Getting to the Basics: Dimensão VC e Generalização em Machine Learning

Peng Yaohao e Mateus Hiro Nagata

LAMFO

Outline

Framework da Aprendizagem

Machine Learning: queremos uma resposta

1. Existe função ideal:variáveis explicativas $\rightarrow$ resposta.
2. Disponibilidade dos dados: temos dados que informam tanto as variáveis explicativas como sua respectiva resposta.
3. Objetivo: Usar certas hipóteses e escolher um algoritmo que aproxima àquela função ideal

A Prova

1. Função ideal
2. Disponibilidade dos dados
3. Objetivo

Tal como temos várias questões de provas anteriores e suas respostas. Precisamos APRENDER o padrão e GENERALIZÁ-lo para a prova. Essa nos dá perguntas nunca vistas antes, mas aprendemos o padrão. O âmago da questão é sabermos responder as perguntas novas da prova.

A Matemática da Aprendizagem

- Função ideal f:$\mathcal{X} \to \mathcal{Y}$
- Dados de treinamento $\mathcal{D} = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$
- Função aprendida h:$\mathcal{X} \to \mathcal{Y}$
- Queremos f $\approx$ h nos dados de treinamento (good fitting) $\Leftrightarrow$ $E_{in}(h) \approx 0$
- Queremos f $\approx$ h fora dos dados de treinamento (generalização) $\Leftrightarrow$ $E_{out}(h) \approx E_{out}(f)$

Desafios

Desafios

1. Dados com ruído
2. Amostra não representa a população
3. Algoritmo não generaliza bem

Desafios

Desafios

1. Dados com ruído $\rightarrow$ Temos que lidar
2. Amostra não representa a população $\rightarrow$ Estatística!
3. Algoritmo não generaliza bem $\rightarrow$ Overfitting

Outline

Inferindo sobre o Inexplorado

Dados Ruins: amostra que informa muito pouco sobre a população

- Precisamos de uma garantia
- $E_{\text{in}}(h) =$ Erro da função h dentro dos dados de treino
- $E_{\text{out}}(h) =$ Erro da função h fora
- $\epsilon =$ Tolerância do erro
- N = Tamanho amostral

CUIDADO

ATENÇÃO! O PRÓXIMO SLIDE CONTÉM MATEMÁTICA

Desigualdade de Hoeffding

$$\mathbb{P}\left[\left|E_{\text{in}}\left(h\right)-E_{\text{out}}\left(h\right)\right|>\epsilon\right]\leq 2\exp\left(-2\epsilon^{2}N\right)$$

Desigualdade de Hoeffding

$$\mathbb{P}\left[\left|E_{\text{in}}\left(h\right)-E_{\text{out}}\left(h\right)\right|>\epsilon\right]\leq 2\exp\left(-2\epsilon^{2}N\right)$$

$$E_{\text{out}}(g)\leq E_{\text{in}}(g)+\sqrt{\frac{1}{2N}\ln\frac{2}{\delta}}$$

Desigualdade de Hoeffding

$$\mathbb{P}\left[\left|E_{\text{in}}\left(h\right)-E_{\text{out}}\left(h\right)\right|>\epsilon\right]\leq 2\exp\left(-2\epsilon^{2}N\right)$$

"Avaliando uma hipótese, quando o tamanho amostral N aumenta, torna-se exponencialmente improvavél que $E_{in}(h)$ e $E_{out}(h)$ se distem mais que ϵ"

Analogia da Prova

Teste

$$\mathbb{P}\left[|E_{\text{in}} - E_{\text{out}}| > \epsilon\right] \leq 2\exp\left(-2\epsilon^2 N\right)$$

- E_{in} é o quão bem você foi na prova
- E_{out} o quão bem você entendeu o conteúdo (generalizou)
- Quanto mais questões na prova (N), mais próximo

Analogia da Prova

Teste

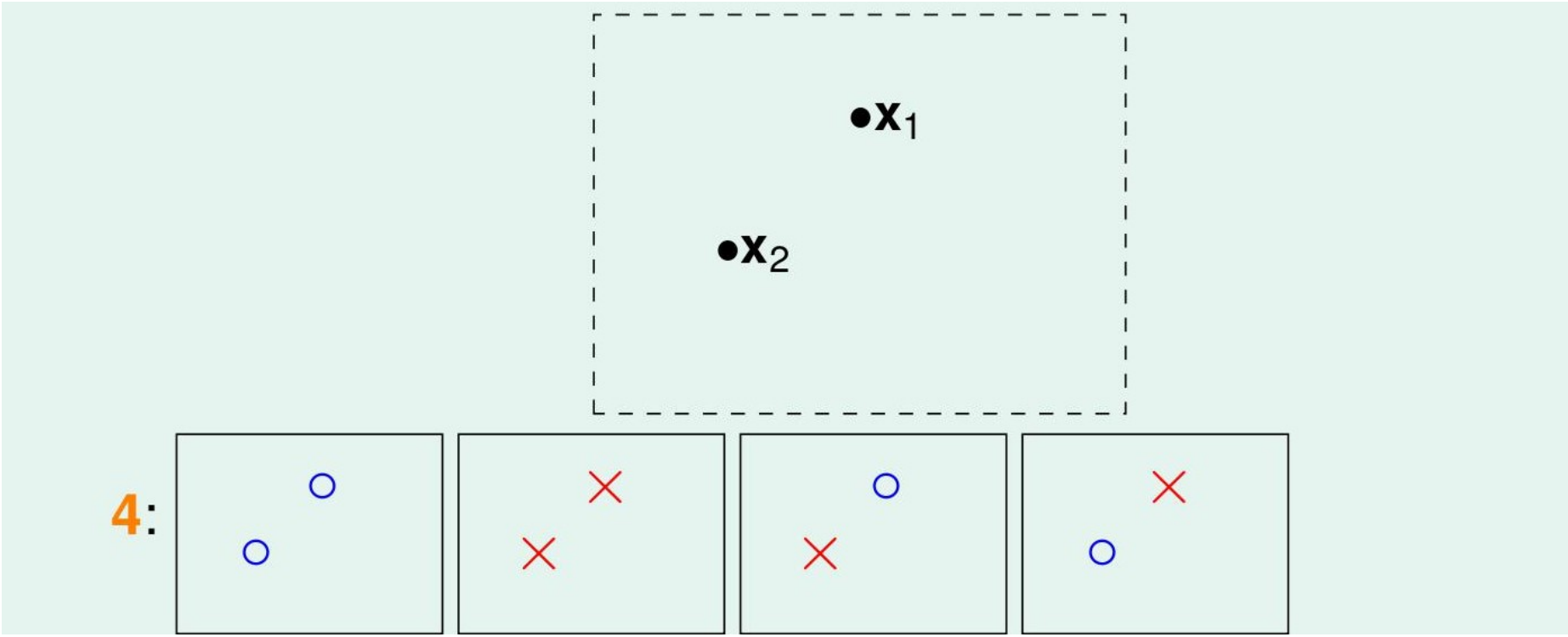
$$\mathbb{P}\left[|E_{\text{in}} - E_{\text{out}}| > \epsilon\right] \leq 2\exp\left(-2\epsilon^2 N\right)$$

Treino

$$\mathbb{P}\left[|E_{\text{in}} - E_{\text{out}}| > \epsilon\right] \leq 2 \cdot M \cdot \exp\left(-2\epsilon^2 N\right)$$

- E_{in} é o quão bem você foi nos treinos
- E_{out} o quão bem você entendeu o conteúdo
- Treino contaminado! Memorizou algumas questões, então discrepância entre resultado e conteúdo é maior que no teste
- Preço pago = O quanto você explorou! Quantidade de hipóteses que são possíveis M!

Dicotomias 2^N



Problemas Binários

Quantidade de possíveis resultados: 2^N
Dicotomias:

$$\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \{(h(\mathbf{x}_1), \cdots, h(\mathbf{x}_N)) \,|h \in \mathcal{H}\}$$

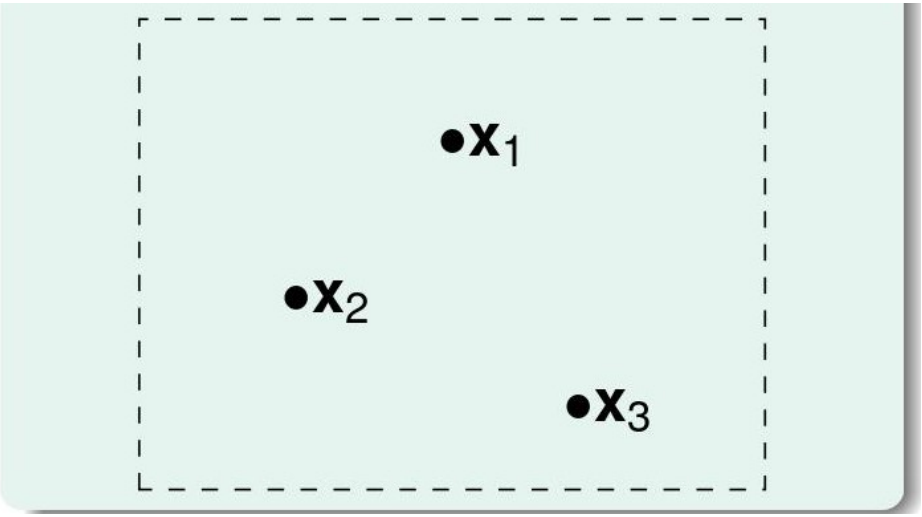
Growth Function:

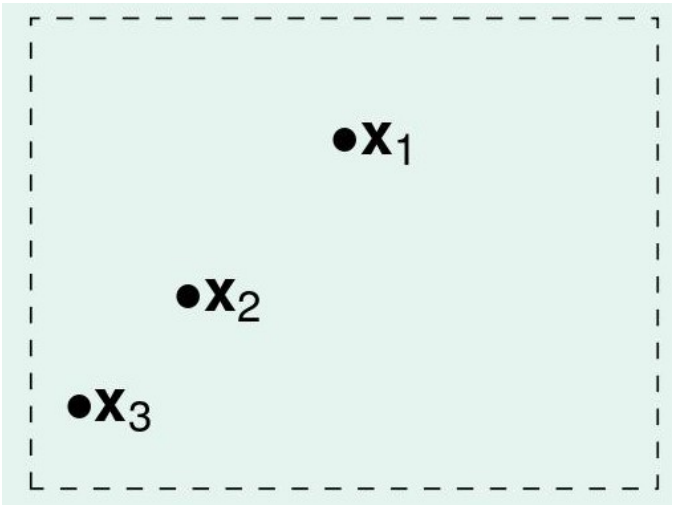
$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)|$$

Número Máximo de Dicotomias:

$$m_{\mathcal{H}}(N) \leq 2^N$$

Dicotomias em Perceptron





Outline

Dimensão VC

Definição. A **Dimensão VC** de um conjunto de hipóteses H, escrito d_{vc}, é o maior valor de N que $m_{\mathcal{H}}(N) = 2^N$.

- Quantidade de bolinhas que a gente pode usar sem criar dicotomias impossíveis

Exemplos

Examples

- $\mathcal{H}$ is positive rays:

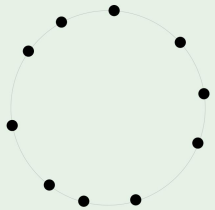
$$d_{\text{VC}} = 1$$

- $\mathcal{H}$ is 2D perceptrons:

$$d_{\text{VC}} = 3$$

- $\mathcal{H}$ is convex sets:

$$d_{\text{VC}} = \infty$$



O Teorema Mais Importante da Aprendizagem Estatística

- $m_{\mathcal{H}}(N) \leq N^{d_{\text{vc}}} + 1$

Teorema. Para qualquer tolerância $\delta > 0$,

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

com probabilidade $\geq 1 - \delta$.
Então, com dados suficientes, toda e qualquer hipótese no $\mathcal{H}$ infinito com dimensão VC finita vai generalizar.

Dimensão VC

d_{vc} **finito** $\Rightarrow$ função aprendida g vai generalizar!!

- Independente do algoritmo
- Independente da distribuição
- Independente da função ideal

Desigualdade de Hoeffding Atualizada

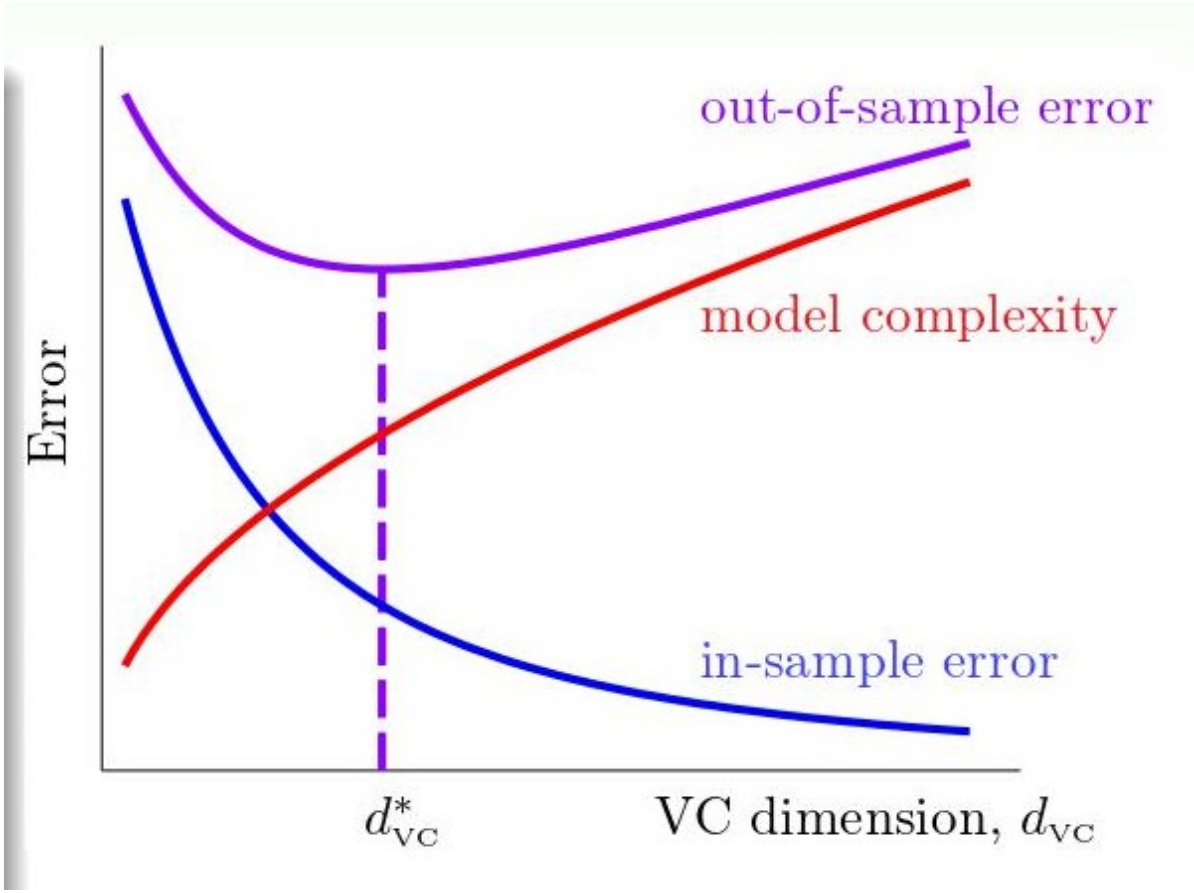
$$\mathbb{P}\left[|E_{\text{in}} - E_{\text{out}}| > \epsilon\right] \leq 2Me^{-2\epsilon^2 N}$$

$$\Downarrow$$

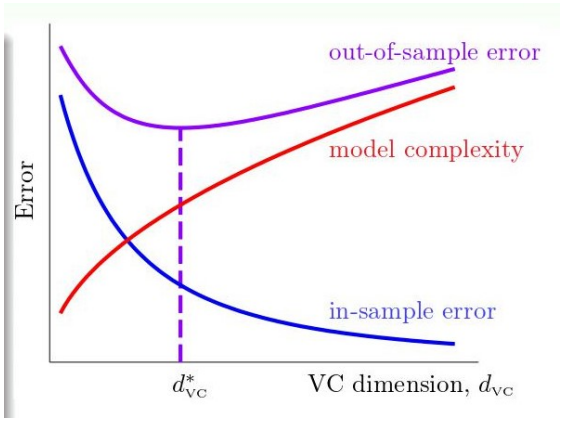
$$\mathbb{P}\left[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\right] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

Outline

Dilema Viés-Variância

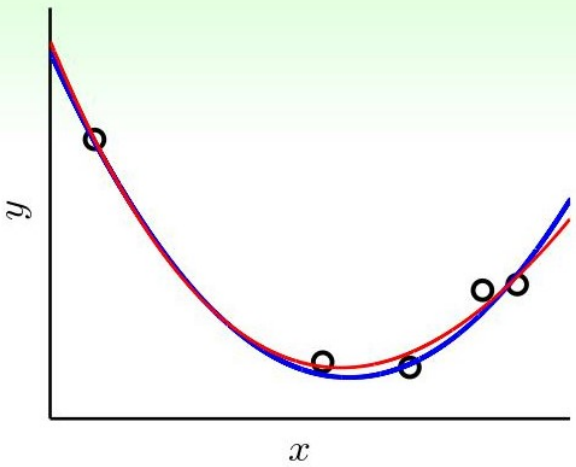


Dilema Viés-Variância



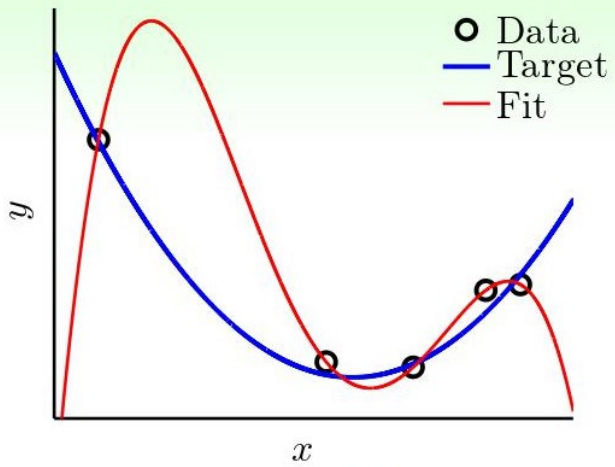
1. Modelo complexo $(\uparrow d_{vc}) \rightarrow E_{in}(g) \approx 0$
2. Modelo simples $(\downarrow d_{vc}) \rightarrow E_{in}(g) \approx E_{out}(g)$
3. O bom seria um nível intermediário que resulta em mínimo erro no dado teste

O Bom Intermediário



'good fit'

$\Longrightarrow$



overfit

O Pavor do Overfitting

Quantidade de Dados	↑	Overfitting	↓
Ruído	↑	Overfitting	↑
Complexidade Alvo	↑	Overfitting	↑

- Soluções: Bagging, Boosting, Regularization
- Validação
- Feature Transform
- Começar com modelo simples e ir aumentando a complexidade

Getting to the Basics: Dimensão VC e Generalização em Machine Learning

Peng Yaohao e Mateus Hiro Nagata

LAMFO