ACH2016 - Inteligência Artificial
Aula 05 - Maximização de Verossimilhança

Valdinei Freire da Silva

valdinei.freire@usp.br - Bloco A1 100-O

Estimador de Máxima Verossimilhança

Considere a p.d.f (probability density function) conjunta $f_n(\mathrm{x}|\theta)$. Se essa função é interpretada como uma função de θ com parâmetros $\mathrm{x} = (x_1, \ldots, x_n)$, então ela é chamada de função de Verossimilhança (likelihood) e é denotada por $L(\theta; \mathrm{x})$.

Suponha que as n variáveis aleatórias $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição para qual a p.d.f. condicional é $f(X|\theta)$. Então:

$$L(\theta; \mathrm{x}) = f(x_1|\theta) f(x_2|\theta) \cdots f(x_{n-1}|\theta) f(x_n|\theta).$$

Para cada possível vetor de observação $\mathrm{x} = (x_1, \ldots, x_n)$, defina $\hat{\theta} = \arg\max_{\theta \in \Omega} L(\theta; \mathrm{x})$. A estimativa $\hat{\theta}$ é a estimativa de máxima verossimilhança (M.L.E. - maximum likelihood estimator).

Função Log-likelihood

Seja $\hat{\theta}$ o M.L.E. de θ, se $g : \mathbb{R} \to \mathbb{R}$ é uma função estritamente crescente, então $\hat{\theta} = \arg\max_{\theta \in \Omega} g[L(\theta; x)]$.

Para encontrar o M.L.E. usualmente considera-se a transformação $\ell(\theta; x) = \log L(\theta; x)$ e resolve-se a seguinte equação:

$$\nabla_\theta \ell(\theta; x) = 0.$$

O estimador M.L.E. não necessariamente é único e também pode não existir dependendo da classe de distribuição.

Função Log-likelihood - Caso Binomial

Considera-se uma variável aleatória $Y \in \{0, 1\}$ condicionada em X.

Se temos N amostras $e_i = (x_i, y_i)$, e uma função hipótese $h_w(x)$ tal que:

$$\Pr(Y = 1|X = x, w) = h_w(x) \qquad e \qquad \Pr(Y = 0|X = x, w) = 1 - h_w(x)$$

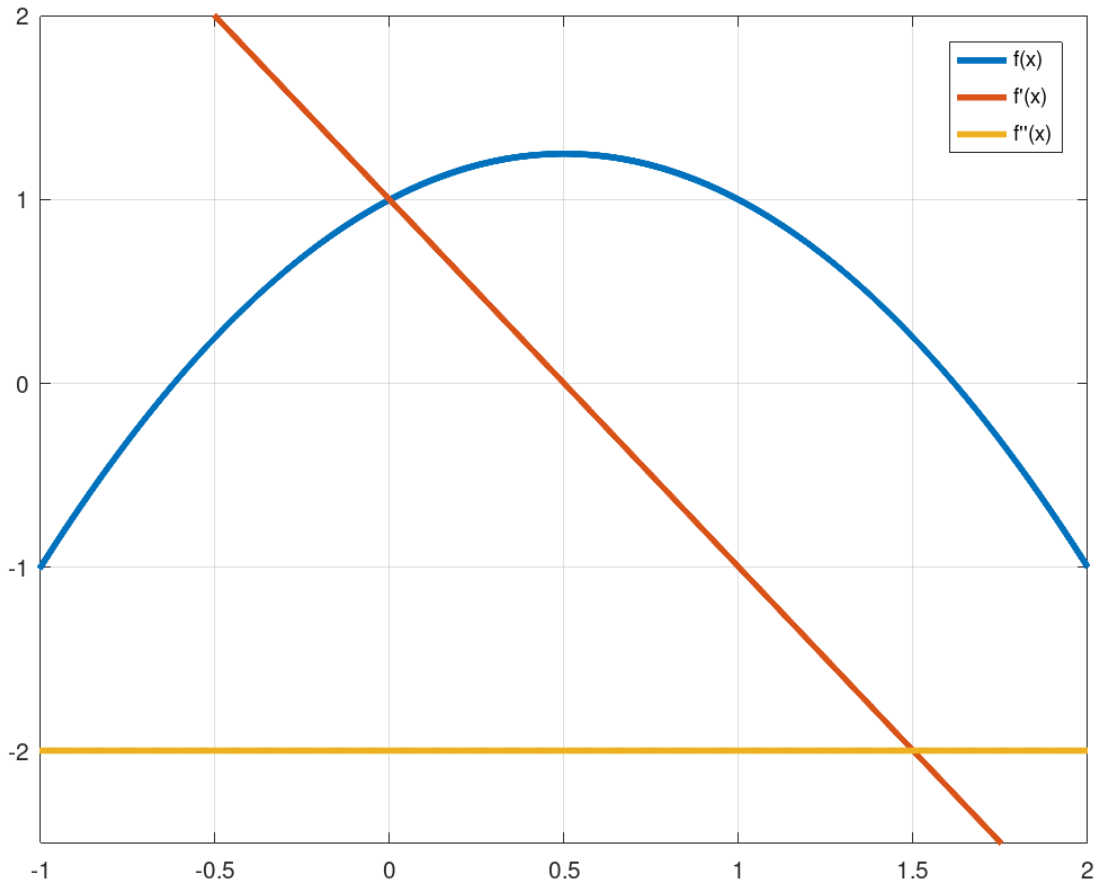
Então:

$$\begin{aligned} \ell(w; x) &= \sum_{i=1}^{N} \log \Pr(Y = y_i|X = x_i, w) \\ &= \sum_{i=1}^{N} y_i \log h_w(x_i) + (1 - y_i) \log(1 - h_w(x_i)) \end{aligned}$$

Exemplos

Exercício 1: encontro o ponto máximo para a função:

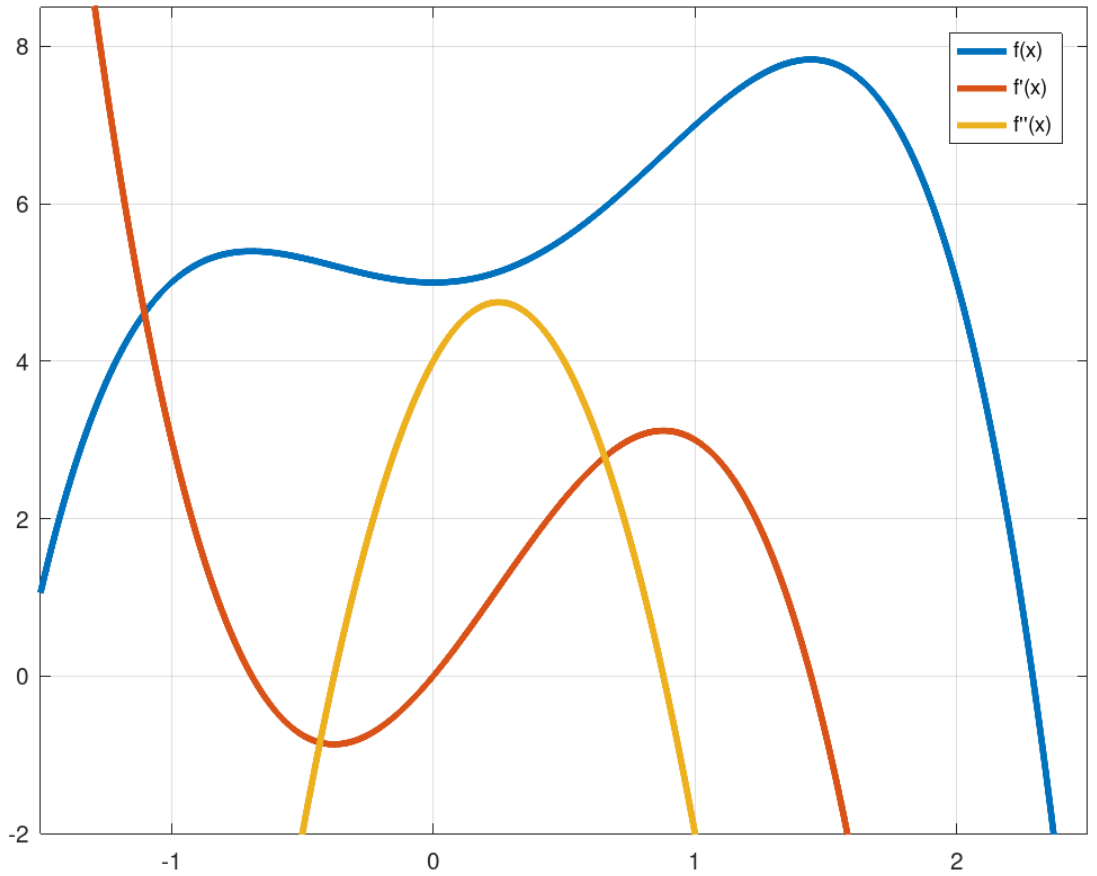
$$f(x) = -x^2 + x + 1$$



Exemplos

Exercício 2: encontro o ponto máximo para a função:

$$f(x) = -x^4 + x^3 + 2x^2 + 5$$



Otimização

Problema de Otimização

Dada uma função $g : \mathbb{R}^d \to \mathbb{R}$, encontre $x^* \in \mathbb{R}^d$ tal que $g(x^*) \geq g(x)$ para todo $x \in \mathbb{R}$.

Teorema 1. Se $g : \mathbb{R}^d \to \mathbb{R}$ é contínua e diferenciável, a solução x^* para o problema de otimização deve atender a seguinte equação:

$$\frac{\partial g(x^*)}{\partial x_i} = g'(x^*) = 0 \Leftrightarrow \nabla_x g(x^*) = 0.$$

onde x_i representa a i-ésima dimensão da entrada x.

Método de Newton

Teorema 2 (Série de Taylor). Seja $f : \mathbb{R} \to \mathbb{R}$ uma função infinitamente diferenciável definida em um intervalo aberto $(a - r, a + r)$, então:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x - a)^n,$$

onde $f^{(n)(a)}$ é a n-ésima derivada de f no ponto a.

Definition 1 (Método de Newton). Dada uma função $g : \mathbb{R} \to \mathbb{R}$ contínua e duas vezes diferenciáveis tal que $g''(x^*) \neq 0$, o método de Newton itera em valores $x^{(t)}$ seguindo:

$$x^{(t+1)} = x^{(t)} - \frac{g'(x^{(t)})}{g''(x^{(t)})}.$$

Método do Gradiente Ascendente

Definition 2 (Escalar). Dada uma função $g : \mathbb{R} \to \mathbb{R}$ contínua e diferenciável. O método do gradiente Ascendente itera nos valores $x^{(t)}$ seguindo:

$$x^{(t+1)} = x^{(t)} + \beta^{(t)} g'(x^{(t)}).$$

Definition 3 (Vetorial). Dada uma função $g : \mathbb{R}^d \to \mathbb{R}$ contínua e diferenciável. O método do gradiente Ascendente itera nos valores $\mathrm{x}^{(t)}$ seguindo:

$$\mathrm{x}^{(t+1)} = \mathrm{x}^{(t)} + \beta^{(t)} \nabla_{\mathrm{x}} g(\mathrm{x}^{(t)}).$$

Usualmente $\beta^{(t)} \to 0$.

Método do Gradiente Ascendente

1. Escolha $x^{(0)}$ arbitrário
2. Escolha $\beta^{(0)} > 0$ arbitrário
3. Enquanto não atende critério de parada

   3.1 Faça:

$$\mathrm{x}^{(t+1)} \leftarrow \mathrm{x}^{(t)} + \beta^{(t)} \nabla_{\mathrm{x}} g\left(\mathrm{x}^{(t)}\right)$$

   3.2 Se: $g\left(\mathrm{x}^{(t+1)}\right) > g\left(\mathrm{x}^{(t)}\right)$

   (a) Então:

$$\beta^{(t+1)} \leftarrow r\beta^{(t)}$$

   (b) Caso contrário:

$$\beta^{(t+1)} \leftarrow \frac{1}{r}\beta^{(t)}$$

$$\mathrm{x}^{(t+1)} \leftarrow \mathrm{x}^{(t)}$$

Critérios de Parada: gradiente mínimo, quantidade de iterações máximas

Busca de β: $r > 1$, mas existem vários outros métodos