Aprendizagem Estatística Aula 2 : 16/08/2017 (1)

Clasificação
- Sem supervisão → sequências de 0's e 1's geradas por uma fonte probabilística
- Com supervisão → último mes Deep Learning.

O que era ξ_n pasa a ser X_n

Espaço $\mathcal{X}$ de "entradas"
(conjunto) "sinais"

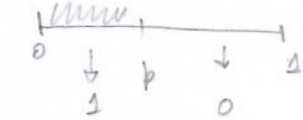
$\mathcal{Y}$ conjunto de "saídas"
"etiquetas"
"categorias"

Exemplo 1ª aula:

$\mathcal{X} = [0,1]$

$\mathcal{Y} = \{0,1\}$ clasificação binaria

$X \in [0,1]$ $Y = f(X) = \mathbb{1}_{\{X \leq p\}}$



$(X,Y) \in \mathcal{X} \times \mathcal{Y}$ são aleatórias
Não conhecemos a distribução conjunta

$\mathcal{F}$ = conjunto de funções de $\mathcal{X}$ en $\mathcal{Y}$
com certas restições

objetivo: encontrar $f^* \in \mathcal{F}$ que minimize o "risco" da classificação

$f \in \mathcal{F}$

$R(f) = \mathbb{P}(f(X) \neq Y)$ risco

No exemplo da última aula X é aleatorio e Y é uma função determinística de X.

$$P(X=x, Y=1) = \mathbb{P}(X=x, f(X)=1)$$
$$= \mathbb{P}(X=x, f(x)=1) = \mathbb{P}(X=x)\, \mathbb{1}_{\{f(x)=1\}}$$

aberta ou fechada

$\tilde{Y} = 1$ se $X \leq p$ $\qquad Y = \mathbb{1}_{\{X \leq p\}} + Z$

↑ ruido da linha

Hipótese: Ruido Z é independente da entrada X

$$\mathbb{P}(\underbrace{X \leq p}_{\text{sem ruido na classificação seria 1}}, Y=1) = \mathbb{P}(X \leq p)\, \mathbb{P}(Z=1)$$
$$= p \cdot \mathbb{P}(Z=1) = p(1-\varepsilon)$$

Vamos supor que $\mathbb{P}(Z=1) = 1-\varepsilon$ onde ε é pequeno $(\varepsilon < 1/2)$

Escolho $\mathcal{F}$ como sendo a clase de funcões da forma:

$$g : [0,1] \longrightarrow \{0,1\}$$

$$g(x) = \mathbb{1}_{\{x \leq q\}}$$

↖ parámetro

$$\mathbb{P}(Y \neq g(X)) = ?$$

Na aula anterior $\varepsilon = 0$, $\mathbb{P}(Z=1) = 1$

$$\mathbb{P}(g(X) \neq Y) = |q - p|$$

$$\inf_{g \in \mathcal{F}} P(g(X) \neq Y) = 0$$

↑ é atingido por $f^*(x) = 1_{\{x \leq p\}}$

Se $\varepsilon \in (0, 1/2)$ quero calcular

$$\mathbb{P}(g(X) \neq Y) \text{ quando } Y = \mathbb{1}_{\{X \leq p\}} \cdot Z$$

$g \in \mathcal{F}$

Definimos:

$$R(g) = \mathbb{P}(g(X) \neq Y)$$

$$R^* = \inf_{g \in \mathcal{F}} R(g) \qquad \text{Risco de Bayes}$$

f^* = clasificador de Bayes satifaz $\boxed{R(f^*) = R^*}$

Texto: Bousquet, Bouchieron, Lugosi
Introduction to Statistical Learning theory

Função de regressão:

$$x \in \mathcal{X} \longmapsto \eta(x) \in \mathcal{Y} = \{0,1\}$$

qualquer

$$\eta(x) = \mathbb{1}\{\mathbb{P}(Y=1 \mid X=x) > 1/2\}$$

Teorema
$\eta(x)$ = clasificador de Bayes

Exemplo

$\mathcal{X} = [0,1]$

$X \in [0,1]$ e tem distribuição uniforme

isto é $\mathbb{P}(X \leq r) = r$

$Z \in \{0,1\}$ e $\mathbb{P}(Z=1) = 1-\varepsilon$, onde $0 \leq \varepsilon \leq 1$

Vamos supor que X, Z são independentes

Defino $Y = \mathbb{1}_{\{X \leq p\}} \cdot Z$

Exercício: Calcular

$$\eta(x) = \mathbb{1}_{\{\mathbb{P}(Y=1 \mid X=x) > 1/2\}}$$

Na prática NÃO conhecemos a lei conjunta de X e Y e portanto não temos como calcular $R^* = \inf_{g \in \mathcal{F}} \mathbb{P}(g(X) \neq Y)$

nem $\eta(x) = \mathbb{1}\{\mathbb{P}(Y=1 \mid X=x) > 1/2\}$

Porem temos amostras $(X_1, Y_1), \dots, (X_n, Y_n)$ com n sorteios independentes de (X, Y)

Dado $g \in \mathcal{F}$, calculo $\hat{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(X_i) \neq Y_i\}$

Vimos na última aula no caso $\varepsilon = 0$

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i = 1\} \xrightarrow{n \to \infty} p$$

parâmetro correto

Aplicação da ley dos grandes números.

Então se chamamos $\hat{f}_n(x) = \mathbb{1}\{x \leq \hat{p}_n\}$ então

$$\hat{f}_n \xrightarrow[n \to \infty]{} f(x) = \mathbb{1}\{x \leq p\}$$

$$\mathbb{P}\left(\left|\hat{R}_n(g) - R(g)\right| > \varepsilon\right) = \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^{n} \overbrace{\mathbb{1}\{g(X_i) \neq Y_i\}}^{Z_i \in \{0,1\}} - \underbrace{\mathbb{P}(g(X) \neq Y)}_{\mathbb{E}(\mathbb{1}\{g(X) \neq Y\}) = \mathbb{E}(Z)}\right| > \varepsilon\right)$$

$$= \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^{n} Z_i - \mathbb{E}(Z)\right| > \varepsilon\right) \leq 2 e^{-2n\varepsilon^2}$$

v.a iid $Z_i \in \{0,1\}$

Hoeffdings.

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^{n} Z_i - \mathbb{E}(Z)\right| > \varepsilon\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i - n\mathbb{E}(Z)\right| > n\varepsilon\right)$$

$$= \mathbb{P}\left(\left\{\sum_{i=1}^{n} Z_i > n\left[\mathbb{E}(Z) + \varepsilon\right]\right\} \cup \left\{\sum_{i=1}^{n} Z_i < n\left[\mathbb{E}(Z) - \varepsilon\right]\right\}\right)$$

$$= \underbrace{\mathbb{P}\left(\sum_{i=1}^{n} Z_i > n\left[\mathbb{E}(Z)+\varepsilon\right] \right)} + \mathbb{P}\left(\sum_{i=1}^{n} Z_i < n\left[\mathbb{E}(Z)-\varepsilon\right] \right)$$

$$= \mathbb{P}\left(e^{\lambda \sum_{i=1}^{n} Z_i} > e^{\lambda n\left[\mathbb{E}(Z)+\varepsilon\right]} \right)$$

$r \mapsto e^{\lambda r}$

função crescente

→ método de Cramer-Chernoff ~ 1940

Desigualdade de Markov ~ 1905

v.a W positiva

$$\mathbb{P}(W \geq a) \leq \frac{\mathbb{E}(W)}{a}$$

$$W = e^{\lambda \sum_{i=1}^{n} Z_i}, \quad a = e^{\lambda n(\mathbb{E}(Z)+\varepsilon)}$$

→ usando Markov segue que:

$$\mathbb{P}\left(e^{\lambda \sum_{i=1}^{n} Z_i} > e^{\lambda n\left[\mathbb{E}(Z)+\varepsilon\right]} \right) \leq \frac{\mathbb{E}\left(e^{\lambda \sum_{i=1}^{n} Z_i}\right)}{e^{\lambda n(\mathbb{E}(Z)+\varepsilon)}}$$

$$= \frac{\prod_{i=1}^{n} \mathbb{E}\left(e^{\lambda Z_i}\right)}{e^{\lambda n(\mathbb{E}(Z)+\varepsilon)}} = \frac{\mathbb{E}\left(e^{\lambda Z}\right)^n}{e^{\lambda n(\mathbb{E}(Z)+\varepsilon)}}$$

isso vale para todo $\lambda > 0$

Final da conta: achar "o melhor" λ para essa majoração

Hoeffding diz que:

$$\mathbb{P}\left(|\hat{R}_n(g) - R(g)| > \varepsilon \right) \leq \underbrace{2e^{-2n\varepsilon^2}}_{\delta}$$

Eu quero $\delta = 2e^{-2n\varepsilon^2}$

$$\frac{\delta}{2} = e^{-2n\varepsilon^2}$$

$$\ln\frac{\delta}{2} = -2n\varepsilon^2, \quad -\frac{1}{2n}\ln\frac{\delta}{2} = \varepsilon^2, \quad \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}} = \varepsilon$$

Se queremos que a probabilidade de erro seja δ, tomamos $\varepsilon = \sqrt{\dfrac{\ln 2/\delta}{2n}}$

Isto acontece

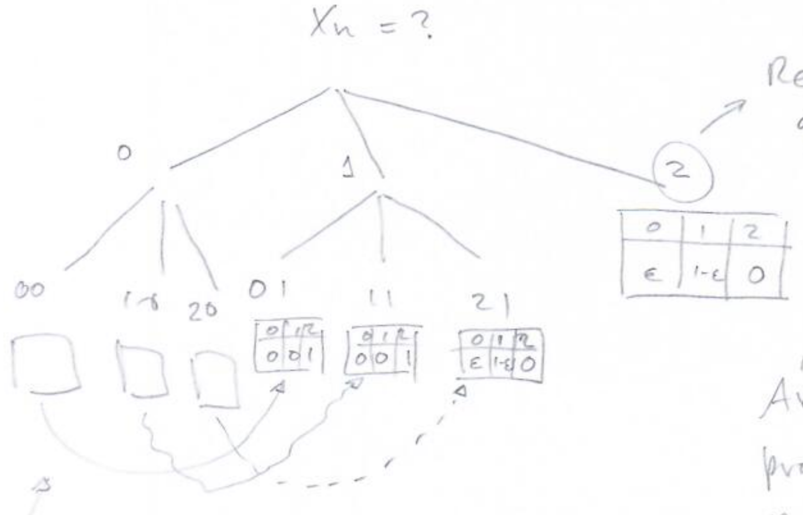
$$R(g) \leq \hat{R}_n(g) + \sqrt{\frac{\ln 2/\delta}{2n}} \quad \text{com probabilidade} \geq 1-\delta$$

Jogo do Goleiro:

Batedor: Começa com a sequência deterministica

211 211 211 ...

cada símbolo 1 pode ser ou transformado em 0 com probabilidade ε ou mantido com prob $1-\varepsilon$.
Cada uma das escolhas é feita independentemente das anteriores.



Representa o conjunto das sequenças que acabam com o símbolo 2.

Temos
Árvore, uma família de probabilidades de transição indexada pelas folhas da árvore

usamos

Algoritmo:

1. Como começar?
2. Como scolher próximo paso.

Árvore τ define uma partição no conjunto de todas as sequências de símbolos passados.

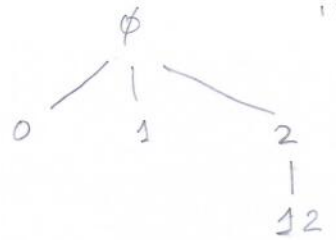
O que é Árvore: 1ª resposta: Grafo sem ciclos
com etiqueta. Grafo orientado (por laço de hereditariedade)

Alfabeto $A = \{0, 1, 2\}$

Vamos representar τ por suas folhas

$\tau = \{2, 01, 11, 21, 00, 10, 20\}$

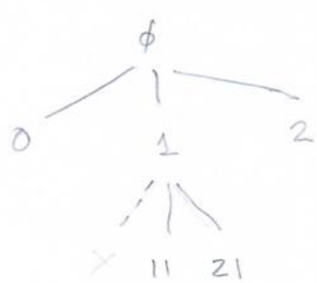
Outro exemplo, $\tau' = \{\underbrace{2, 12}_{\text{imcompativel}}, 0, 1\}$



Nenhum elemento de τ pode ser sufixo propio de outro.

Candidato a árvore

$\tilde{\tau} = \{0, 11, 21, 2\}$



Problema E se a sequencia terminar com 01?

Ela pode ser usada só para algoritmos que nunca genam pares 01

Exemplo (exercício)

1. Tomo a sequencia periodica

   2101 2101 2101 ...

2. Transformo símbolos 1 em cero com prob ε e mantenho 1 com prob. $1-\varepsilon$. Faço isso de maneira iid

τ árvore que define uma partição de todos os pasados que podem aparecer no arquivo.

Represento τ por as folhas, CONTEXTO = FOLHA

$(x_{-m}, \dots, x_{-2}, x_{-1}) = x_{-m}^{-1}$

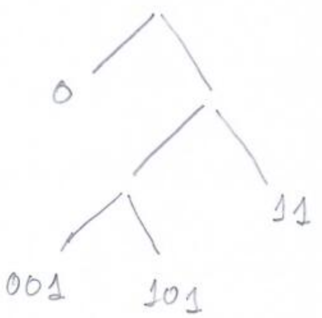
$\dots x_{-m}, x_{-(m-1)}, \dots, x_{-1} = x_{-\infty}^{-1}$

$c_\tau : x_{-\infty}^{-1} \longmapsto c_\tau(x_{-\infty}^{-1})$ que é o unico sufixo de $x_{-\infty}^{-1}$ que pertence a τ

Goleiro

Escolhe o próximo símbolo usando a probabilidade de transiçao asociada ao contexto que termina no símbolo anterior.

Cómo começar: Começo com um contexto de comprimento máximo



$\ell(x_{-k}^{-1}) = k$

comprimento da sequencia

Altura da árvore

$\max\{\ell(w) : w \in \tau\} = h$

comprimento máxima

· Associamos uma árvore a uma sequencia $x_0, x_1, \ldots, x_n$