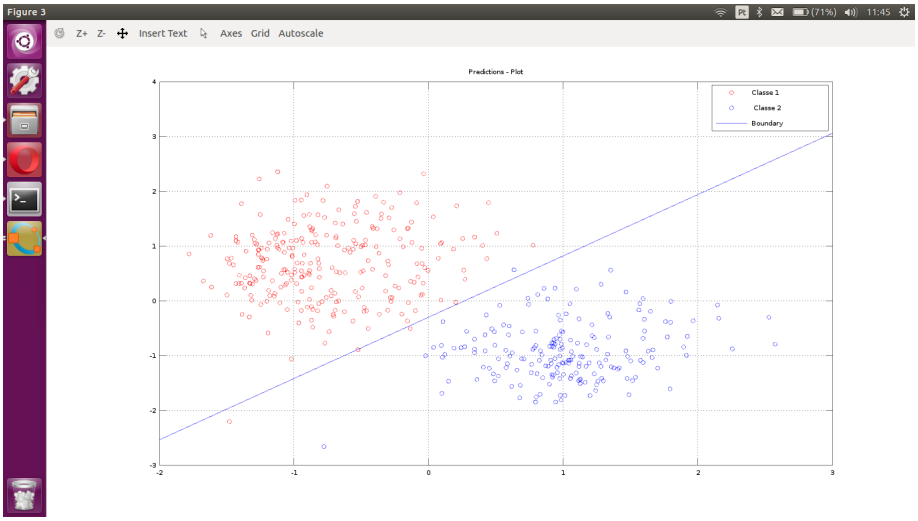


Figura 4: Contornos X-Y

# 4 Conclusão

O sistema mostrou uma grande taxa de acerto e grande capacidade para classificação. O limiar gerado assume comportamento linear, mas para trabalhos futuros pode ser generalizado para curvas com outros comportamentos. Outras implementações podem incluir aumento do número de dimensões, classes e geração de limiares com comportamentos cônicos.