Avaliação de Desempenho

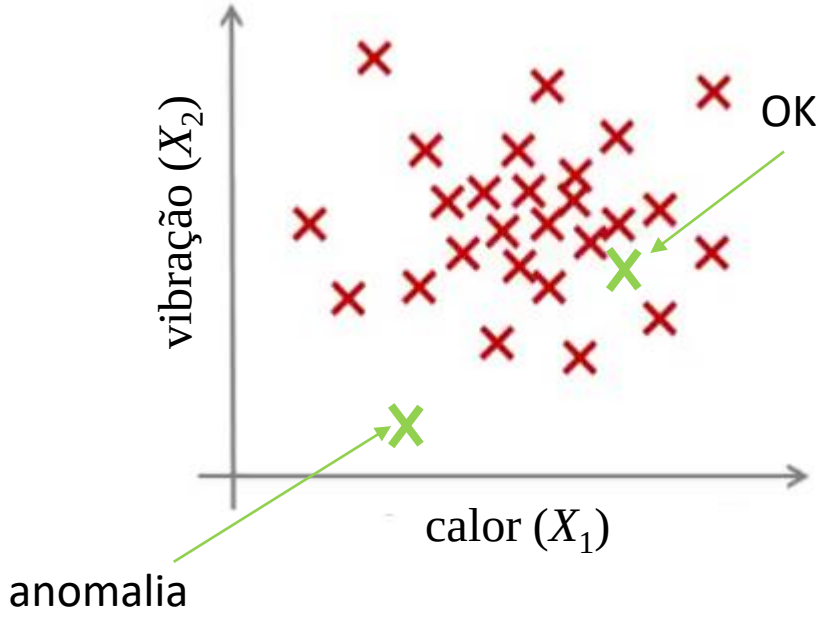
Aplicação da Distribuição Normal Multivariada

Detecção de Outliers

Detecção de outliers

- Classificação ocorrências como comportamento fora do comum

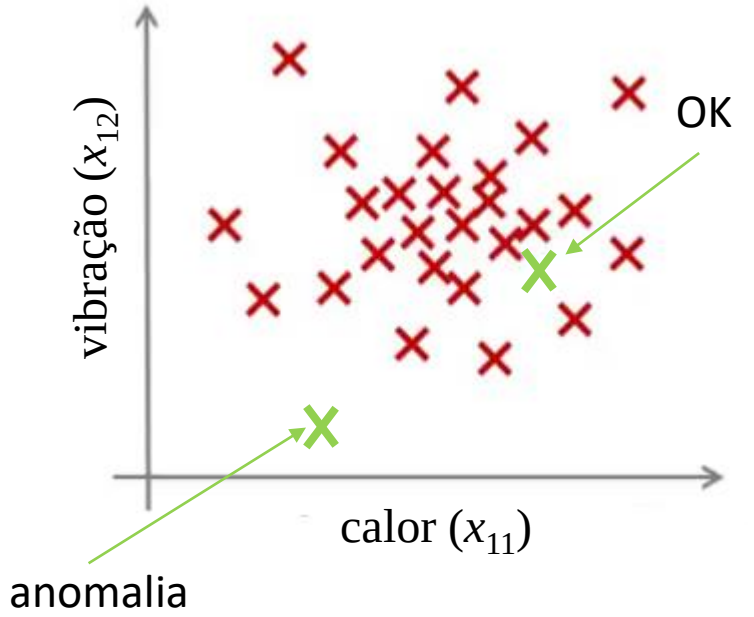
- Funcionamento anormal de um motor
  - Características
    - aquecimento, intensidade de vibração
  - Observações $X = [\ X_1\ ;\ X_2\ ]$
  - Novo motor: $X_{teste} = [\ x_1\ ;\ x_2\ ]$



Detecção de outliers

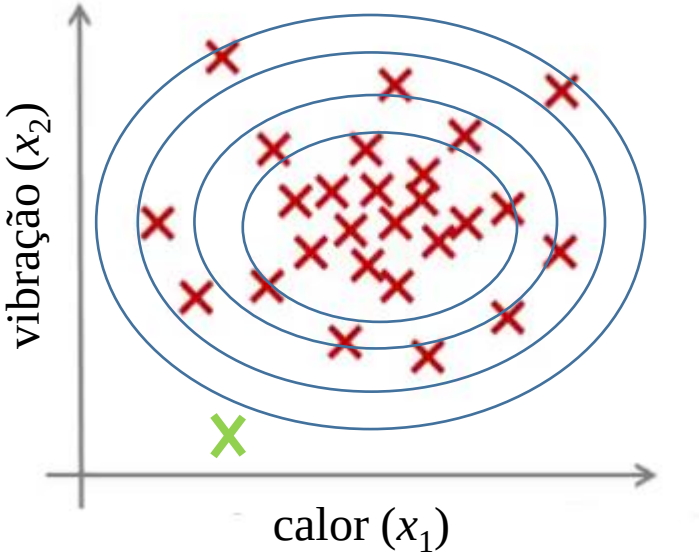
	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$		$\boldsymbol{X}_n$
$\boldsymbol{E}_1$	x_{11}	x_{12}	...	x_{1n}
$\boldsymbol{E}_2$	x_{21}	x_{22}	...	x_{2n}
$\boldsymbol{E}_m$	x_{m1}	x_{m2}	...	x_{mn}

- Dados de treinamento $\boldsymbol{X} = [\ \boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n\ ]$
- Cada característica é X_i é uma coluna de $\boldsymbol{X}$ com m valores $\boldsymbol{X}_j = [x_{1,j}\ ;\ x_{2,j}\ ;\ ...\ ;\ x_{m,j}\ ]$
- Cada exemplo é a uma linha de X com n componentes $\boldsymbol{E}_i = [x_{i,1}\ ,\ x_{i,2}\ ,\ ...\ ,\ x_{i,n}\ ]$
- No exemplo E_i = [calor , vibração]



Detecção de outliers

- Formulação
  - Dado um elemento E_{teste}
  - E_{teste} é uma anomalia?
  - Determinado pela probabilidade de ocorrência
  - Se $f(E_{teste}) < \varepsilon$ então E_{teste} é uma anomalia



Detecção de outliers

- Outros exemplos
  - Detecção de fraude
    - X_1 quantidade de vezes que faz o login
    - X_2 quantas páginas visita
    - X_3 quantidade de posts em fórums
    - ...
  - Monitoração de computadores em um data center
    - X_1 uso de memória
    - X_2 carga da CPU
    - X_3 tráfego de rede
    - ...

Detecção de outliers

Quando usar

Detecção de outliers (não supervisionado)

- Número pequeno de exemplos positivos
- Muitos tipos diferentes de anomalia
- Novas anomalias que não são semelhantes aos exemplos conhecidos

Aprendizado supervisionado

- Número grande de exemplos positivos e negativos
- Anomalias não vistas serão semelhantes àquelas encontradas no conjunto de treinamento
- Muitos exemplos rotulados

Detecção de outliers

Aplicações

Detecção de outliers (não supervisionado)

- Detecção de fraude
- Processos de fabricação
- Monitoração de máquinas em data center
- ...

Aprendizado supervisionado

- Classificação de e-mails (span)
- Previsão do tempo
- Classificação de tumor
- ...

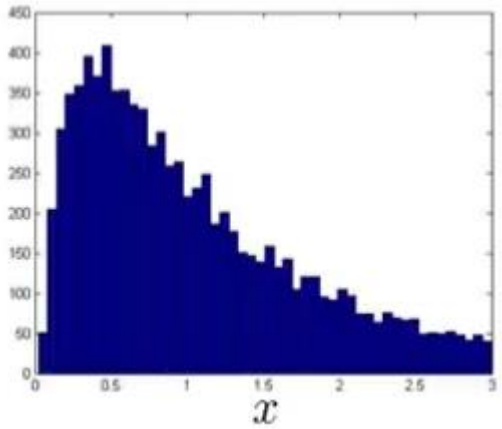
Dificuldades com o modelo

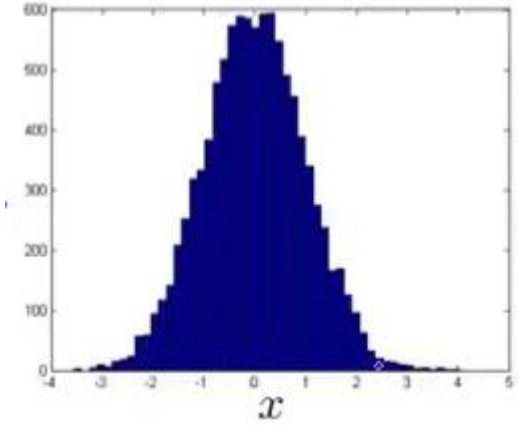
- Características não Gaussianas

$$x_1 \leftarrow \log(x_1)$$

$$x_2 \leftarrow \log(x_2 + c)$$

$$x_3 \leftarrow x_3^{1/k}$$

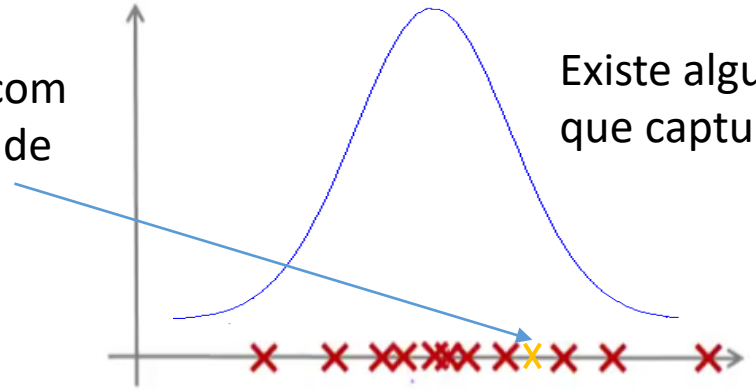




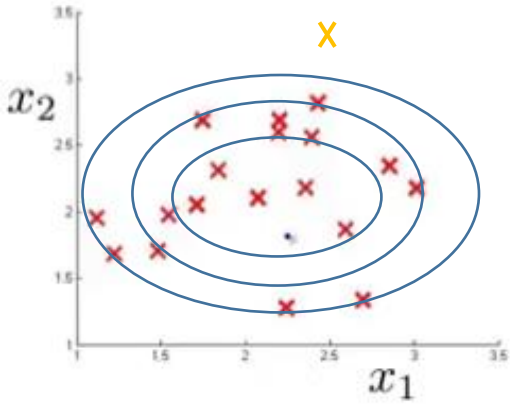
Dificuldades com o modelo

- Objetivo
  - $f(x)$ é grande para exemplos normais
  - $f(x)$ é pequeno para exemplos com anomalia
- Problema comum
  - $f(x)$ é semelhante para exemplos normais e anomalias

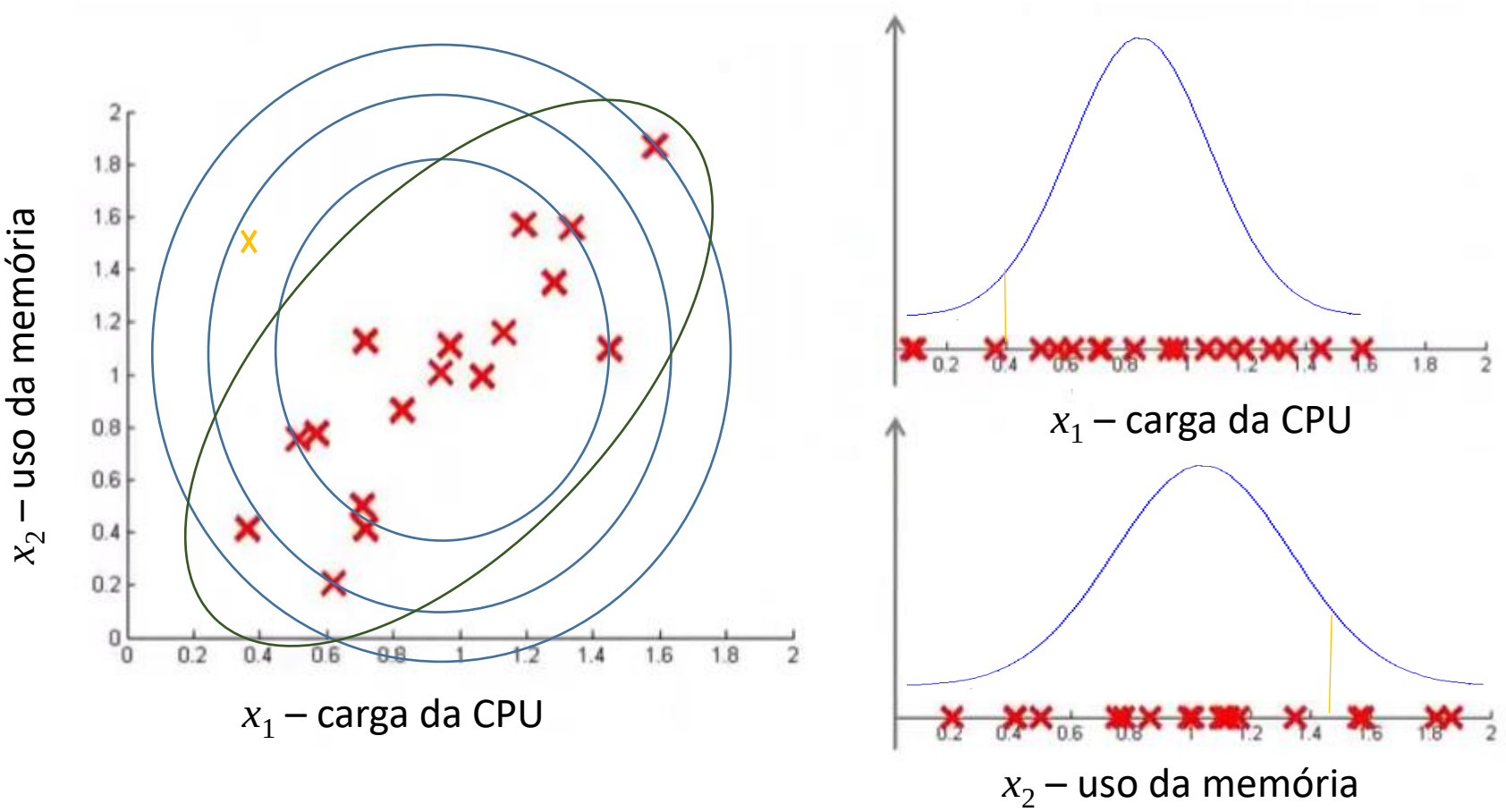
Exemplo anômalo com grande probabilidade



Existe alguma característica que captura o outlier?



Distribuição Gaussiana Maultivariada



- Usar a distribuição Gaussiana multivariada
- Captura possível correlação entre características

Distribuição Gaussiana Multivariada

- Um argumento ($\boldsymbol{X}$) e dois parâmetros ($\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$)

$$f(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right)$$

$$\boldsymbol{X} = [\,\boldsymbol{X}_1 \quad \dots \quad \boldsymbol{X}_n] \qquad \boldsymbol{\mu} = \begin{bmatrix} E[\boldsymbol{X}_1] \\ E[\boldsymbol{X}_2] \\ \dots \\ E[\boldsymbol{X}_n] \end{bmatrix} \qquad \Sigma = \begin{bmatrix} COV[\boldsymbol{X}_1,\boldsymbol{X}_1] & COV[\boldsymbol{X}_1,\boldsymbol{X}_2] & \cdots & COV[\boldsymbol{X}_1,\boldsymbol{X}_n] \\ COV[\boldsymbol{X}_2,\boldsymbol{X}_1] & COV[\boldsymbol{X}_2,\boldsymbol{X}_2] & \cdots & COV[\boldsymbol{X}_2,\boldsymbol{X}_n] \\ \vdots & \vdots & \ddots & \vdots \\ COV[\boldsymbol{X}_n,\boldsymbol{X}_1] & COV[\boldsymbol{X}_n,X_2] & \cdots & COV[\boldsymbol{X}_n,\boldsymbol{X}_n] \end{bmatrix}$$

- Estimativa dos parâmetros

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

No MatLab

```
mu = mean(X)
```

```
sigma = cov(X)
```

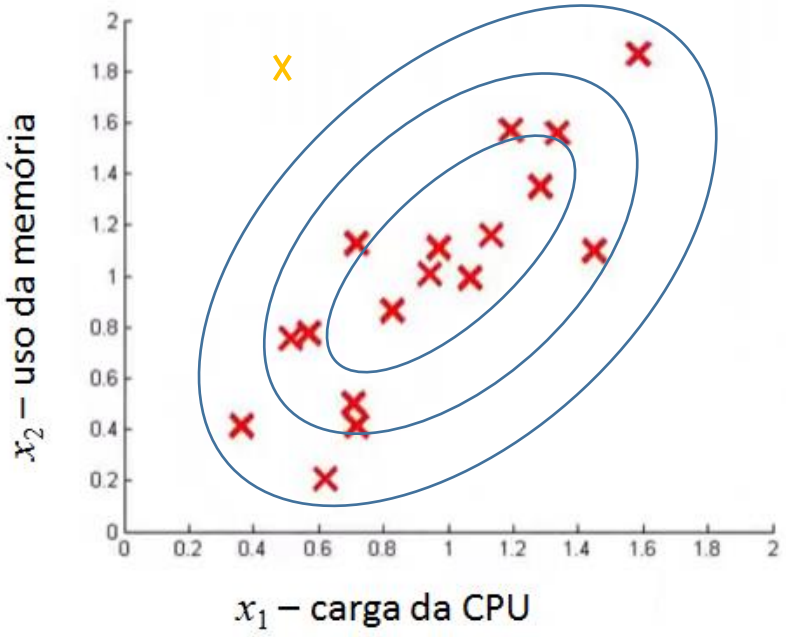
Algoritmo

1. Ajustar o modelo $f(\mathbf{X})$ fazendo

$$\boldsymbol{\mu} = \frac{1}{m}\sum_{i=1}^{m} X_i \qquad \boldsymbol{\Sigma} = \frac{1}{m}\sum_{i=1}^{m} (\mathbf{X} - \boldsymbol{\mu}) \cdot (\mathbf{X} - \boldsymbol{\mu})^T$$

2. Dado um novo exemplo $\mathbf{X}$ calcular

$$f(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right)$$



`mvnpdf(X, mu, sigma)`

3. Anomalia se $f(\mathbf{X}) < \varepsilon$ não anomalia caso contrário

Métricas de avaliação

- O foco deve estar na capacidade preditiva do modelo e não no tempo que leva para classificar ou criar um modelo, na escalabilidade, etc.

		Categoria prevista	
		Sim	Não
Categoria real	Sim	Verdadeiro positivo	Falso negativo
	Não	Falso positivo	Verdadeiro negativo

Métricas de avaliação

- Precisão $$prec = \frac{vp}{vp + fp}$$
- Revocação (recall) $$rec = \frac{vp}{vp + fn}$$
- F1 $$\mathrm{F}_1 = \frac{2 \cdot prec. \cdot rec}{prec + rec}$$

Ajuste e avaliação do modelo

- Assume-se que temos exemplos que sabemos ser normais (maior parte) e exemplos que são anômalos (poucos)
- Conjunto de treinamento (não rotulado): precisa ter exemplos não anômalos (mais de 80%) e exemplos anômalos (menos de 80%)
- $X_{trei} = [\ X_1, X_2, \ldots , X_n\ ]$
- Conjunto de validação $X_{val} = [X{val}_1, X{val}_2, \ldots, X{val}_n , Y_{val}]$
- Conjunto teste $X_{test} = [X{test}_1, X{test}_2, \ldots, X{test}_n , Y_{test}]$

Ajuste e avaliação do modelo

- Ajustar o modelo usando o conjunto de validação
- Usar o conjunto de validação para selecionar ε (otimizar a métrica F1)

$$y = \begin{cases} 1 & \text{se } f(x) < \varepsilon \\ 2 & \text{se } f(x) \geq \varepsilon \end{cases}$$

- Realizar previsões no conjunto de teste

Avaliação de Desempenho

Aplicação da Distribuição Normal Multivariada

Detecção de Outliers