CT 720 Tópicos em Aprendizagem de Máquina e Classificação de Padrões

2-Fundamentos Matemáticos

Conteúdo

1. Introdução
2. Álgebra linear
3. Otimização de Lagrange
4. Teoria de probabilidade
5. Teste de hipóteses
6. Teoria da informação
7. Complexidade computacional

1-Introdução

- Objetivos
  - introduzir/comentar notação
  - revisar conceitos
  - contextualizar o conteúdo ao curso
- Referências

2-Algebra linear

- Notação e preliminares

$$\mathbf{x} \in \mathbf{R}^d$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \mathbf{x}^t = (x_1\, x_2 \cdots x_d)$$

$$\mathbf{M} = \left[m_{ij} \right] (n \times d), \;\; \mathbf{M} : (n \times d)$$

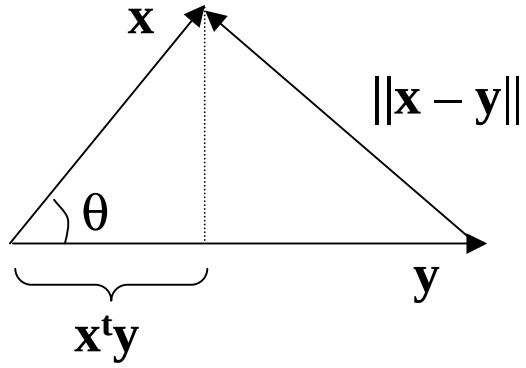
- Produto interno (escalar)

$$\mathbf{x}^t\mathbf{y} = \sum_{i=1}^{d} x_i y_i = \mathbf{y}^t\mathbf{x}$$

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^t\mathbf{x}}$$

$$\theta = \frac{\mathbf{x}^t\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

$$\left|\mathbf{x}^t\mathbf{y}\right| \leq \|\mathbf{x}\|\|\mathbf{y}\| \quad \text{(Cauchy} - \text{Schwarz)}$$



- Independência linear

$\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ LI varre espaço dimensão n

- Produto externo

$$\mathbf{x}\mathbf{y}^t = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_d y_1 & x_d y_2 & \cdots & x_d y_n \end{bmatrix}$$

- Gradiente

$$y = f(\mathbf{x}), \quad f : \mathbf{U}^d \rightarrow \mathbf{V}, \, \mathbf{x} \in \mathbf{U}^d$$

$$\nabla f(\mathbf{x}) = \text{grad} f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_d} \end{pmatrix}$$

- Jacobiano

$$\mathbf{f}:\mathbf{U}^d \rightarrow \mathbf{V}^n,\ \mathbf{x}\in\mathbf{U}^d$$

$$\mathbf{J}(\mathbf{x})=\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}=\begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_d} \\ \vdots & \ddots & \\ \frac{\partial f_n(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_n(\mathbf{x})}{\partial x_d} \end{pmatrix}$$

- $\mathbf{M} = \mathbf{M}(\theta)$

$$\frac{\partial \mathbf{M}}{\partial \theta}=\begin{pmatrix} \frac{m_{11}}{\partial \theta} & \cdots & \frac{m_{1d}}{\partial \theta} \\ \vdots & \ddots & \\ \frac{m_{n1}}{\partial \theta} & \cdots & \frac{m_{nd}}{\partial \theta} \end{pmatrix}$$

- Derivadas de matrizes

$$\frac{\partial}{\partial \theta}[\mathbf{M}^{-1}(\theta)\} = -\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial \theta}\mathbf{M}^{-1}$$

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{M}\mathbf{x}\} = \mathbf{M}$$

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{y}^t\mathbf{x}\} = \frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{y}\} = \mathbf{y}$$

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{M}\mathbf{x}\} = [\mathrm{M} + \mathbf{M}^t\}\mathbf{x}$$

- Inversão de matrizes e pseudo-inversas

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$$

$$\mathbf{M}^{-1} = \frac{\mathrm{Adj}[\mathbf{M}]}{|\mathbf{M}|}$$

$$\mathbf{M}^{*} = [\mathbf{M}^{t}\mathbf{M}]^{-1}\mathbf{M}^{t}$$

$$\mathbf{M}^{*}\mathbf{M} = \mathbf{I}$$

$$[\mathbf{M}\mathbf{N}]^{-1} = \mathbf{N}^{-1}\mathbf{M}^{-1}$$

- Série de Taylor

$$f(\mathbf{x}) = f(\mathbf{x}_o) + \left[\frac{\partial f}{\partial \mathbf{x}}\right]^{t}_{\mathbf{x}=\mathbf{x}_o} (\mathbf{x}-\mathbf{x}_o) + \frac{1}{2!}(\mathbf{x}-\mathbf{x}_o)^{t}\left[\frac{\partial^2 f}{\partial \mathbf{x}^2}\right]_{\mathbf{x}=\mathbf{x}_o} (\mathbf{x}-\mathbf{x}_o) + O(\|\mathbf{x}-\mathbf{x}_o\|^3)$$

- Autovalores e autovetores

$$\mathbf{M}\mathbf{x} = \lambda\mathbf{x} \qquad \mathbf{M}(d \times d),\ \mathbf{x} \in \mathbf{R}^d$$

$$(\mathbf{M} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

$\mathbf{x} = \mathbf{e}_j$ autovetor correspondente a λ_j

$$|\mathbf{M} - \lambda\mathbf{I}| = \lambda^d + a_1\lambda^{d-1} + \cdots + a_{d-1}\lambda + a_d$$

$$\mathrm{tr}[\mathbf{M}] = \sum_{i=1}^{d} \lambda_i$$

$$|\mathbf{M}| = \prod_{i=1}^{d} \lambda_i$$

3-Otimização de Lagrange

$$\min(\max)\ f(\mathbf{x})$$

$$\text{s.a} \qquad g(\mathbf{x}) = \mathbf{0}$$

$$L(\mathbf{x}, \lambda) = f(\mathrm{x}) + \lambda g(\mathbf{x})$$

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \lambda \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = 0$$

4-Teoria de probabilidade

- Variável aleatória
  - variável cujo valor depende de uma probabilidade
  - processo de atribuir um número real $x(\zeta)$ a cada $\zeta \in \chi$

- Variável aleatória discreta

$$\chi = \{v_1, v_2, ..., v_m\}$$
$$p_i = \Pr\{x = v_i\},\ i = 1, ..., m$$
$$p_i \geq 0$$
$$\Sigma_i p_i = 1$$

- Valores esperados

$$E[x] = \mu = \sum_{x \in \chi} xP(x) = \sum_{i=1}^{m} v_i p_i$$

$$E[f(x)] = \sum_{x \in \chi} f(x)P(x)$$

$$E[\alpha_1 f_1(x) + \alpha_2 f_2(x)] = \alpha_1 E[f_1(x)] + \alpha_2 E[f_2(x)]$$ Operador linear

$$E[x^2] = \sum_{x \in \chi} x^2 P(x)$$ segundo momento

$$\mathrm{Var}[x] = \sigma^2 = E[(x-\mu)^2] = \sum_{x \in \chi} (x-\mu)^2 P(x)$$

$$\mathrm{Var}[x] = E[x^2] - (E[x])^2$$

Se $v_1 = 0$, $v_2 = 1$ e $p = \Pr[x = 1]$ então $\mu = p$ e $\sigma = \sqrt{p(1-p)}$

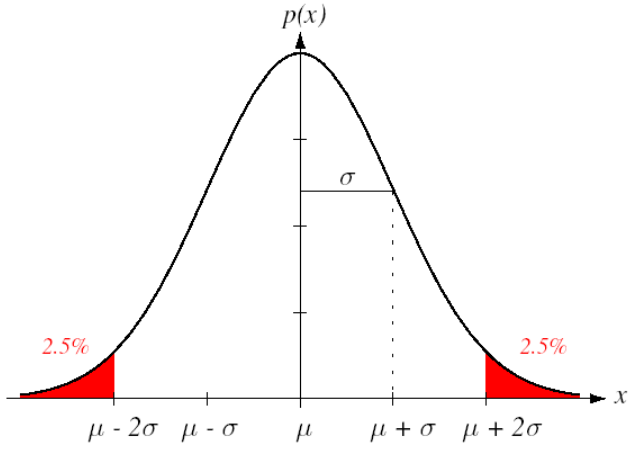
- Desigualdade de Chebyshev

$$\Pr[|x-\mu| > n\sigma] \leq \frac{1}{n^2}, \quad n > 1$$

$$\Pr[|x-\mu| \leq \sigma] \approx 0.68$$

$$\Pr[|x-\mu| \leq 2\sigma] \approx 0.95$$

$$\Pr[|x-\mu| \leq 3\sigma] \approx 0.997$$



Distribuição Normal

- Pares de variáveis aleatórias discretas

$X = \{v_1, v_2, ..., v_m\} \quad Y = \{w_1, w_2, ..., w_n\}$

(x, y): vetor (ponto) espaço $X \times Y$

$p_{ij} = \Pr[x = v_i, y = w_j]$

$P(x, y) =$ função distribuição conjunta de probabilidade

(*joint probability mass function*)

$$P(x,y) \geq 0 \qquad \sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$$

$$\left.\begin{aligned} P_x(x) &= \sum_{y \in Y} P(x, y) \\ P_y(y) &= \sum_{x \in X} P(x, y) \end{aligned}\right\} \quad \text{Distribuições marginais}$$

- Independência estatística

$$P(x,y) = P(x)P(y)$$
$$p_i = \Pr[x = v_i]$$
$$q_j = \Pr[x = w_j]$$
$$p_i q_j = P(v_i)P(w_j)$$

- Valores esperados:funções de duas variáveis

$$E[f(x,y)] = \sum_{x \in X} \sum_{y \in Y} f(x,y)P(x,y)$$
$$E[\alpha_1 f_1(x,y) + \alpha_2 f_2(x,y)] = \alpha_1 E[f_1(x,y)] + \alpha_2 E[f_2(x,y)]$$

- Primeiros momentos

$$\mu_x = E[x] = \sum_{x \in X} \sum_{y \in Y} x P(x,y)$$

$$\mu_y = E[y] = \sum_{x \in X} \sum_{y \in Y} y P(x,y)$$

- Variâncias

$$\sigma_x^2 = \mathrm{Var}[x] = E[(x-\mu_x)^2] = \sum_{x \in X} \sum_{y \in Y} (x-\mu_x)^2 P(x,y)$$

$$\sigma_y^2 = \mathrm{Var}[y] = E[(x-\mu_y)^2] = \sum_{x \in X} \sum_{y \in Y} (x-\mu_y)^2 P(x,y)$$

- Covariância

$$\sigma_{xy} = E[(x-\mu_x)(y-\mu_y)] = \sum_{x\in X}\sum_{y\in Y}(x-\mu_x)(y-\mu_y)P(x,y)$$

- Generalização

$$\boldsymbol{\mu} = E[\mathbf{z}] = \sum_{\mathbf{z}\in\mathbf{Z}}\mathbf{z}P(\mathbf{z}) \qquad \mathbf{z} = (x,y) \in \mathbf{Z} = X \times Y$$

$$\boldsymbol{\Sigma} = E[(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^t]$$

- Coeficiente de correlação

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \qquad -1 \le \rho \le 1$$

- Observações

  x e y estatisticamente independentes $\Rightarrow \ \sigma_{xy} = 0$

  $\sigma_{xy} = 0 \Rightarrow x$ e y não correlatas

  x e y não correlatas $\not\Rightarrow x$ e y são independentes

  x e y não correlatas $\Rightarrow x$ e y são independentes se distribuição for normal

  x e y estatisticamente independentes $\Rightarrow \ E[f(x)g(y)] = E[f(x)]E\{g(y)]$

- Cauchy Schwarz

  $\sigma_{xy} \leq \sigma_x^2 \sigma_y^2$

- Probabilidade condicional

$$\Pr[x = v_i \mid y = w_j] = \frac{\Pr[x = v_i, y = w_j]}{\Pr[y = w_j]}$$

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

- Probabilidade total

evento A ocorre em m diferentes maneiras $A_1, A_2, \ldots, A_m$

sub-eventos $A_1, A_2, \ldots, A_m$ mutuamente exclusivos

$$P(y) = \sum_{x \in X} P(x, y)$$

- Teorema de Bayes

$$P(x, y) = P(y \mid x)P(x)$$

$$P(x \mid y) = \frac{P(x, y)}{P(y)} = \frac{P(y \mid x)P(x)}{\sum_{x \in X} P(y \mid x)P(x)}$$

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{\sum_{x \in X} P(y \mid x)P(x)}$$

Regra de Bayes

posteriori = (likelihood $\times$ priori) / evidência

probabilidade a posteriori de uma causa, uma vez observado o efeito

- Vetor de variáveis aleatórias

$$\mathbf{x} = (x_1 x_2 \cdots x_d)^t$$

$$P(\mathbf{x}) \geq 0 \qquad \sum P(\mathbf{x}) = 1$$

$$P(\mathbf{x}) = P_{x_1}(x_1) P_{x_2}(x_2) \cdots P_{x_d}(x_d) = \prod_{i=1}^{d} P_{x_i}(x_i)$$ Variáveis independentes

$$P_{x_i}(x_i) = \sum_{j \neq i}^{d} P(x_1, x_2, \cdots, x_d)$$ Distribuição marginal

$$P(x_1, x_4) = \sum_{x_2} \sum_{x_3} \sum_{x_5} P(x_1, x_2, x_3, x_4, x_5)$$

- Vetor de variáveis aleatórias (cont.)

$$\mathbf{x} = (x_1 x_2 \cdots x_d)$$

$$P(x_1, x_2 \mid x_3) = \frac{P(x_1, x_2, x_3)}{P(x_3)} \quad , \quad P(x_1, x_2, x_3) = \sum_{x_4} \sum_{x_5} P(x_1, x_2, x_3, x_4, x_5)$$

$$P(\mathbf{x}_1, \mathbf{x}_2) = \frac{P(\mathbf{x}_2 \mid \mathbf{x}_1)}{\sum_{\mathbf{x}_1} P(\mathbf{x}_2 \mid \mathbf{x}_1) P(\mathbf{x}_1)}$$

$\mathbf{x}_1$, $\mathbf{x}_2$ vetores

- Esperanças e vetores médios

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{bmatrix} \rightarrow E(\mathbf{f}) = \begin{bmatrix} E[f_1(\mathbf{x})] \\ E[f_2(\mathbf{x})] \\ \vdots \\ E[f_n(\mathbf{x})] \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{f}(\mathbf{x}) P(\mathbf{x})$$

$$\boldsymbol{\mu} = E[\mathbf{x}] = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x})$$

- Matriz de covariância

$$\sigma_{ij} = \sigma_{ji} = E[(x_i - \mu_i)(x_j - \mu_j)] \quad i, j = 1, ..., d$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t]$$

$\boldsymbol{\Sigma}$ matriz positiva semidefinida (nenhum autovalor negativo)

- Variáveis aleatórias contínuas

Probabilidade $x \in$ (a, b)

$$\Pr[x \in (a,b)] = \int_a^b p(x)dx$$

$$p(x) \geq 0 \quad \text{e} \quad \int_{-\infty}^{\infty} p(x)dx = 1$$

$p(x)$ função densidade de probabilidade

- Valor esperado, média e variância

$$E[f(x)] = \int_{-\infty}^{\infty} f(x)\, p(x) dx$$

$$\mu = E[x] = \int_{-\infty}^{\infty} x p(x) dx$$

$$\mathrm{Var}[x] = \sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\, p(x) dx$$

$$\sigma^2 = E[x^2] - (E[x])^2$$

- Caso multivariável

$$p(\mathbf{x}) \geq 0 \quad \text{e} \quad \int_{-\infty}^{\infty} p(\mathbf{x}) d\mathbf{x} = 1$$

$$E[\mathbf{f}(\mathbf{x})] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x})\, p(\mathbf{x}) dx_1 dx_2 \cdots dx_d = \int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x})\, p(\mathbf{x}) d\mathbf{x}$$

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

$$\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int_{-\infty}^{\infty} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\, p(\mathbf{x}) d\mathbf{x}$$

- Componentes dos vetores estatisticamente independentes

$$P(\mathbf{x}) = \prod_{i=1}^{d} p_{x_i}(x_i)$$

matriz de covariância é diagonal

- Funções densidade de probabilidade condicional

$$P(x \mid y) = \frac{p(x, y)}{p(y)}$$

- Regra de Bayes

$$P(x \mid y) = \frac{p(y \mid x)\, p(x)}{\int_{-\infty}^{\infty} p(y \mid x) p(x) dx}$$

- Esperança em função de variáveis particulares

$$E_{x_1}[f(x_1, x_2)] = \int_{-\infty}^{\infty} f(x_1, x_2)\, p(x_1) dx_1$$

- Distribuição de soma de variáveis aleatórias independentes

x, y : variáveis aleatórias independentes e seja $z = x + y$

$$\mu_z = E[z] = E[x+y] = E[x] + E[y] = \mu_x + \mu_y$$

$$\sigma_z^2 = E[(z-\mu_z)^2] = E[(x+y-(\mu_x+\mu_y))^2] = \sigma_x^2 + \sigma_y^2$$

$$p(z) = p_x(x) * p_y(y) = \int_{-\infty}^{\infty} p_x(x)\, p_y(z-y)\, dx$$

média da soma é a soma das médias

variância da soma é a soma das variâncias

função densidade de probabilidade da soma é a convolução das densidades

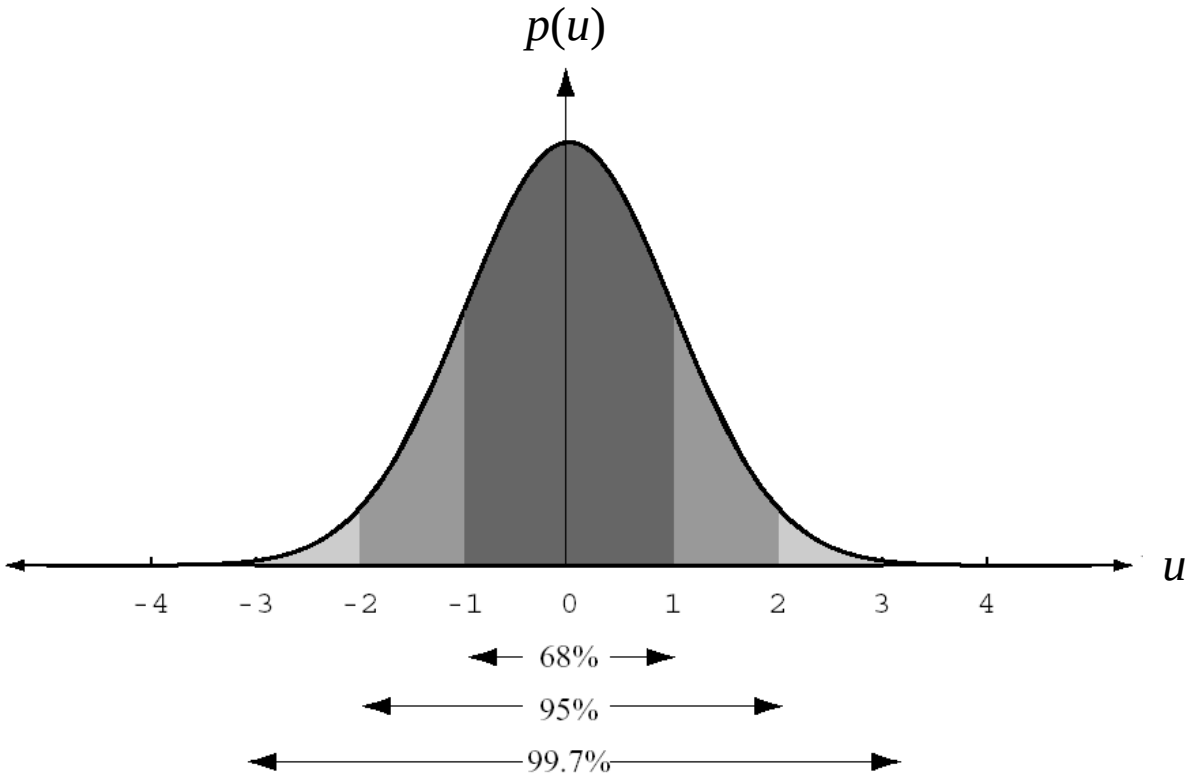
- Distribuição normal

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

$$E[1] = \int_{-\infty}^{\infty} p(x)dx = 1$$

$$E[x] = \int_{-\infty}^{\infty} xp(x)dx = \mu$$

$$E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2$$



$$\Pr[|x-\mu| \leq \sigma] \approx 0.68$$

$$\Pr[|x-\mu| \leq 2\sigma] \approx 0.95$$

$$\Pr[|x-\mu| \leq 3\sigma] \approx 0.997$$

- Distância de Mahalanobis

$$r = \frac{|x - \mu|}{\sigma}$$

- Variável aleatória padronizada

$$u = \frac{(x - \mu)}{\sigma}$$

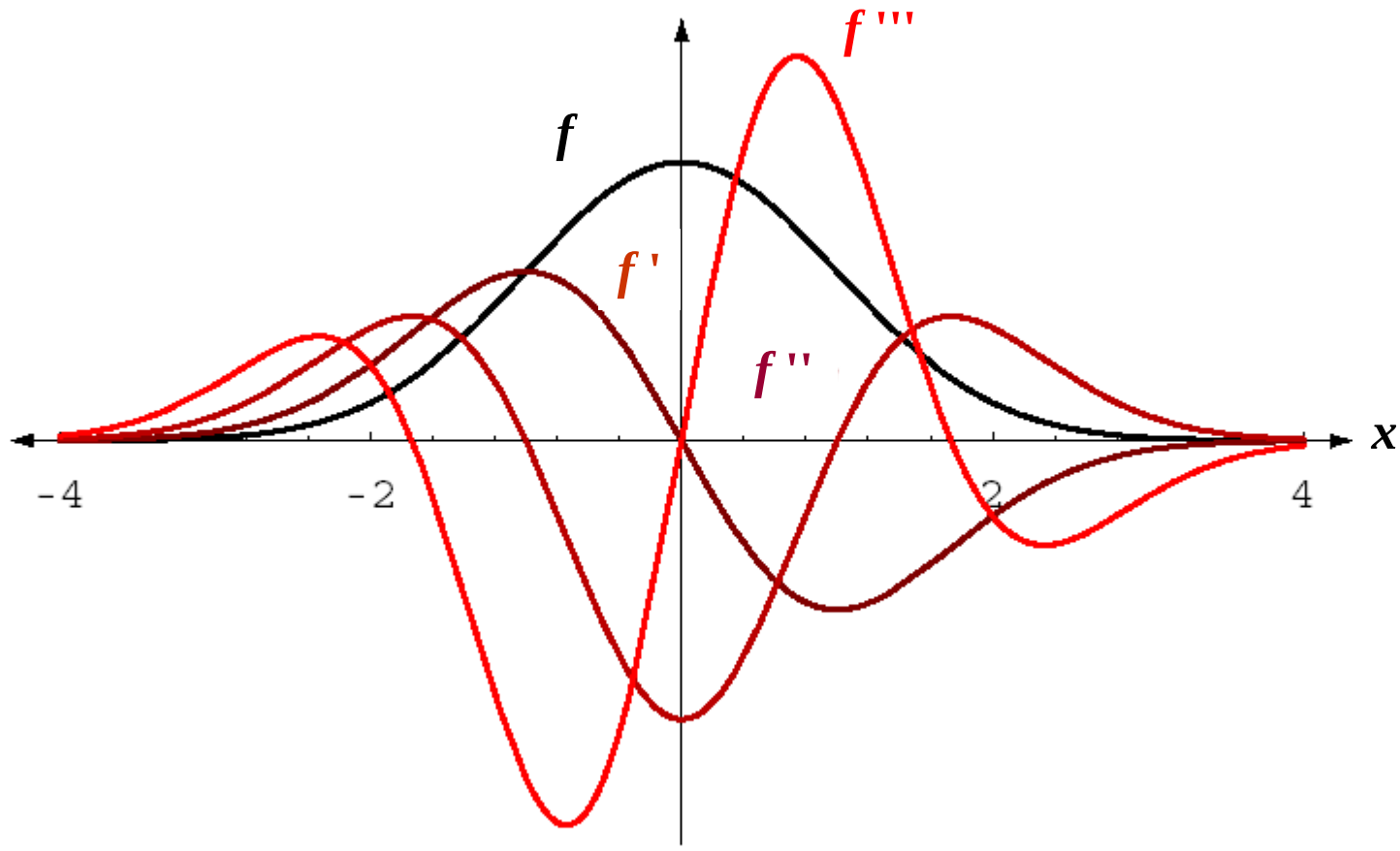
$$p(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \sim N(0,1)$$

- Derivadas de Gaussianas

$$\frac{\partial}{\partial x}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-1/2(x)^2/\sigma^2)}\right]=\frac{-x}{\sqrt{2\pi}\sigma^3}e^{-1/2(x^2/\sigma^2)}=\frac{-x}{\sigma^2}p(x)$$

$$\frac{\partial^2}{\partial x^2}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-1/2(x^2/\sigma^2)}\right]=\frac{1}{\sqrt{2\pi}\sigma^5}(-\sigma^2+x^2)e^{-1/2(x^2/\sigma^2)}=\frac{-\sigma^2+x^2}{\sigma^4}p(x)$$

$$\frac{\partial^3}{\partial x^3}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-1/2(x^2/\sigma^2)}\right]=\frac{1}{\sqrt{2\pi}\sigma^7}(3x\sigma^2-x^3)e^{-1/2(x^2/\sigma^2)}=\frac{-3x\sigma^2-x^3}{\sigma^6}p(x)$$


f '''
f
f '
f ''
x
-4
-2
2
4

- Integrais de Gaussianas

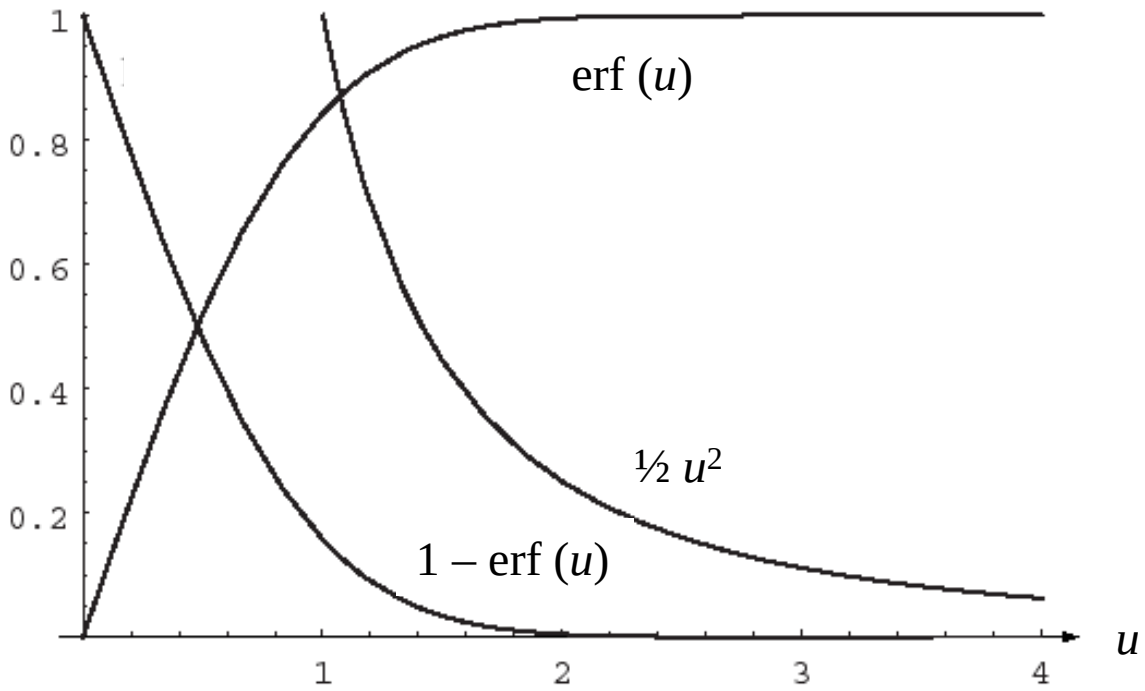
$$\mathrm{erf}(u) = \frac{1}{\sqrt{\pi}} \int_0^u e^{-x^2} dx$$ Função erro

$$\Gamma(n+1) = \int_0^\infty x^n e^{-x} dx$$ Função Gama

$$\Gamma(n+1) = n\Gamma(n)$$

$$\Gamma(1/2) = \sqrt{\pi}$$

$$\Gamma(n+1) = n! \quad n \text{ inteiro}$$


1
0.8
0.6
0.4
0.2
1
2
3
4
erf (u)
½ u2
1 – erf (u)
u

- Densidades normais multivariáveis

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x}) d\mathbf{x}$$

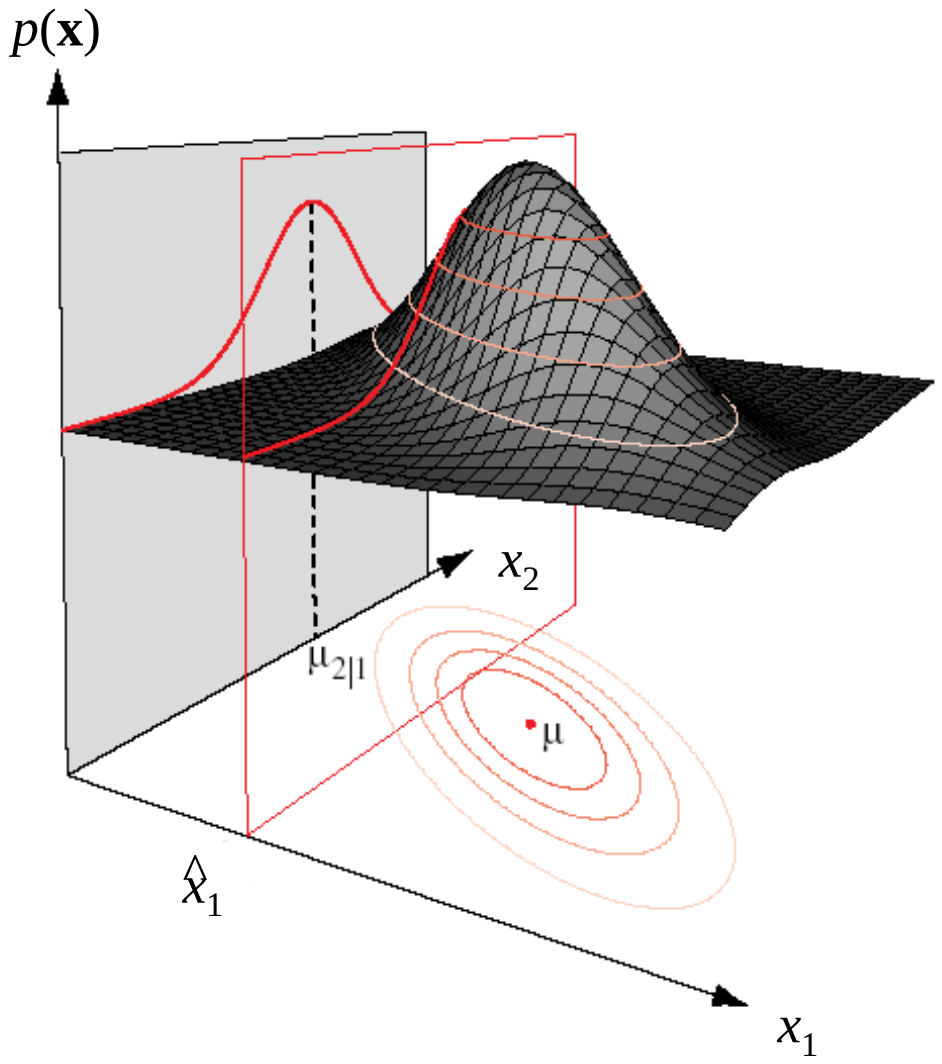
$$\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t] = \int_{-\infty}^{\infty} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t\, p(\mathbf{x}) d\mathbf{x}$$

$$d = 2$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \qquad \rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$$

$$|\boldsymbol{\Sigma}| = \sigma_1^2\sigma_2^2(1-\rho^2)$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1/\sigma_1^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & 1/\sigma_2^2 \end{bmatrix}$$


$p(\mathbf{x})$
x_2
$\mu_{2|1}$
μ
x_1
x_1

5-Teste de hipóteses

- Conceito
  - Mecanismo formal para decidir se os resultados de um experimento são significativos ou acidentais.
- Conjunto de amostras (observações)

$$\chi_n = \{x_1, x_2, \ldots, x_n\}$$

x_i: amostras com distribuição D_0 ou outra distribuição

Classificação: qual distribuição é a fonte das amostras? D_0 ou a outra distribuição?

- Teste de hipóteses

  Assume inicialmente que D_0 é a fonte das amostras: hipótese nula H_0
  A partir do valor de uma amostra pergunta-se se podemos rejeitar
  a hipótese nula, isto é, afirmar com certo grau de confiança (expresso
  como uma probabilidade) que a amostra não foi gerada por D_0.

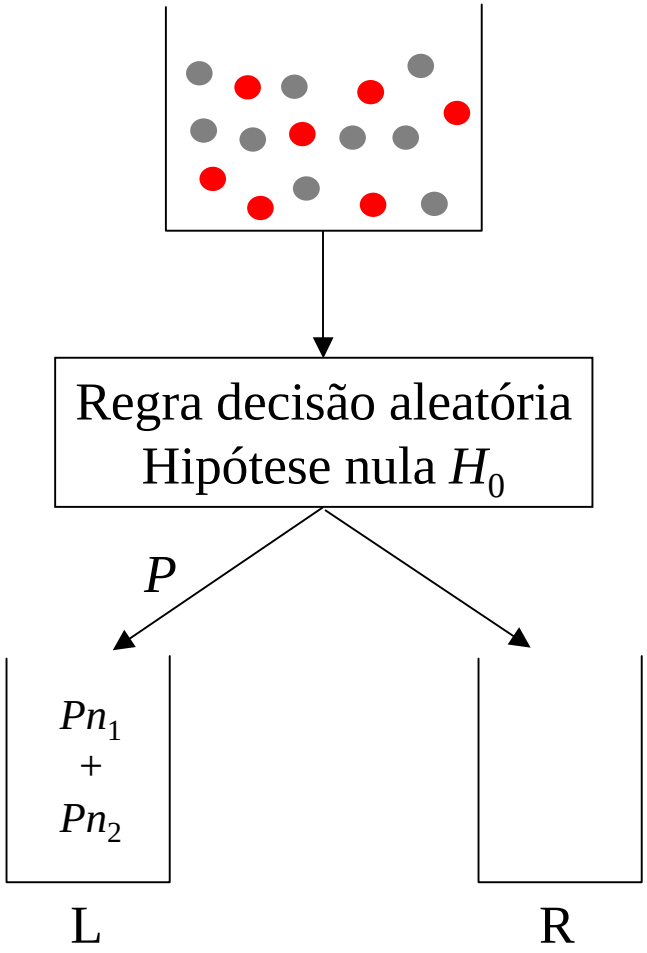
- Exemplo: suponha que D_0 seja $p(x) \sim N(0,1)$ e $x = 0.3$

  $$\Pr[|x - \mu| \leq \sigma] \approx 0.68$$

  é provável que $x = 0.3$ tenha sido gerado por D_0 ?
  e se $x = 5$?

  H_0: amostra é proveniente de uma Gaussiana com $\mu = 0$

- Teste Qui-quadrado



- classe ω_1 (n_1)
- classe ω_2 (n_2)

$n_1 + n_2 = n$

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_{iL} - n_{ie})^2}{n_{ie}}$$

$n_{iL} = \text{número padrões classe } \omega_i \text{ colocadas em } L \text{ pela regra candidata}$

$n_{ie} = Pn_i$

$\chi^2 \geq 0$

$\chi^2_{\alpha,df} > v_{cri}$ hipótese nula é rejeitada

$\alpha = \text{nível de significância;} \quad df = \text{graus de liberdade}$

Valores críticos Qui-quadrado para $\alpha = 0.05$ e 0.02

df	.05	.01	*df*	.05	.01	*df*	.05	.01
1	3.84	6.64	11	19.68	24.72	21	32.67	38.93
2	5.99	9.21	12	21.03	26.22	22	33.92	40.29
3	7.82	11.34	13	22.36	27.69	23	35.17	41.64
4	9.49	13.28	14	23.68	29.14	24	36.42	42.98
5	11.07	15.09	15	25.00	30.58	25	37.65	44.31
6	12.59	16.81	16	26.30	32.00	26	38.88	45.64
7	14.07	18.48	17	27.59	33.41	27	40.11	46.96
8	15.51	20.09	18	28.87	34.80	28	41.34	48.28
9	16.92	21.67	19	30.14	37.57	29	42.56	49.59
10	18.31	23.21	20	31.41	37.57	30	43.77	50.89

- Exemplo teste Qui-quadrado

Em 180 lances de um par de dados, foram observados 37 *setes* e 12 *três*. Testar a hipótese do dado ser honesto, para um nível de significância $\alpha = 0.05$

H_0: dado é honesto

H_1: dado não é honesto

$\alpha = 0.05$

$df = k - 1 = 2 - 1 = 1$ (k = número classes/valores observados)

Eventos	*sete*	*três*
n_{iL} (observado)	37	12
n_{ie} (esperado)	30	10

$P(sete) = 6/36 = 1/6 \quad \rightarrow n_{ie}(\text{sete}) = 180(1/6) = 30$

$P(três) = 2/36 = 1/18 \rightarrow n_{ie}(\text{três}) = 180(1/18) = 10$

$$\chi^2 = \frac{(37-30)^2}{30} + \frac{(12-10)^2}{10} = \frac{49}{30} + \frac{4}{10} = 2.03$$

$$\chi^2 < v_{crit} = 3.84$$

Aceita-se H_0

6-Teoria da informação

- Entropia

$\{v_1, v_2, ..., v_m\}$ conjunto de símbolos

$\{P_1, P_2, ..., P_m\}$ probabilidades associadas

$$H = -\sum_{i=1}^{m} P_i \log_2 P_i \quad \text{bits}$$

$$H(f(x)) \leq H(x)$$

$$H = -\int_{-\infty}^{+\infty} p(x) \ln p(x) dx$$

$$H = 0.5 + \log_2(\sqrt{2\pi}\sigma) \quad p/\text{Gaussiana}$$

$$\sigma \to 0$$

$$\delta(x-a) = \begin{cases} 0 & x \neq a \\ \infty & x = a \end{cases} \qquad \int_{-\infty}^{\infty} \delta(x) dx = 1$$

- Entropia relativa

variável aleatória x

distribuições de probabilidades $p(x)$ e $q(x)$

$$D_{KL}(p(x),q(x)) = \sum_x q(x) \ln \frac{q(x)}{p(x)}$$

Distância de Kullback-Leibler

$$D_{KL}(p(x),q(x)) = \int_{-\infty}^{\infty} q(x) \ln \frac{q(x)}{p(x)} dx$$

distância entre distribuições de probabilidade

- Informação mútua

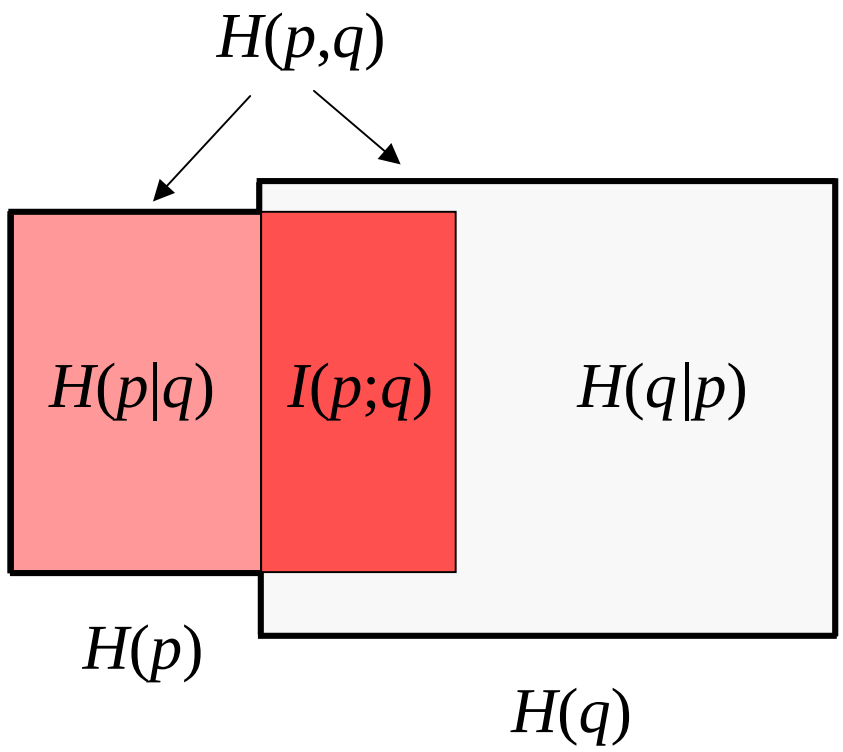
variáveis aleatórias diferentes x e y

distribuições de probabilidades $p(x)$ e $q(y)$

$$I(p;q) = H(p) - H(p \mid q) = \sum_{x,y} r(x,y) \log_2 \frac{r(x,y)}{p(x)q(y)}$$

redução da incerteza sobre uma variável devido ao conhecimento de outra variável

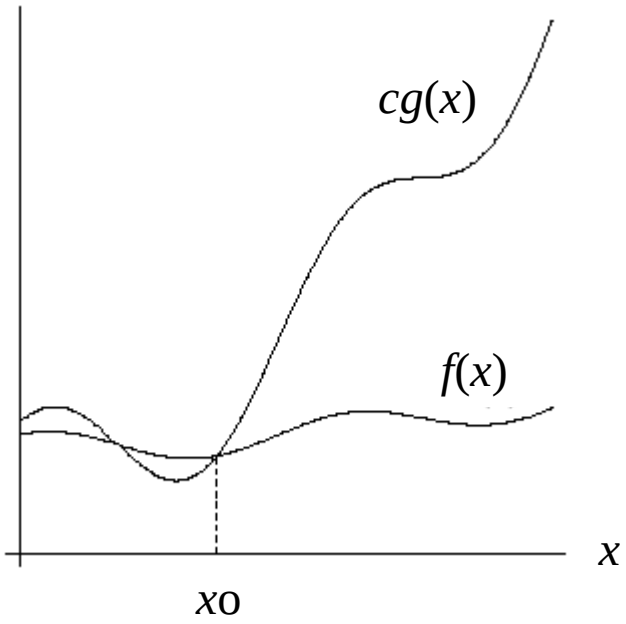
- Relações entre entropia, entropia relativa e informação mútua



7-Complexidade computacional

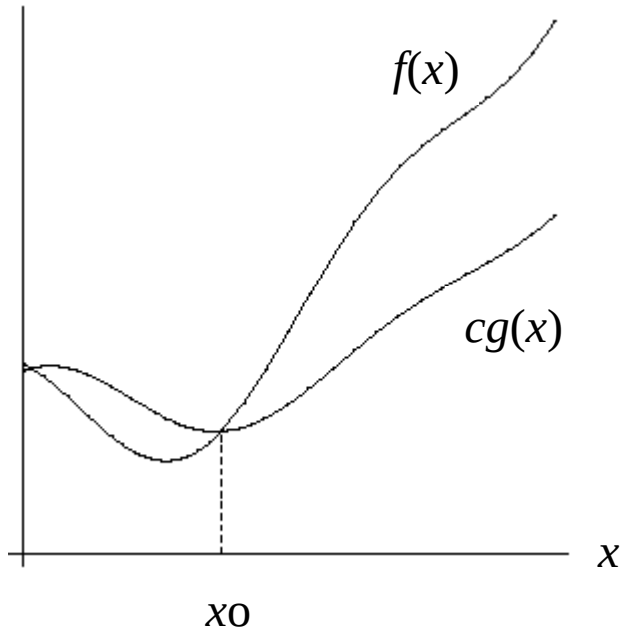
- Limitante superior assintótico

$$O(g(x)) = \{ f(x) : \exists\, c > 0 \text{ e } x_o > 0 \text{ tal que } 0 \leq f(x) \leq cg(x),\ \forall x \geq x_o \}$$



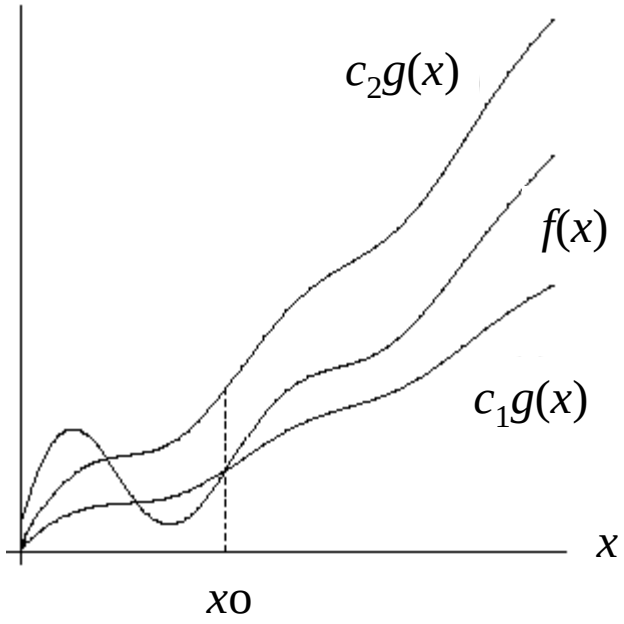
- Limitante inferior assintótico

$\Omega(g(x)) = \{ f(x) : \exists\, c > 0 \text{ e } x_o > 0 \text{ tal que } 0 \leq cg(x) \leq f(x),\ \forall x \geq x_o\}$



- Limitantes assintóticos

$\Theta(g(x)) = \{ f(x) : \exists\ c_1, c_2 > 0 \text{ e } x_o > 0 \text{ tal que } 0 \leq c_1 g(x) \leq f(x) \leq c_2 g(x)\ ,\ \forall x \geq x_o \}$



Observação

Este material refere-se às notas de aula do curso CT 720 Tópicos Especiais em Aprendizagem de Máquina e Classificação de Padrões da Faculdade de Engenharia Elétrica e de Computação da Unicamp e do Centro Federal de Educação Tecnológica do Estado de Minas Gerais. Não substitui o livro texto, as referências recomendadas e nem as aulas expositivas. Este material não pode ser reproduzido sem autorização prévia dos autores. Quando autorizado, seu uso é exclusivo para atividades de ensino e pesquisa em instituições sem fins lucrativos.