Detecção de Anomalias

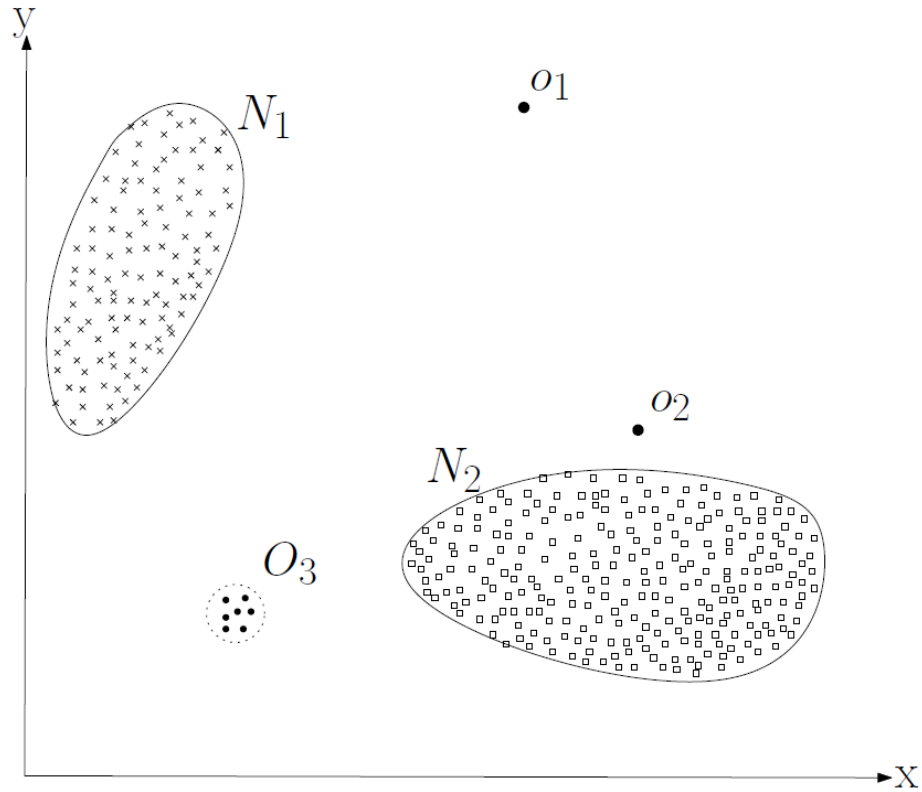
Prof. Danilo Silva

EEL7514/EEL7513 - Tópico Avançado em Processamento de Sinais
EEL410250 - Aprendizado de Máquina

EEL / CTC / UFSC

Introdução

Detecção de Anomalias (ou Outliers)

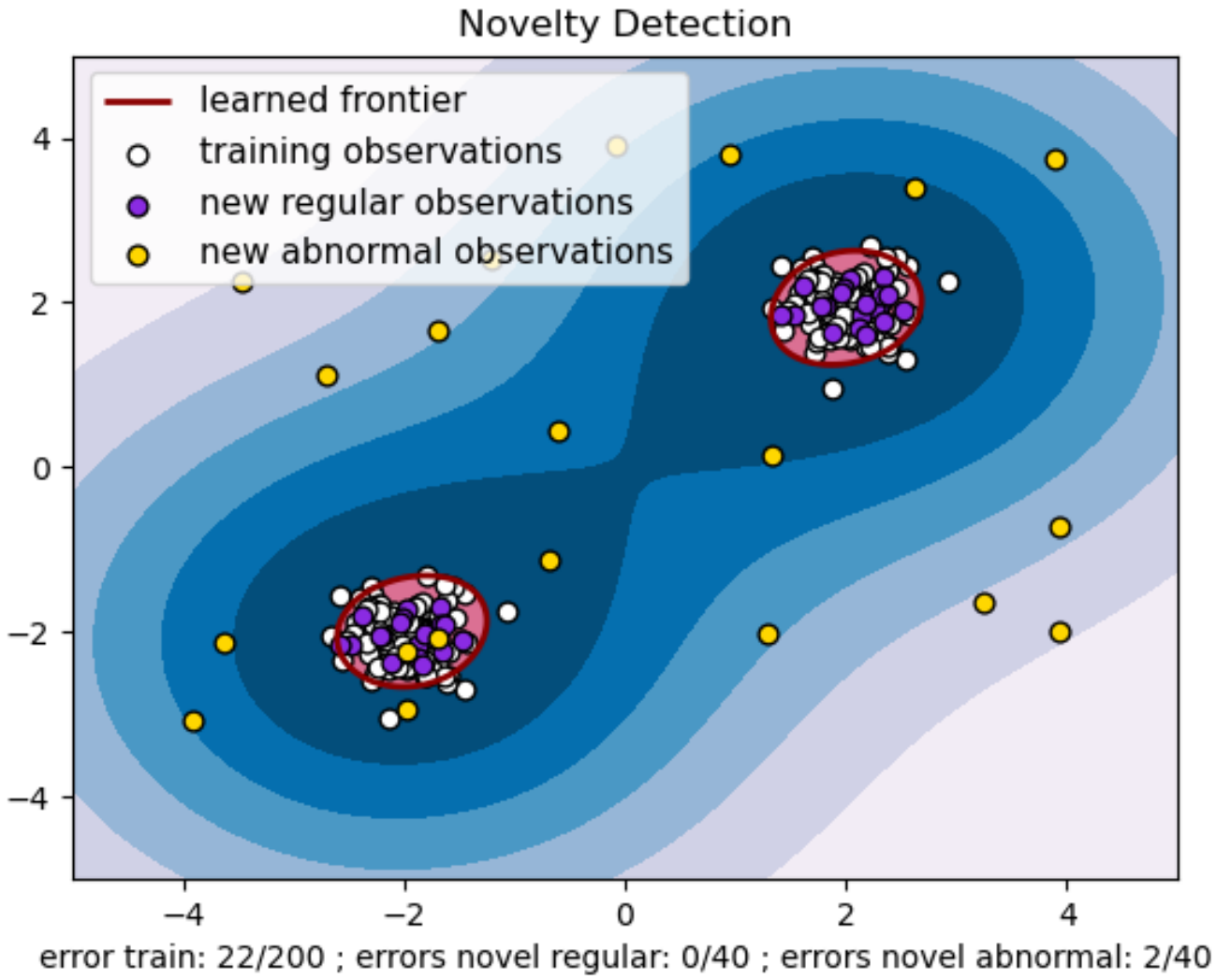


- Objetivo: detectar casos que fogem ao “normal” (comum / esperado)
- Aplicações:
  - Detecção de intrusos / atividade maliciosa / fraudes
  - Detecção de falhas em sistemas
  - Monitoramento de saúde

Detecção de Anomalias (ou Outliers)

- Tipos de aprendizado:
  - Supervisionado: Requer amostras rotuladas como normais e anômalas
    - Equivalente a classificação binária com alto desbalanceamento
  - Semi-Supervisionado: Requer apenas amostras normais
    - Também conhecido como detecção de novidade
  - Não-Supervisionado: Requer apenas amostras não-rotuladas
    - Assume-se que a proporção de anomalias é pequena
- Abordagens:
  - Baseada em classificação (ex: SVM de única classe)
  - Vizinhança/clustering
  - Redução de dimensionalidade (ex: PCA, redes neurais *autoencoders*)
  - Modelamento estatístico (estimação de densidade de probabilidade)

One-Class SVM



Estimação de Densidade

- Princípio básico: uma amostra é classificada como anômala se possui baixa probabilidade de ocorrência (abaixo de um limiar pré-estabelecido)

$$p(\mathbf{x}) < \epsilon$$

Tipos:

- Estimação paramétrica: assume um modelo específico caracterizado por uma densidade de probabilidade com parâmetros livres a serem ajustados pelos dados (ex: modelo gaussiano)
- Estimação não-paramétrica: não faz hipóteses sobre o modelo, ao invés disso determina a densidade a partir dos dados (ex: histograma)

Modelos Gaussianos

Modelo Gaussiano (Univariável)

- $n = 1 \implies \mathbf{x} = (x_1) = x \in \mathbb{R}$
- Densidade de probabilidade ($x \sim \mathcal{N}(\mu, \sigma^2)$):

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

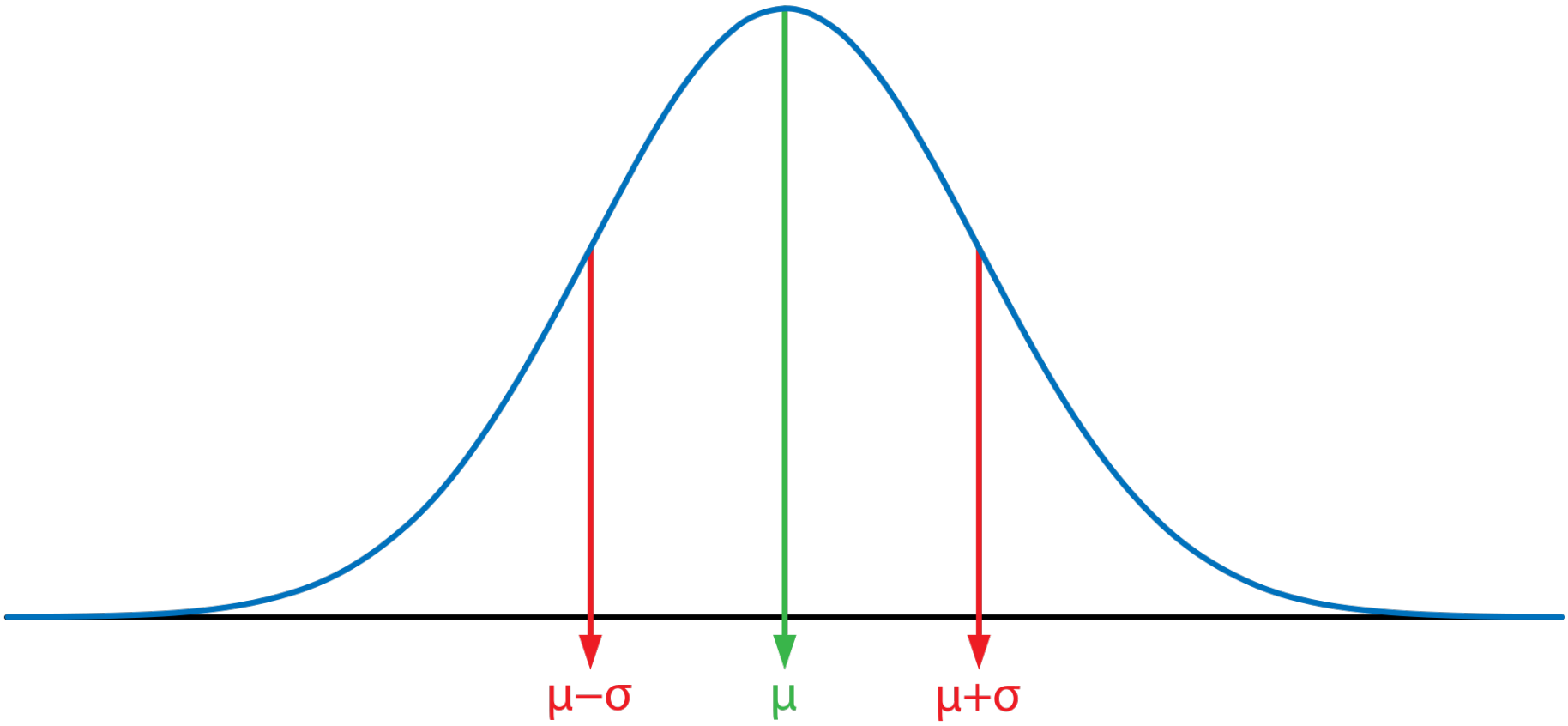
- Estimação de máxima verossimilhança:

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu)^2$$

  - Estimação não-enviesada de σ^2: divida por $m - 1$ ao invés de m
    - Obs: irrelevante para m suficientemente grande

Exemplo



Modelo Gaussiano Multivariável

- Densidade de probabilidade ($\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$):

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

- Estimação de máxima verossimilhança:

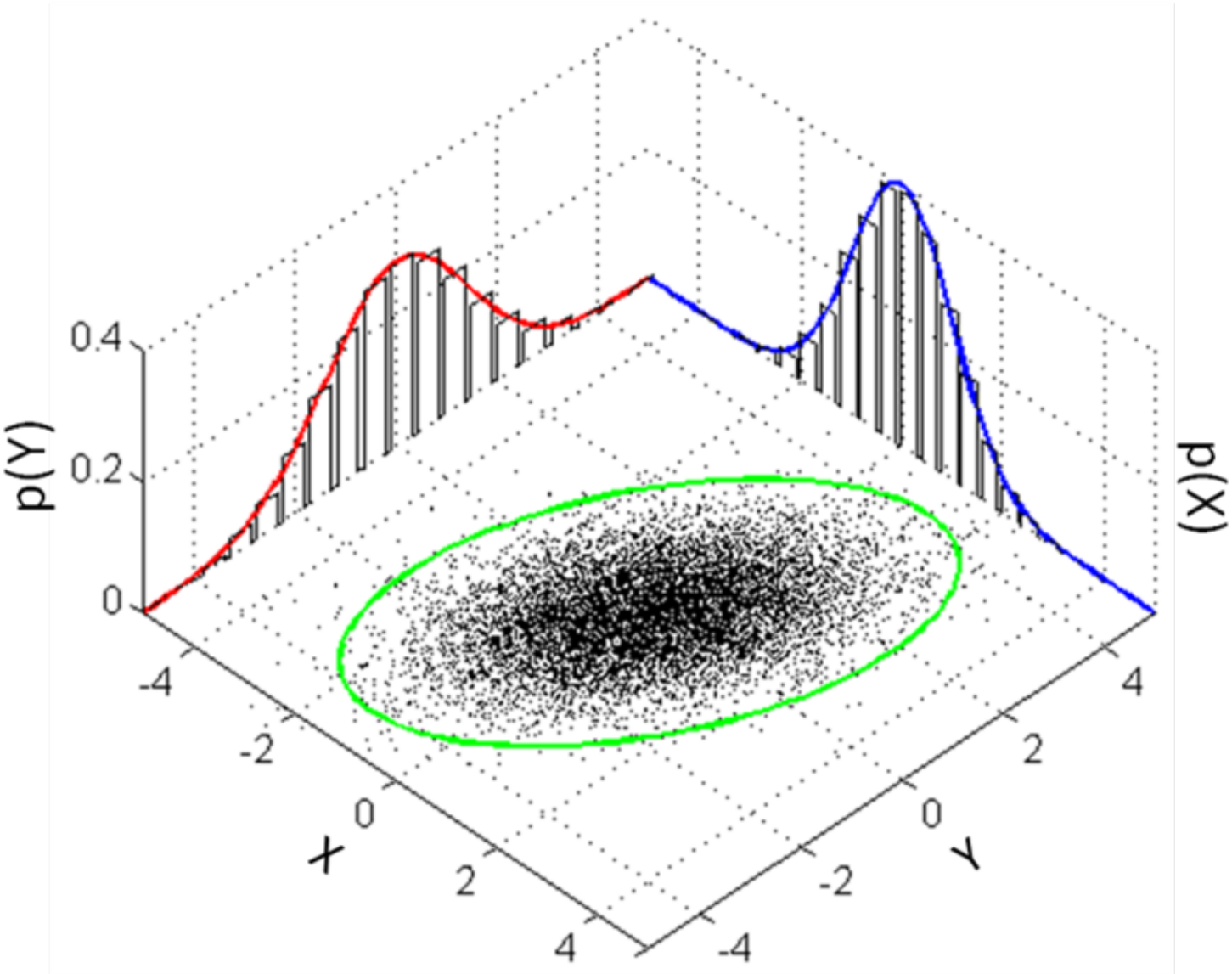
$$\boldsymbol{\mu} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}^{(i)}$$

$$\boldsymbol{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}^{(i)} - \boldsymbol{\mu})(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T$$

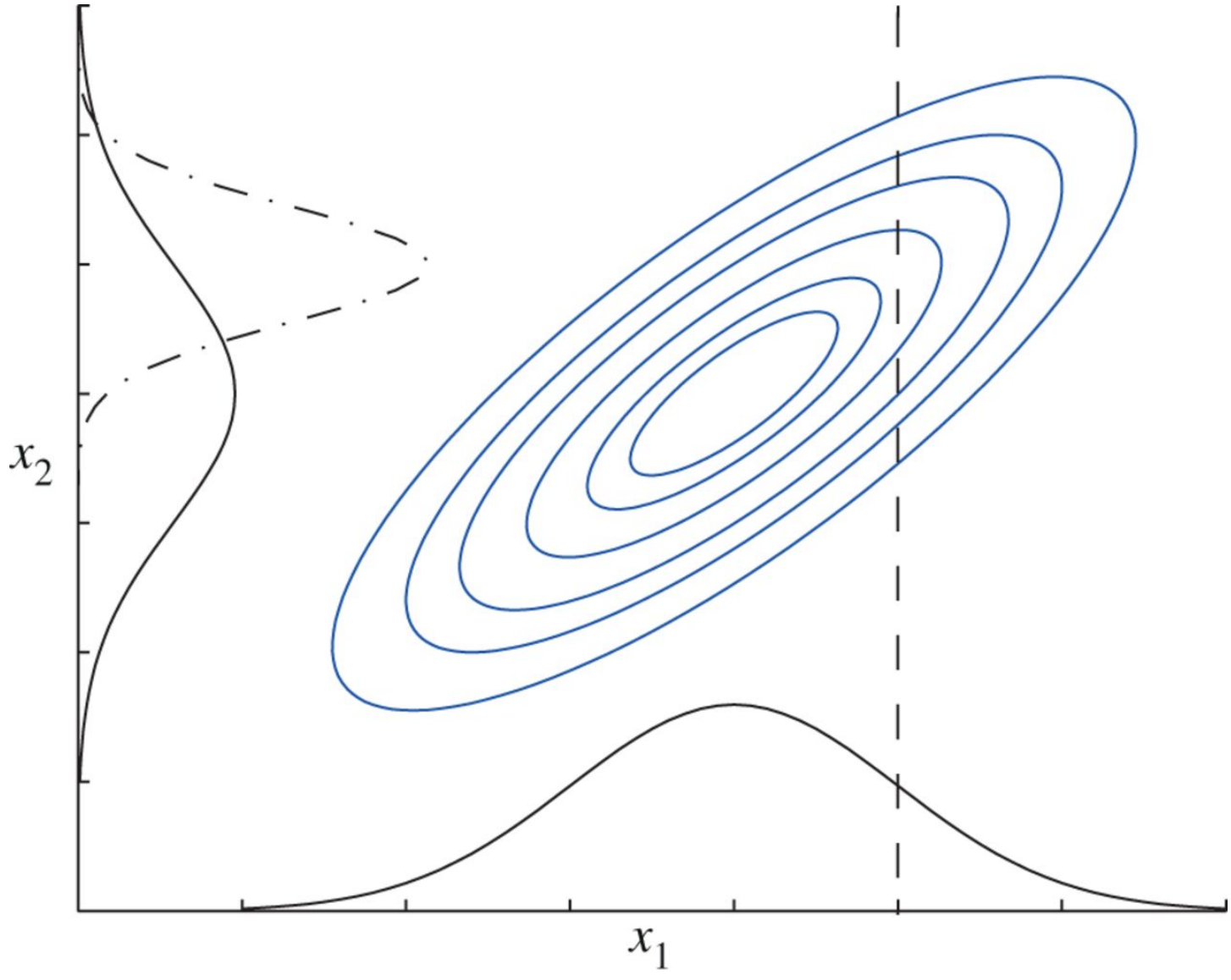
  - Obs: requer $m > n$ para que $\boldsymbol{\Sigma}$ seja inversível

- Modelar cada x_j como uma variável gaussiana independente equivalente a modelar $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ onde $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$
  - Nesse caso, se houver correlação nos dados ela não será identificada
  - Em compensação, requer menor complexidade e menos dados

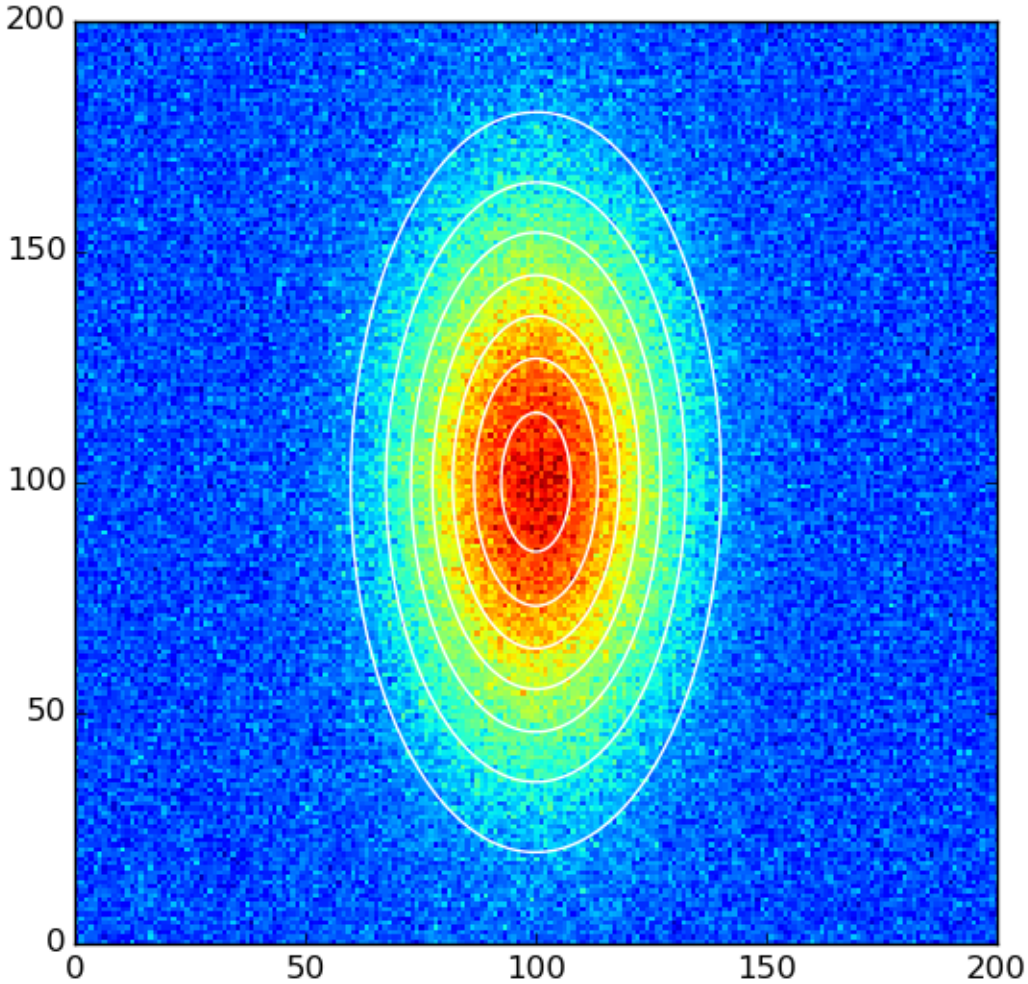
Exemplo



Exemplo



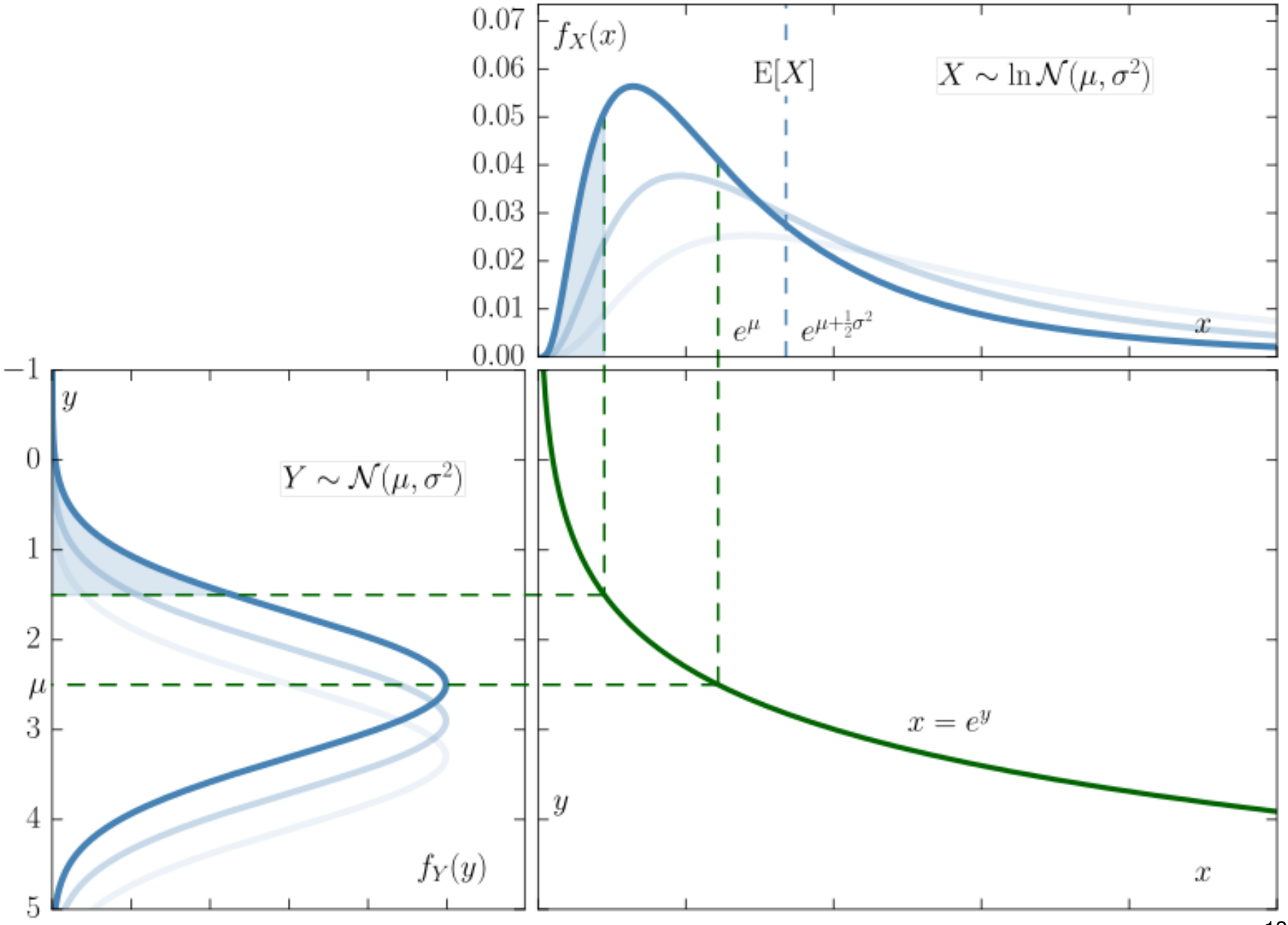
Exemplo



Recomendações

- Visualize os dados antes de modelar
- Mesmo que a distribuição não seja aproximadamente gaussiana, é possível que se torne aproximadamente gaussiana após alguma transformação
  - Ex: X é log-normal $\implies$ $Y = \log(X)$ é gaussiana
- Escolha atributos que possam variar significativamente no caso de uma anomalia
- Se possível, ajuste o modelo em amostras normais; se possuir amostras anômalas, guarde-as para validação e/ou teste
- Caso possua apenas amostras não-rotuladas para treinamento (possivelmente "contaminadas" com anomalias), torna-se importante usar um método de estimação de covariância robusto

Exemplo


0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
$f_X(x)$
E[X]
$X \sim \ln \mathcal{N}(\mu, \sigma^2)$
e^{μ}
$e^{\mu+\frac{1}{2}\sigma^2}$
x
−1
0
1
2
μ
3
4
5
y
$Y \sim \mathcal{N}(\mu, \sigma^2)$
$f_Y(y)$
$x = e^y$
y
x

Avaliação do Modelo

- Amostras rotuladas como anômalas podem ser usadas para avaliação do modelo
- Deve ser usada uma métrica robusta a desbalanceamento das classes (não usar acurácia):
  - Curva ROC (TPR x FPR)
  - Curva Precision-Recall

$$P = \frac{T_p}{T_p + F_p}, \qquad R = \frac{T_p}{T_p + F_n}$$

  - F_1 score (obs: "Positivo" = "Anômalo"):

$$F_1 = 2\frac{PR}{P + R} = \frac{2T_p}{2T_p + F_n + F_p}$$

    - Vantagem de ser um único número

Como Determinar o Limiar?

- Amostras rotuladas como anômalas também podem ser usadas para escolher o limiar ϵ, através de um conjunto de validação
- Escolha ϵ que maximiza o desempenho no conjunto de validação