

Aula 20 - Detecção de Anomalia

João Florindo

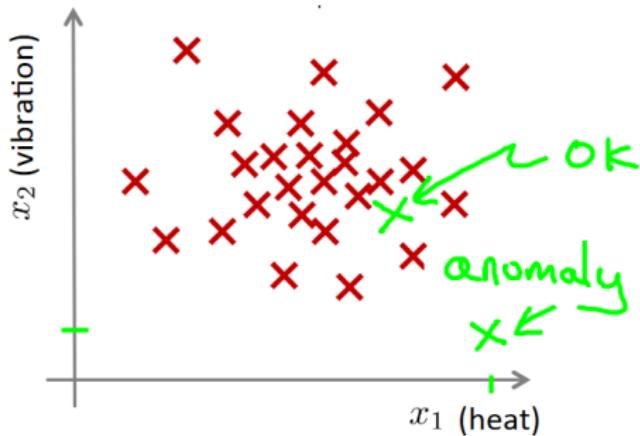
Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas - Brasil
florindo@unicamp.br

Outline

- 1 Motivação
- 2 Distribuição Gaussiana
- 3 Algoritmo
- 4 Avaliação do Modelo
- 5 Detecção de Anomalia vs Aprendizado Supervisionado
- 6 Transformação de Atributos
- 7 Distribuição Gaussiana Multivariada
- 8 Algoritmo

Exemplo

- Motores de avião - grande maioria funciona corretamente, uma minoria não funciona.
- Atributos:
 - x_1 : calor gerado
 - x_2 : intensidade de vibração
 - ...



- Anomalia em x_{teste} com base em uma **distribuição de probabilidade** $p(x)$:
 - $p(x) < \epsilon \Rightarrow$ anomalia
 - $p(x) \geq \epsilon \Rightarrow$ OK
- EXEMPLOS:
 - Linha de produção: motor do avião, etc.
 - Detecção de fraude
 - Computadores em um *data center*

- Anomalia em x_{teste} com base em uma **distribuição de probabilidade** $p(x)$:
 - $p(x) < \epsilon \Rightarrow$ anomalia
 - $p(x) \geq \epsilon \Rightarrow$ OK
- EXEMPLOS:
 - Linha de produção: motor do avião, etc.
 - Detecção de fraude
 - Computadores em um *data center*

- Anomalia em x_{teste} com base em uma **distribuição de probabilidade** $p(x)$:
 - $p(x) < \epsilon \Rightarrow$ anomalia
 - $p(x) \geq \epsilon \Rightarrow$ OK
- EXEMPLOS:
 - Linha de produção: motor do avião, etc.
 - Detecção de fraude
 - Computadores em um *data center*

- Anomalia em x_{teste} com base em uma **distribuição de probabilidade** $p(x)$:
 - $p(x) < \epsilon \Rightarrow$ anomalia
 - $p(x) \geq \epsilon \Rightarrow$ OK
- EXEMPLOS:
 - Linha de produção: motor do avião, etc.
 - Detecção de fraude
 - Computadores em um *data center*

- Anomalia em x_{teste} com base em uma **distribuição de probabilidade** $p(x)$:
 - $p(x) < \epsilon \Rightarrow$ anomalia
 - $p(x) \geq \epsilon \Rightarrow$ OK
- EXEMPLOS:
 - Linha de produção: motor do avião, etc.
 - Detecção de fraude
 - Computadores em um *data center*

- Anomalia em x_{teste} com base em uma **distribuição de probabilidade** $p(x)$:
 - $p(x) < \epsilon \Rightarrow$ anomalia
 - $p(x) \geq \epsilon \Rightarrow$ OK
- EXEMPLOS:
 - Linha de produção: motor do avião, etc.
 - Detecção de fraude
 - Computadores em um *data center*

- Anomalia em x_{teste} com base em uma **distribuição de probabilidade** $p(x)$:
 - $p(x) < \epsilon \Rightarrow$ anomalia
 - $p(x) \geq \epsilon \Rightarrow$ OK
- EXEMPLOS:
 - Linha de produção: motor do avião, etc.
 - Detecção de fraude
 - Computadores em um *data center*

Outline

- 1 Motivação
- 2 Distribuição Gaussiana
- 3 Algoritmo
- 4 Avaliação do Modelo
- 5 Detecção de Anomalia vs Aprendizado Supervisionado
- 6 Transformação de Atributos
- 7 Distribuição Gaussiana Multivariada
- 8 Algoritmo

Distribuição Gaussiana

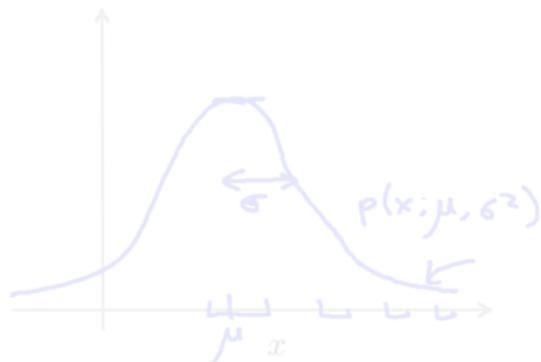
- Diz-se que $x \in \mathbb{R}$ segue uma distribuição Gaussiana (Normal) com média μ e variância σ^2 :

$$x \sim \mathcal{N}(\mu, \sigma^2).$$

- Probabilidade:

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

em que $p(x; \mu, \sigma^2)$ indica que $p(x)$ é **parametrizado** por μ e σ^2 .



Distribuição Gaussiana

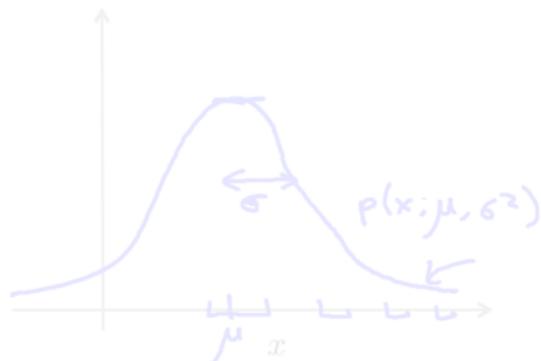
- Diz-se que $x \in \mathbb{R}$ segue uma distribuição Gaussiana (Normal) com média μ e variância σ^2 :

$$x \sim \mathcal{N}(\mu, \sigma^2).$$

- Probabilidade:

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

em que $p(x; \mu, \sigma^2)$ indica que $p(x)$ é **parametrizado** por μ e σ^2 .



Distribuição Gaussiana

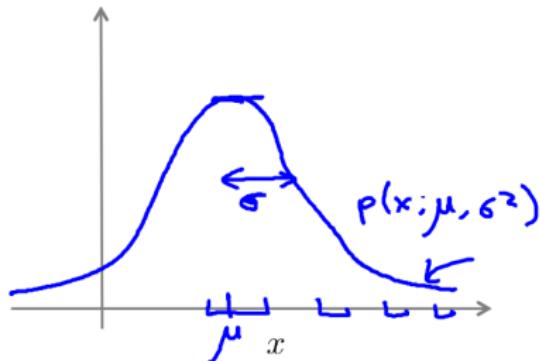
- Diz-se que $x \in \mathbb{R}$ segue uma distribuição Gaussiana (Normal) com média μ e variância σ^2 :

$$x \sim \mathcal{N}(\mu, \sigma^2).$$

- Probabilidade:

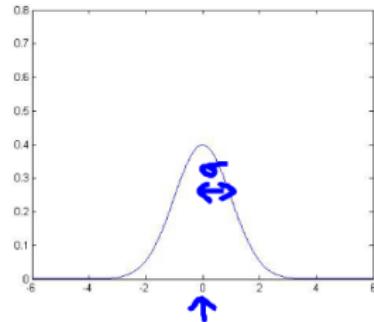
$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right),$$

em que $p(x; \mu, \sigma^2)$ indica que $p(x)$ é **parametrizado** por μ e σ^2 .

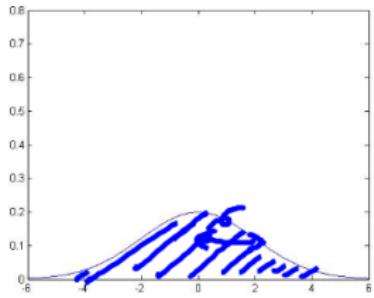


Distribuição Gaussiana

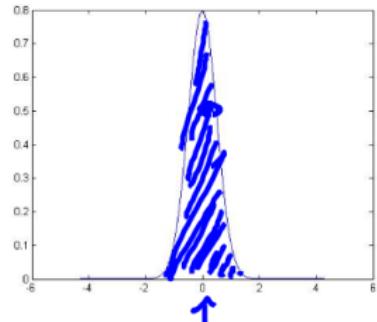
→ $\mu = 0, \sigma = 1$



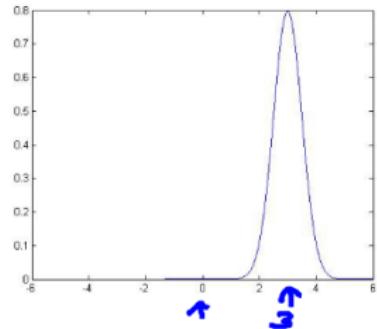
→ $\mu = 0, \sigma = 2$



→ $\mu = 0, \sigma = 0.5$



→ $\mu = 3, \sigma = 0.5$



Distribuição Gaussiana

- Parâmetros μ e σ obtidos a partir do conjunto de treinamento $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$, $x^{(i)} \in \mathbb{R}$:

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

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

NOTA

Existe na estatística uma outra definição de σ^2 que usa $\frac{1}{m-1}$. Essas têm propriedades teóricas diferentes, mas, para fins de machine learning, em que usualmente m é grande, não há diferença relevante e o comum em machine learning é a definição com $\frac{1}{m}$.

Outline

- 1 Motivação
- 2 Distribuição Gaussiana
- 3 Algoritmo
- 4 Avaliação do Modelo
- 5 Detecção de Anomalia vs Aprendizado Supervisionado
- 6 Transformação de Atributos
- 7 Distribuição Gaussiana Multivariada
- 8 Algoritmo

Algoritmo

- Cada atributo segue uma distribuição Gaussiana própria:

$$x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$

$$x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

$$x_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$$

...

- 1^a abordagem: Assume que as distribuições de cada atributo são independentes (sem correlação) e então a distribuição geral (conjunta) é o produto:

$$\begin{aligned} p(x) &= p(x_1; \mu_1, \sigma_1^2) p(x_2; \mu_2, \sigma_2^2) p(x_3; \mu_3, \sigma_3^2) \cdots p(x_n; \mu_n, \sigma_n^2) \\ &= \prod_{j=1}^n p(x_j; \mu_j, \sigma_j^2). \end{aligned}$$

Algoritmo

- Cada atributo segue uma distribuição Gaussiana própria:

$$x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$

$$x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

$$x_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$$

...

- 1^a abordagem: Assume que as distribuições de cada atributo são independentes (sem correlação) e então a distribuição geral (conjunta) é o produto:

$$\begin{aligned} p(x) &= p(x_1; \mu_1, \sigma_1^2) p(x_2; \mu_2, \sigma_2^2) p(x_3; \mu_3, \sigma_3^2) \cdots p(x_n; \mu_n, \sigma_n^2) \\ &= \prod_{j=1}^n p(x_j; \mu_j, \sigma_j^2). \end{aligned}$$

Algoritmo

- ① Escolher atributos x_i que poderiam indicar anomalias
- ② Ajustar parâmetros $\mu_1, \dots, \mu_n, \sigma_1^2, \dots, \sigma_n^2$:

$$\mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)}$$

$$\sigma_j^2 = \frac{1}{m} \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

- ③ Dado um novo exemplo x , calcular $p(x)$:

$$p(x) = \prod_{j=1}^n p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

Temos uma anomalia se $p(x) < \epsilon$.

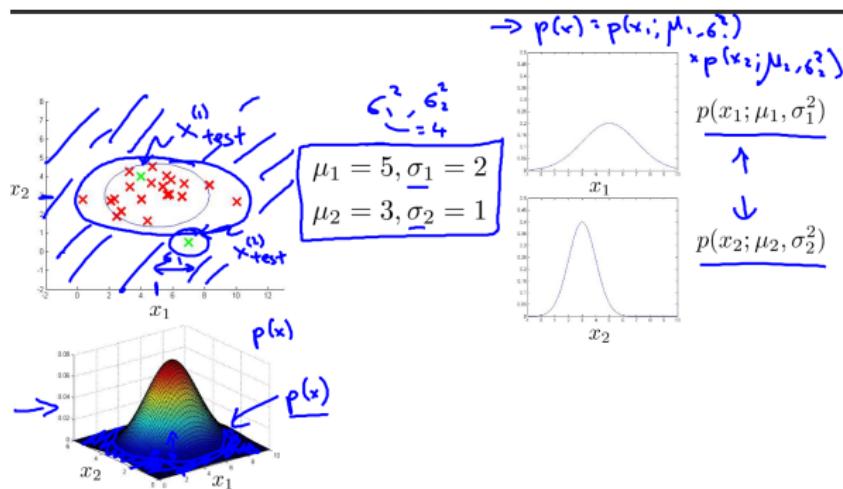
Algoritmo

Observação

O Passo 2 pode ser vetorizado:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \frac{1}{m} \sum_{i=1}^m x^{(i)}.$$

Exemplo



Suponha que $\epsilon = 0.02$ e quando calculamos temos

$$p(x_{\text{teste}}^{(1)}) = 0.0426 \geq \epsilon$$

$$p(x_{\text{teste}}^{(2)}) = 0.0021 < \epsilon,$$

de modo que $x^{(2)}$ é uma anomalia.

Outline

- 1 Motivação
- 2 Distribuição Gaussiana
- 3 Algoritmo
- 4 Avaliação do Modelo
- 5 Detecção de Anomalia vs Aprendizado Supervisionado
- 6 Transformação de Atributos
- 7 Distribuição Gaussiana Multivariada
- 8 Algoritmo

- Seja $y = 0$ se o exemplo for normal e $y = 1$ se for anômalo.
- Colocamos apenas exemplos normais / não anômalos no conjunto de treinamento $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$.
- Separamos ainda um conjunto de validação $\{(x_{cv}^{(1)}, y_{cv}^{(1)}), \dots, (x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})\}$ e um de teste $\{(x_{teste}^{(1)}, y_{teste}^{(1)}), \dots, (x_{teste}^{(m_{teste})}, y_{teste}^{(m_{teste})})\}$.
- EXEMPLO: Problema dos motores de avião, com
 - 10000 motores bons (normais)
 - 20 motores falhos (anômalos) (número comparativamente pequeno, na faixa 2 ~ 50)

- Seja $y = 0$ se o exemplo for normal e $y = 1$ se for anômalo.
- Colocamos apenas exemplos normais / não anômalos no conjunto de treinamento $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$.
- Separamos ainda um conjunto de validação $\{(x_{cv}^{(1)}, y_{cv}^{(1)}), \dots, (x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})\}$ e um de teste $\{(x_{teste}^{(1)}, y_{teste}^{(1)}), \dots, (x_{teste}^{(m_{teste})}, y_{teste}^{(m_{teste})})\}$.
- EXEMPLO: Problema dos motores de avião, com
 - 10000 motores bons (normais)
 - 20 motores falhos (anômalos) (número comparativamente pequeno, na faixa 2 ~ 50)

- Seja $y = 0$ se o exemplo for normal e $y = 1$ se for anômalo.
- Colocamos apenas exemplos normais / não anômalos no conjunto de treinamento $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$.
- Separamos ainda um conjunto de validação $\{(x_{cv}^{(1)}, y_{cv}^{(1)}), \dots, (x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})\}$ e um de teste $\{(x_{teste}^{(1)}, y_{teste}^{(1)}), \dots, (x_{teste}^{(m_{teste})}, y_{teste}^{(m_{teste})})\}$.
- EXEMPLO: Problema dos motores de avião, com
 - 10000 motores bons (normais)
 - 20 motores falhos (anômalos) (número comparativamente pequeno, na faixa 2 ~ 50)

- Seja $y = 0$ se o exemplo for normal e $y = 1$ se for anômalo.
- Colocamos apenas exemplos normais / não anômalos no conjunto de treinamento $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$.
- Separamos ainda um conjunto de validação $\{(x_{cv}^{(1)}, y_{cv}^{(1)}), \dots, (x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})\}$ e um de teste $\{(x_{teste}^{(1)}, y_{teste}^{(1)}), \dots, (x_{teste}^{(m_{teste})}, y_{teste}^{(m_{teste})})\}$.
- EXEMPLO: Problema dos motores de avião, com
 - 10000 motores bons (normais)
 - 20 motores falhos (anômalos) (número comparativamente pequeno, na faixa $2 \sim 50$)

- Seja $y = 0$ se o exemplo for normal e $y = 1$ se for anômalo.
- Colocamos apenas exemplos normais / não anômalos no conjunto de treinamento $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$.
- Separamos ainda um conjunto de validação $\{(x_{cv}^{(1)}, y_{cv}^{(1)}), \dots, (x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})\}$ e um de teste $\{(x_{teste}^{(1)}, y_{teste}^{(1)}), \dots, (x_{teste}^{(m_{teste})}, y_{teste}^{(m_{teste})})\}$.
- EXEMPLO: Problema dos motores de avião, com
 - 10000 motores bons (normais)
 - 20 motores falhos (anômalos) (número comparativamente pequeno, na faixa 2 ~ 50)

- Seja $y = 0$ se o exemplo for normal e $y = 1$ se for anômalo.
- Colocamos apenas exemplos normais / não anômalos no conjunto de treinamento $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$.
- Separamos ainda um conjunto de validação $\{(x_{cv}^{(1)}, y_{cv}^{(1)}), \dots, (x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})\}$ e um de teste $\{(x_{teste}^{(1)}, y_{teste}^{(1)}), \dots, (x_{teste}^{(m_{teste})}, y_{teste}^{(m_{teste})})\}$.
- EXEMPLO: Problema dos motores de avião, com
 - 10000 motores bons (normais)
 - 20 motores falhos (anômalos) (número comparativamente pequeno, na faixa $2 \sim 50$)

- Divisão usual:
 - Treino: 6000 motores bons
 - Validação: 2000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 2000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)
- Alternativa menos recomendada (repetir os motores bons da validação no teste):
 - Treino: 6000 motores bons
 - Validação: 4000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 4000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)

- Divisão usual:
 - Treino: 6000 motores bons
 - Validação: 2000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 2000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)
- Alternativa menos recomendada (repetir os motores bons da validação no teste):
 - Treino: 6000 motores bons
 - Validação: 4000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 4000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)

- Divisão usual:
 - Treino: 6000 motores bons
 - Validação: 2000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 2000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)
- Alternativa menos recomendada (repetir os motores bons da validação no teste):
 - Treino: 6000 motores bons
 - Validação: 4000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 4000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)

- Divisão usual:
 - Treino: 6000 motores bons
 - Validação: 2000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 2000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)
- Alternativa menos recomendada (repetir os motores bons da validação no teste):
 - Treino: 6000 motores bons
 - Validação: 4000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 4000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)

- Divisão usual:
 - Treino: 6000 motores bons
 - Validação: 2000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 2000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)
- Alternativa menos recomendada (repetir os motores bons da validação no teste):
 - Treino: 6000 motores bons
 - Validação: 4000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 4000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)

- Divisão usual:
 - Treino: 6000 motores bons
 - Validação: 2000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 2000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)
- Alternativa menos recomendada (repetir os motores bons da validação no teste):
 - Treino: 6000 motores bons
 - Validação: 4000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 4000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)

- Divisão usual:
 - Treino: 6000 motores bons
 - Validação: 2000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 2000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)
- Alternativa menos recomendada (repetir os motores bons da validação no teste):
 - Treino: 6000 motores bons
 - Validação: 4000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 4000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)

- Divisão usual:
 - Treino: 6000 motores bons
 - Validação: 2000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 2000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)
- Alternativa menos recomendada (repetir os motores bons da validação no teste):
 - Treino: 6000 motores bons
 - Validação: 4000 motores bons ($y = 0$) e 10 anômalos ($y = 1$)
 - Teste: 4000 motores bons ($y = 1$) e 10 anômalos ($y = 1$)

Avaliação

- Ajuste o modelo $p(x)$ sobre o treinamento $\{x^{(1)}, \dots, x^{(m)}\}$.
- Para um exemplo x de validação/teste, prever

$$y = \begin{cases} 1 & \text{se } p(x) < \epsilon \text{ (anomalia)} \\ 0 & \text{se } p(x) \geq \epsilon \text{ (normal)} \end{cases}$$

- Classes altamente desbalanceadas:
 - Verdadeiro positivo, falso positivo, falso negativo, verdadeiro negativo
 - *Precision/Recall*
 - F_1 -score
- Pode-se ainda usar o conjunto de validação para escolher ϵ e os atributos ideais.

Avaliação

- Ajuste o modelo $p(x)$ sobre o treinamento $\{x^{(1)}, \dots, x^{(m)}\}$.
- Para um exemplo x de validação/teste, prever

$$y = \begin{cases} 1 & \text{se } p(x) < \epsilon \text{ (anomalia)} \\ 0 & \text{se } p(x) \geq \epsilon \text{ (normal)} \end{cases}$$

- Classes altamente desbalanceadas:
 - Verdadeiro positivo, falso positivo, falso negativo, verdadeiro negativo
 - Precision/Recall
 - F_1 -score
- Pode-se ainda usar o conjunto de validação para escolher ϵ e os atributos ideais.

Avaliação

- Ajuste o modelo $p(x)$ sobre o treinamento $\{x^{(1)}, \dots, x^{(m)}\}$.
- Para um exemplo x de validação/teste, prever

$$y = \begin{cases} 1 & \text{se } p(x) < \epsilon \text{ (anomalia)} \\ 0 & \text{se } p(x) \geq \epsilon \text{ (normal)} \end{cases}$$

- Classes altamente desbalanceadas:
 - Verdadeiro positivo, falso positivo, falso negativo, verdadeiro negativo
 - *Precision/Recall*
 - F_1 -score
- Pode-se ainda usar o conjunto de validação para escolher ϵ e os atributos ideais.

Avaliação

- Ajuste o modelo $p(x)$ sobre o treinamento $\{x^{(1)}, \dots, x^{(m)}\}$.
- Para um exemplo x de validação/teste, prever

$$y = \begin{cases} 1 & \text{se } p(x) < \epsilon \text{ (anomalia)} \\ 0 & \text{se } p(x) \geq \epsilon \text{ (normal)} \end{cases}$$

- Classes altamente desbalanceadas:
 - Verdadeiro positivo, falso positivo, falso negativo, verdadeiro negativo
 - *Precision/Recall*
 - F_1 -score
- Pode-se ainda usar o conjunto de validação para escolher ϵ e os atributos ideais.

Avaliação

- Ajuste o modelo $p(x)$ sobre o treinamento $\{x^{(1)}, \dots, x^{(m)}\}$.
- Para um exemplo x de validação/teste, prever

$$y = \begin{cases} 1 & \text{se } p(x) < \epsilon \text{ (anomalia)} \\ 0 & \text{se } p(x) \geq \epsilon \text{ (normal)} \end{cases}$$

- Classes altamente desbalanceadas:
 - Verdadeiro positivo, falso positivo, falso negativo, verdadeiro negativo
 - *Precision/Recall*
 - F_1 -score
- Pode-se ainda usar o conjunto de validação para escolher ϵ e os atributos ideais.

Avaliação

- Ajuste o modelo $p(x)$ sobre o treinamento $\{x^{(1)}, \dots, x^{(m)}\}$.
- Para um exemplo x de validação/teste, prever

$$y = \begin{cases} 1 & \text{se } p(x) < \epsilon \text{ (anomalia)} \\ 0 & \text{se } p(x) \geq \epsilon \text{ (normal)} \end{cases}$$

- Classes altamente desbalanceadas:
 - Verdadeiro positivo, falso positivo, falso negativo, verdadeiro negativo
 - *Precision/Recall*
 - F_1 -score
- Pode-se ainda usar o conjunto de validação para escolher ϵ e os atributos ideais.

Avaliação

- Ajuste o modelo $p(x)$ sobre o treinamento $\{x^{(1)}, \dots, x^{(m)}\}$.
- Para um exemplo x de validação/teste, prever

$$y = \begin{cases} 1 & \text{se } p(x) < \epsilon \text{ (anomalia)} \\ 0 & \text{se } p(x) \geq \epsilon \text{ (normal)} \end{cases}$$

- Classes altamente desbalanceadas:
 - Verdadeiro positivo, falso positivo, falso negativo, verdadeiro negativo
 - *Precision/Recall*
 - F_1 -score
- Pode-se ainda usar o conjunto de validação para escolher ϵ e os atributos ideais.

Outline

- 1 Motivação
- 2 Distribuição Gaussiana
- 3 Algoritmo
- 4 Avaliação do Modelo
- 5 Detecção de Anomalia vs Aprendizado Supervisionado
- 6 Transformação de Atributos
- 7 Distribuição Gaussiana Multivariada
- 8 Algoritmo

Detecção de anomalia	Aprendizado supervisionado
<p>Número muito pequeno de exemplos positivos ($y = 1$) (normalmente $0 \sim 20$). Insuficiente para treinar um algoritmo supervisionado.</p>	<p>Número grande de exemplos positivos e negativos.</p>
<p>Grande número de exemplos negativos ($y = 0$). Usados para formar $p(x)$.</p>	
<p>Muitos “tipos” diferentes de anomalias. Difícil para um algoritmo aprender como é uma anomalia a partir dos exemplos positivos. Anomalias futuras podem não se parecer em nada com os exemplos de anomalia vistos até então.</p>	<p>Exemplos positivos suficientes para que se tenha uma ideia de como um exemplo positivo deve ser no futuro. EX.: Existem muitos tipos de <i>spam</i>, mas esses tipos podem ser incluídos no treinamento.</p>

Exemplos

Detecção de anomalia	Aprendizado supervisionado
Detecção de fraude	Classificação de <i>spam</i>
Linha de produção	Previsão do tempo
Monitoramento de um <i>data center</i>	Classificação de câncer
:	:

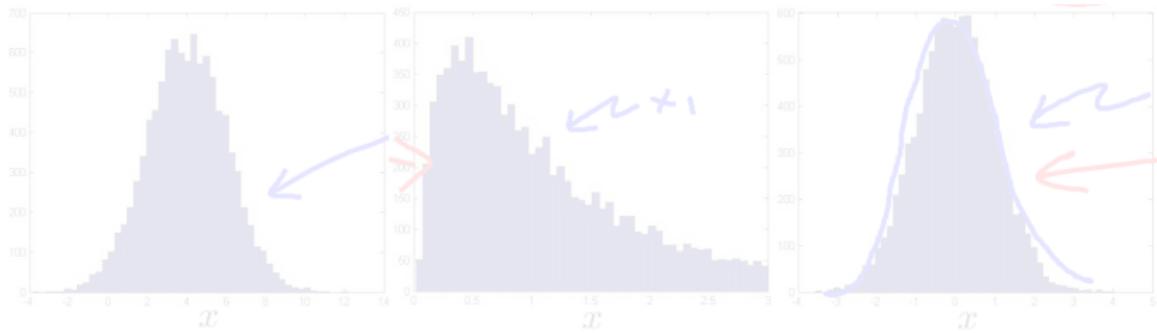
ATENÇÃO

Repare que, por exemplo, o problema de detecção de fraude em um grande banco, com muitos exemplos positivos, pode ser abordado como de aprendizado supervisionado.

Outline

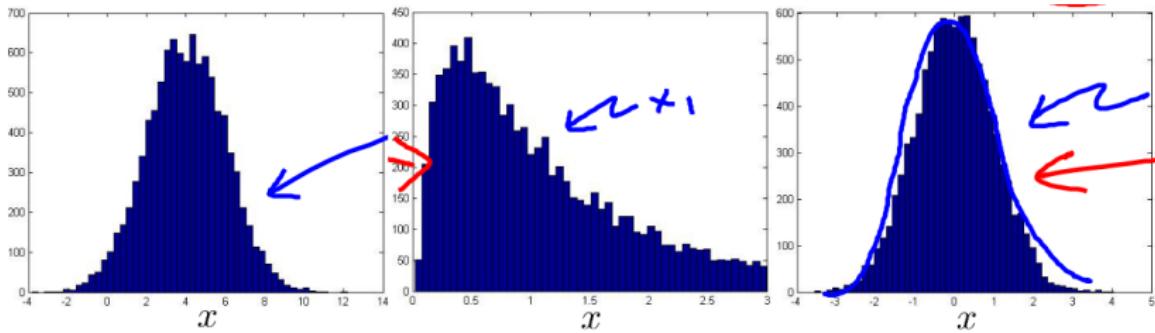
- 1 Motivação
- 2 Distribuição Gaussiana
- 3 Algoritmo
- 4 Avaliação do Modelo
- 5 Detecção de Anomalia vs Aprendizado Supervisionado
- 6 Transformação de Atributos
- 7 Distribuição Gaussiana Multivariada
- 8 Algoritmo

- Até aqui assumimos que x_j é Gaussiano.
- Verificamos pelo formato do histograma.



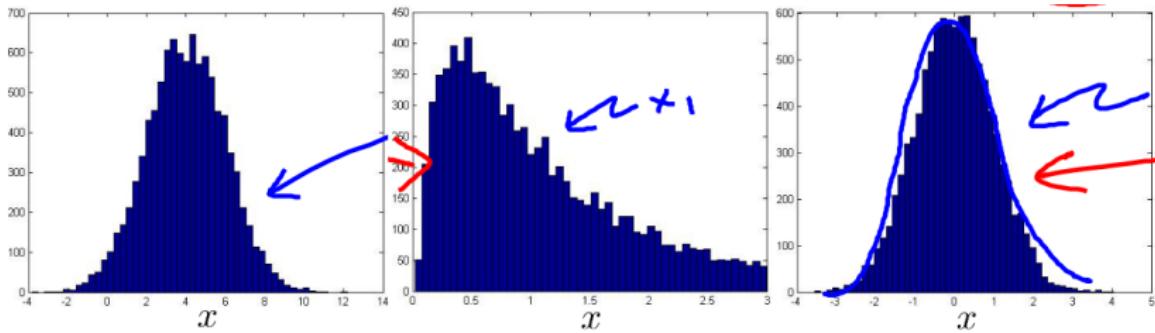
- Se não for Gaussiano (centro), faz-se uma transformação do atributo.
EX.: $\log(x)$, $\log(x + C)$, $x^{\frac{1}{2}}$, $x^{\frac{1}{3}}$, etc.

- Até aqui assumimos que x_j é Gaussiano.
- Verificamos pelo formato do histograma.



- Se não for Gaussiano (centro), faz-se uma transformação do atributo.
EX.: $\log(x)$, $\log(x + C)$, $x^{\frac{1}{2}}$, $x^{\frac{1}{3}}$, etc.

- Até aqui assumimos que x_j é Gaussiano.
- Verificamos pelo formato do histograma.



- Se não for Gaussiano (centro), faz-se uma transformação do atributo.
EX.: $\log(x)$, $\log(x + C)$, $x^{\frac{1}{2}}$, $x^{\frac{1}{3}}$, etc.

Análise de Erro

- Criar atributos com base nos exemplos classificados erroneamente.
- Queremos
 - $p(x)$ grande para exemplos x normais
 - $p(x)$ pequeno para exemplos x anômalos
- Problema mais comum: $p(x)$ comparável para exemplos normais e anômalos.
- Criar atributos que apresentem valores excepcionalmente grandes ou pequenos no caso de anomalia.

Análise de Erro

- Criar atributos com base nos exemplos classificados erroneamente.
- Queremos
 - $p(x)$ grande para exemplos x normais
 - $p(x)$ pequeno para exemplos x anômalos
- Problema mais comum: $p(x)$ comparável para exemplos normais e anômalos.
- Criar atributos que apresentem valores excepcionalmente grandes ou pequenos no caso de anomalia.

Análise de Erro

- Criar atributos com base nos exemplos classificados erroneamente.
- Queremos
 - $p(x)$ grande para exemplos x normais
 - $p(x)$ pequeno para exemplos x anômalos
- Problema mais comum: $p(x)$ comparável para exemplos normais e anômalos.
- Criar atributos que apresentem valores excepcionalmente grandes ou pequenos no caso de anomalia.

Análise de Erro

- Criar atributos com base nos exemplos classificados erroneamente.
- Queremos
 - $p(x)$ grande para exemplos x normais
 - $p(x)$ pequeno para exemplos x anômalos
- Problema mais comum: $p(x)$ comparável para exemplos normais e anômalos.
- Criar atributos que apresentem valores excepcionalmente grandes ou pequenos no caso de anomalia.

Análise de Erro

- Criar atributos com base nos exemplos classificados erroneamente.
- Queremos
 - $p(x)$ grande para exemplos x normais
 - $p(x)$ pequeno para exemplos x anômalos
- Problema mais comum: $p(x)$ comparável para exemplos normais e anômalos.
- Criar atributos que apresentem valores excepcionalmente grandes ou pequenos no caso de anomalia.

Análise de Erro

- Criar atributos com base nos exemplos classificados erroneamente.
- Queremos
 - $p(x)$ grande para exemplos x normais
 - $p(x)$ pequeno para exemplos x anômalos
- Problema mais comum: $p(x)$ comparável para exemplos normais e anômalos.
- Criar atributos que apresentem valores excepcionalmente grandes ou pequenos no caso de anomalia.

Análise de Erro - Exemplo

- Computadores em um *data center*:

- x_1 = uso de memória do computador
- x_2 = número de acessos ao disco/segundo
- x_3 = carga da CPU
- x_4 = tráfego de rede

- Identifica-se que um indicativo consistente de anomalia é quando há grande carga na CPU e um pequeno tráfego de rede.

- Cria-se então:

$$x_5 = \frac{\text{carga de CPU}}{\text{tráfego de rede}}$$

ou

$$x_6 = \frac{(\text{carga de CPU})^2}{\text{tráfego de rede}}.$$

Análise de Erro - Exemplo

- Computadores em um *data center*:

- x_1 = uso de memória do computador
 - x_2 = número de acessos ao disco/segundo
 - x_3 = carga da CPU
 - x_4 = tráfego de rede

- Identifica-se que um indicativo consistente de anomalia é quando há grande carga na CPU e um pequeno tráfego de rede.

- Cria-se então:

$$x_5 = \frac{\text{carga de CPU}}{\text{tráfego de rede}}$$

ou

$$x_6 = \frac{(\text{carga de CPU})^2}{\text{tráfego de rede}}.$$

Análise de Erro - Exemplo

- Computadores em um *data center*:

- x_1 = uso de memória do computador
- x_2 = número de acessos ao disco/segundo
- x_3 = carga da CPU
- x_4 = tráfego de rede

- Identifica-se que um indicativo consistente de anomalia é quando há grande carga na CPU e um pequeno tráfego de rede.

- Cria-se então:

$$x_5 = \frac{\text{carga de CPU}}{\text{tráfego de rede}}$$

ou

$$x_6 = \frac{(\text{carga de CPU})^2}{\text{tráfego de rede}}.$$

Análise de Erro - Exemplo

- Computadores em um *data center*:

- x_1 = uso de memória do computador
- x_2 = número de acessos ao disco/segundo
- x_3 = carga da CPU
- x_4 = tráfego de rede

- Identifica-se que um indicativo consistente de anomalia é quando há grande carga na CPU e um pequeno tráfego de rede.

- Cria-se então:

$$x_5 = \frac{\text{carga de CPU}}{\text{tráfego de rede}}$$

ou

$$x_6 = \frac{(\text{carga de CPU})^2}{\text{tráfego de rede}}.$$

Análise de Erro - Exemplo

- Computadores em um *data center*:
 - x_1 = uso de memória do computador
 - x_2 = número de acessos ao disco/segundo
 - x_3 = carga da CPU
 - x_4 = tráfego de rede
- Identifica-se que um indicativo consistente de anomalia é quando há grande carga na CPU e um pequeno tráfego de rede.
- Cria-se então:

$$x_5 = \frac{\text{carga de CPU}}{\text{tráfego de rede}}$$

ou

$$x_6 = \frac{(\text{carga de CPU})^2}{\text{tráfego de rede}}.$$

Análise de Erro - Exemplo

- Computadores em um *data center*:
 - x_1 = uso de memória do computador
 - x_2 = número de acessos ao disco/segundo
 - x_3 = carga da CPU
 - x_4 = tráfego de rede
- Identifica-se que um indicativo consistente de anomalia é quando há grande carga na CPU e um pequeno tráfego de rede.
- Cria-se então:

$$x_5 = \frac{\text{carga de CPU}}{\text{tráfego de rede}}$$

ou

$$x_6 = \frac{(\text{carga de CPU})^2}{\text{tráfego de rede}}.$$

Análise de Erro - Exemplo

- Computadores em um *data center*:
 - x_1 = uso de memória do computador
 - x_2 = número de acessos ao disco/segundo
 - x_3 = carga da CPU
 - x_4 = tráfego de rede
- Identifica-se que um indicativo consistente de anomalia é quando há grande carga na CPU e um pequeno tráfego de rede.
- Cria-se então:

$$x_5 = \frac{\text{carga de CPU}}{\text{tráfego de rede}}$$

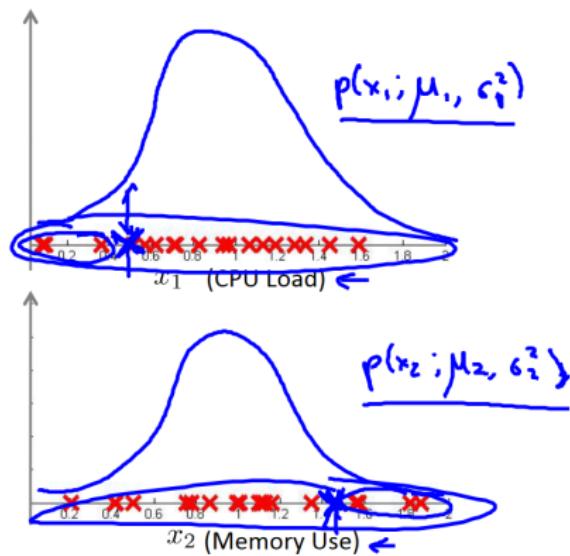
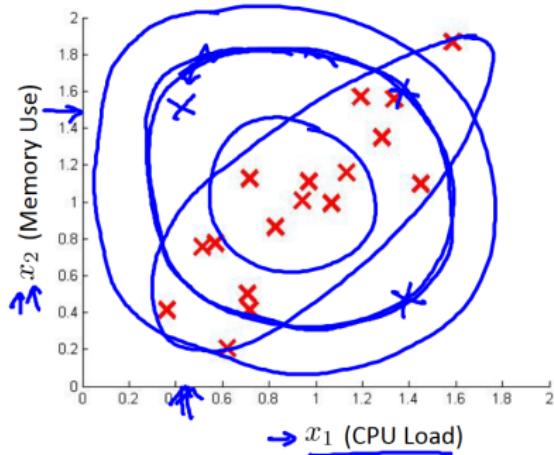
ou

$$x_6 = \frac{(\text{carga de CPU})^2}{\text{tráfego de rede}}.$$

Outline

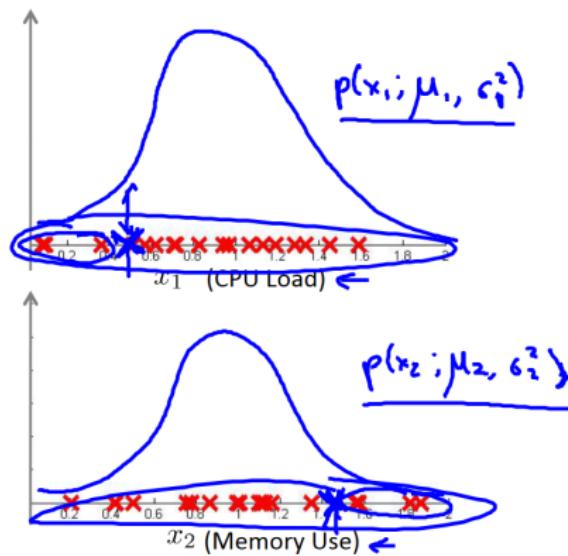
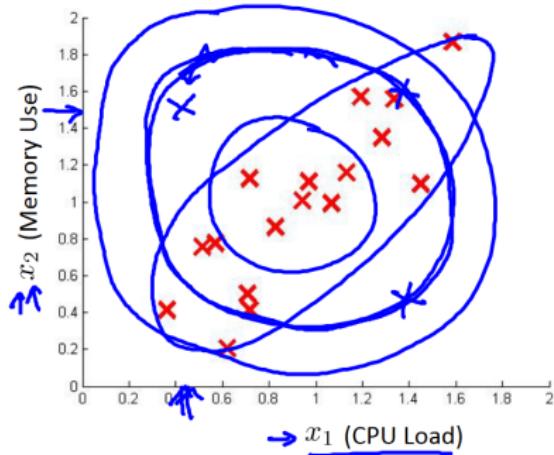
- 1 Motivação
- 2 Distribuição Gaussiana
- 3 Algoritmo
- 4 Avaliação do Modelo
- 5 Detecção de Anomalia vs Aprendizado Supervisionado
- 6 Transformação de Atributos
- 7 Distribuição Gaussiana Multivariada
- 8 Algoritmo

Motivação



- Acima, temos um ponto anômalo com $x_1 = 0.5$ e $x_2 = 1.5$.
- Porém $p(x_1)$ e $p(x_2)$ não são tão pequenos neste ponto (não detectado como anomalia).

Motivação



- Acima, temos um ponto anômalo com $x_1 = 0.5$ e $x_2 = 1.5$.
- Porém $p(x_1)$ e $p(x_2)$ não são tão pequenos neste ponto (não detectado como anomalia).

- Isso ocorre porque x_1 e x_2 são correlacionados.
- Neste caso, NÃO se deve modelar $p(x_1)$, $p(x_2)$, ... separadamente, mas sim de uma vez só.
- Para $x \in \mathbb{R}^n$, definimos $\mu \in \mathbb{R}^n$ e $\Sigma \in \mathbb{R}^{n \times n}$ (matriz de covariância) e então:

$$p(x) = (x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right),$$

em que $|\Sigma|$ é o determinante de Σ .

NOTA

Essa matriz de covariância é exatamente a mesma que vimos na análise de componentes principais.

- Isso ocorre porque x_1 e x_2 são correlacionados.
- Neste caso, NÃO se deve modelar $p(x_1)$, $p(x_2)$, ... separadamente, mas sim de uma vez só.
- Para $x \in \mathbb{R}^n$, definimos $\mu \in \mathbb{R}^n$ e $\Sigma \in \mathbb{R}^{n \times n}$ (matriz de covariância) e então:

$$p(x) = (x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right),$$

em que $|\Sigma|$ é o determinante de Σ .

NOTA

Essa matriz de covariância é exatamente a mesma que vimos na análise de componentes principais.

- Isso ocorre porque x_1 e x_2 são correlacionados.
- Neste caso, NÃO se deve modelar $p(x_1)$, $p(x_2)$, ... separadamente, mas sim de uma vez só.
- Para $x \in \mathbb{R}^n$, definimos $\mu \in \mathbb{R}^n$ e $\Sigma \in \mathbb{R}^{n \times n}$ (matriz de covariância) e então:

$$p(x) = (x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right),$$

em que $|\Sigma|$ é o determinante de Σ .

NOTA

Essa matriz de covariância é exatamente a mesma que vimos na análise de componentes principais.

- Isso ocorre porque x_1 e x_2 são correlacionados.
- Neste caso, NÃO se deve modelar $p(x_1)$, $p(x_2)$, ... separadamente, mas sim de uma vez só.
- Para $x \in \mathbb{R}^n$, definimos $\mu \in \mathbb{R}^n$ e $\Sigma \in \mathbb{R}^{n \times n}$ (matriz de covariância) e então:

$$p(x) = (x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right),$$

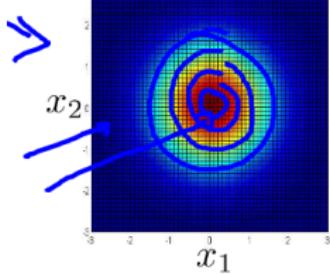
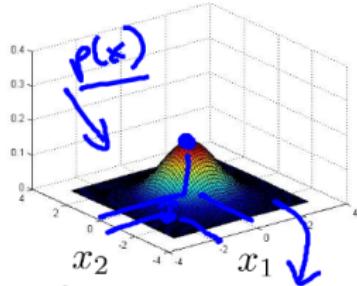
em que $|\Sigma|$ é o determinante de Σ .

NOTA

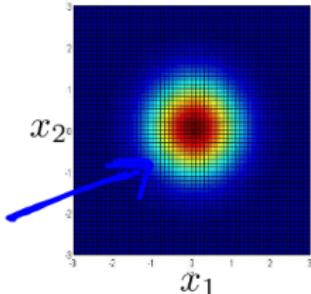
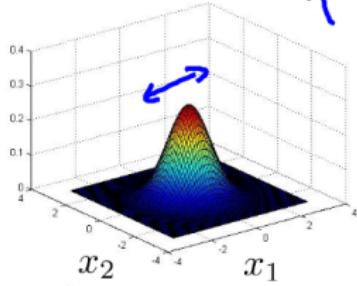
Essa matriz de covariância é exatamente a mesma que vimos na análise de componentes principais.

Exemplos

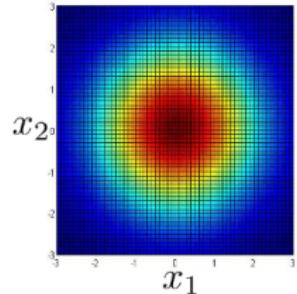
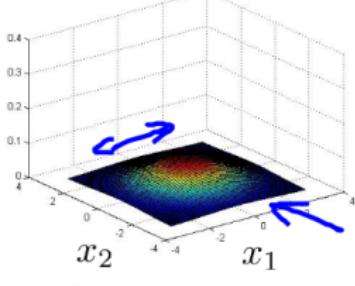
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$$

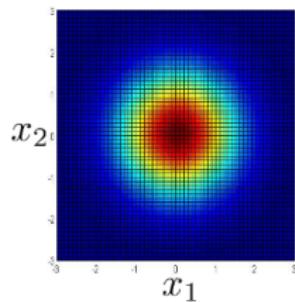
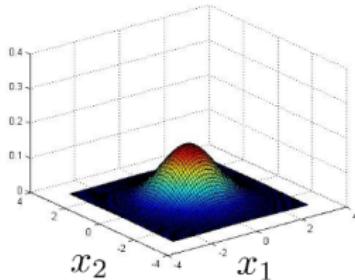


$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

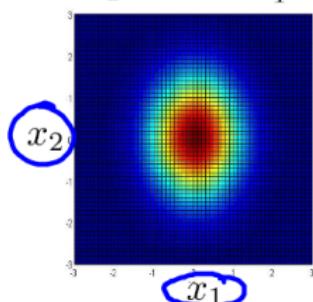
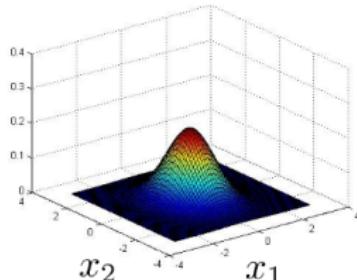


Exemplos

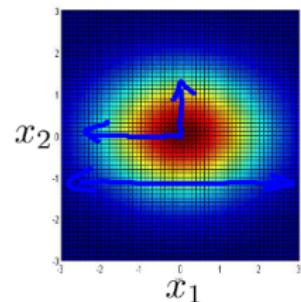
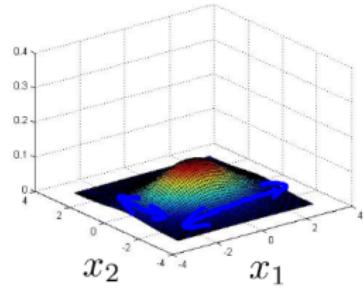
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$

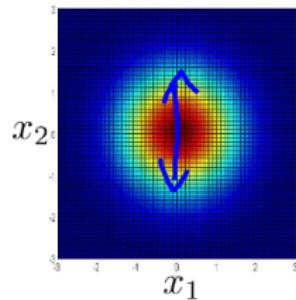
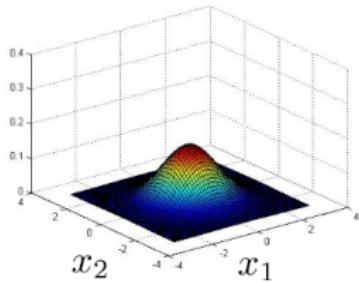


$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

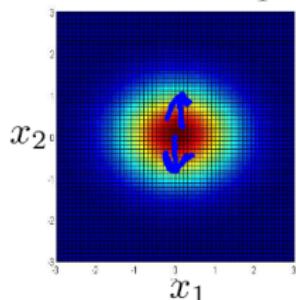
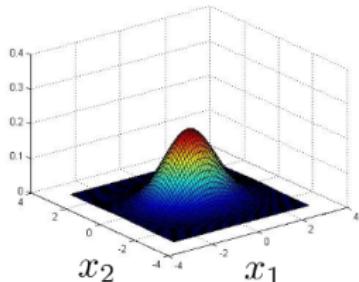


Exemplos

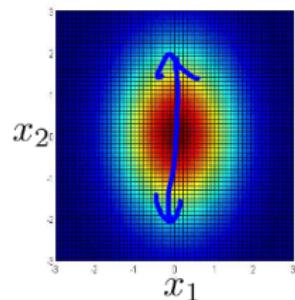
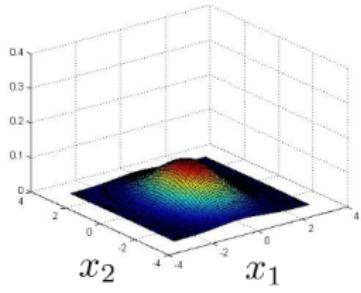
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0.6 \end{bmatrix}$$

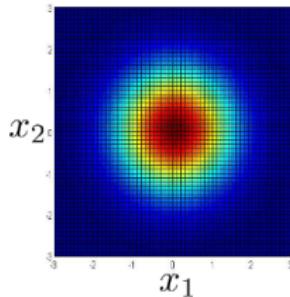
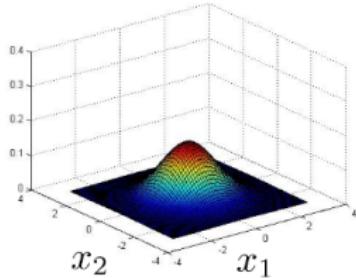


$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

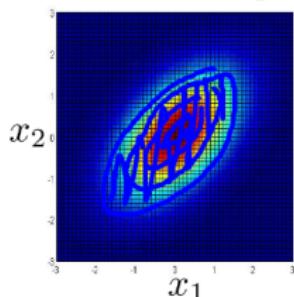
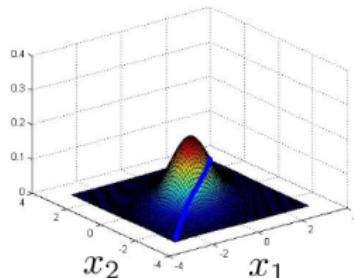


Exemplos

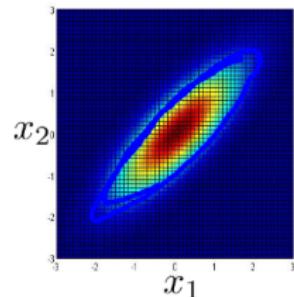
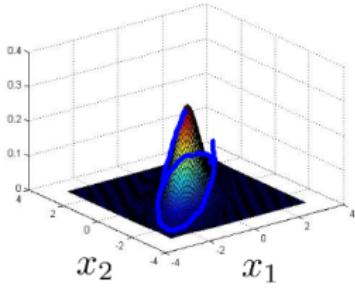
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

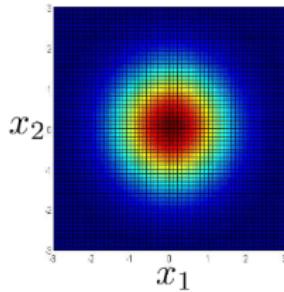
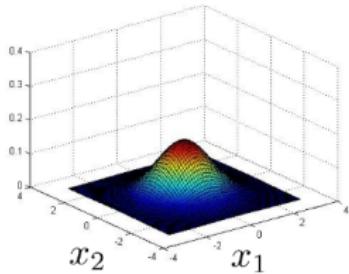


$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

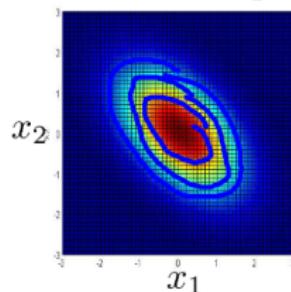
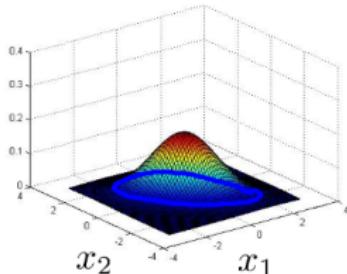


Exemplos

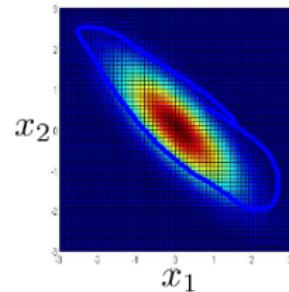
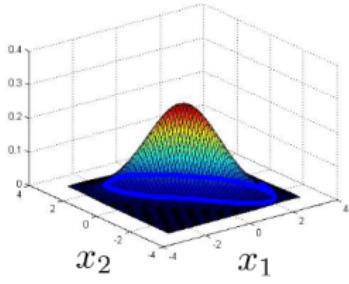
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

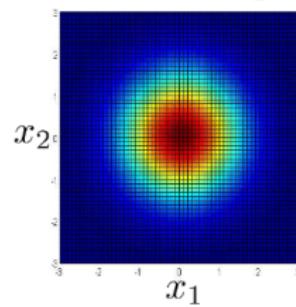
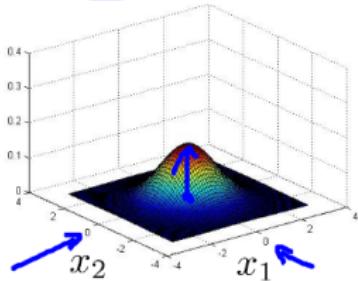


$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

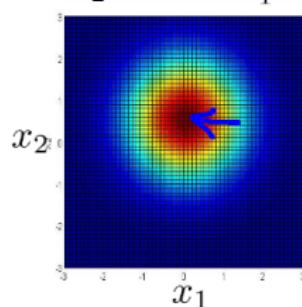
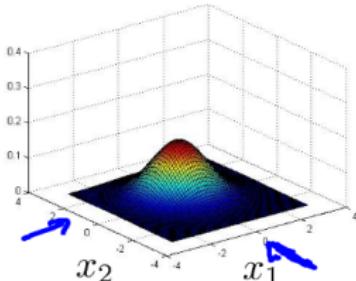


Exemplos

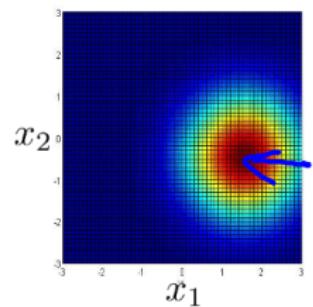
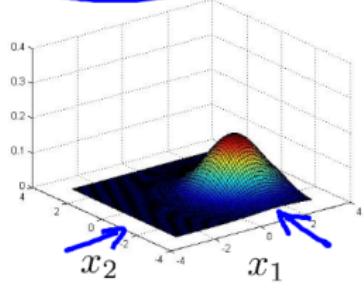
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Outline

- 1 Motivação
- 2 Distribuição Gaussiana
- 3 Algoritmo
- 4 Avaliação do Modelo
- 5 Detecção de Anomalia vs Aprendizado Supervisionado
- 6 Transformação de Atributos
- 7 Distribuição Gaussiana Multivariada
- 8 Algoritmo

- Dados os parâmetros $\mu \in \mathbb{R}^n$ e $\Sigma \in \mathbb{R}^{n \times n}$, temos

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right).$$

- μ e Σ são ajustados a partir do conjunto de treino $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$:

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

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

Note que essa expressão de Σ é exatamente equivalente à que usamos no PCA.

- Dados os parâmetros $\mu \in \mathbb{R}^n$ e $\Sigma \in \mathbb{R}^{n \times n}$, temos

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right).$$

- μ e Σ são ajustados a partir do conjunto de treino $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$:

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

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

Note que essa expressão de Σ é exatamente equivalente à que usamos no PCA.

Em geral:

- ① Ajustar o modelo $p(x)$ por

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

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

- ② Dado um novo exemplo x , calculamos

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

Considerar x anômalo se $p(x) < \epsilon$.

Relação com o Modelo Original

O modelo original

$$p(x) = p(x_1; \mu_1, \sigma_1^2) \times p(x_2; \mu_2, \sigma_2^2) \times p(x_3; \mu_3, \sigma_3^2) \times \cdots \times p(x_n; \mu_n, \sigma_n^2)$$

corresponde ao multivariado

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

se Σ for a matriz diagonal:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Relação com o Modelo Original

Diferenças:

Modelo original

$$p(x_1; \mu_1, \sigma_1^2) \times \cdots \times p(x_n; \mu_n, \sigma_n^2)$$

Criar atributos manualmente que capturem anomalias em que x_1 e x_2 assumem combinações raras de valores, p.ex., $x_3 = \frac{x_1}{x_2} = \frac{\text{carga da CPU}}{\text{memória}}$

Mais barato computacionalmente (escala melhor para n grande, p.ex. $n = 10000$ ou $n = 100000$)

OK mesmo se m (tamanho do treinamento) for pequeno

Gaussiana multivariada

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

Captura correlações automaticamente

Mais caro computacionalmente (calcular Σ^{-1} é muito caro para n grande)

Exige $m > n$, senão Σ não é invertível (singular). Ideal é $m \geq 10n$, já que Σ tem $\frac{n^2}{2}$ valores a serem determinados (já que é simétrica). Para Σ ser invertível é necessário também que não haja atributos redundantes (linearmente dependentes).