Universidade Federal da Paraíba

Coordenação do Curso de Ciência de Dados e Inteligência Artificial

Teoria da Generalização

Prof. Dr. Bruno Pessoa

Roteiro

- Erro de generalização
- Dicotomias
- Função de crescimento
- Break point
- Dimensão VC
- Limitante de generalização VC
- Regras de ouro

Treinamento versus Teste

- E_{in} é uma medida de performance voltada para os dados de treinamento.

- $\boldsymbol{E_{out}}$ mede a capacidade de um modelo de ML de generalizar a aprendizagem obtida na fase de treinamento.

Há como relacionar tais métricas a fim de se obter limitantes para $\boldsymbol{E_{out}}$?

Erro de generalização

- A desigualdade de Hoeffding provê uma forma de caracterizar o erro de generalização:

$$\boldsymbol{P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2Me^{-2\epsilon^2 N}}$$

Fazendo $A = |E_{in}(g) - E_{out}(g)| > \epsilon$ e sabendo que $P(A) = 1 - P(\bar{A})$, temos que:

$$1 - P(\bar{A}) \leq 2Me^{-2\epsilon^2 N}$$

$$\boldsymbol{P(\bar{A}) \geq 1 - 2Me^{-2\epsilon^2 N}}$$

Erro de generalização

Dado que $\bar{A} = |E_{in}(g) - E_{out}(g)| \leq \epsilon$,

$$P(|E_{in}(g) - E_{out}(g)| \leq \epsilon) \geq 1 - 2Me^{-2\epsilon^2 N}.$$

Fazendo $\delta = 2Me^{-2\epsilon^2 N}$, podemos afirmar que, com probabilidade de no mínimo $1 - \delta$,

$$|E_{in}(g) - E_{out}(g)| \leq \epsilon.$$

Uma vez que $E_{out}(g) \geq E_{in}(g)$,

$$E_{out}(g) - E_{in}(g) \leq \epsilon,$$

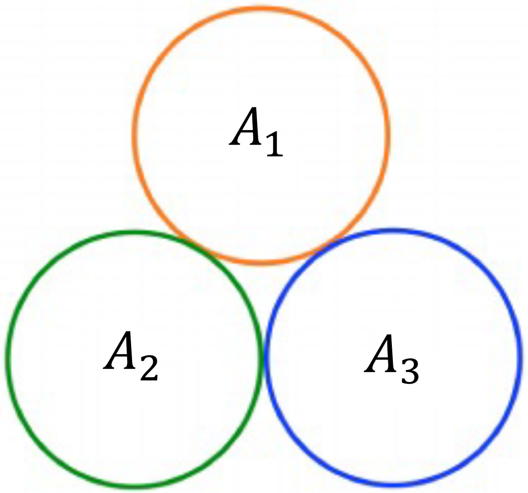
$$\boldsymbol{E_{out}(g) \leq E_{in}(g) + \underbrace{\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}}_{\epsilon}.}$$

A origem de M

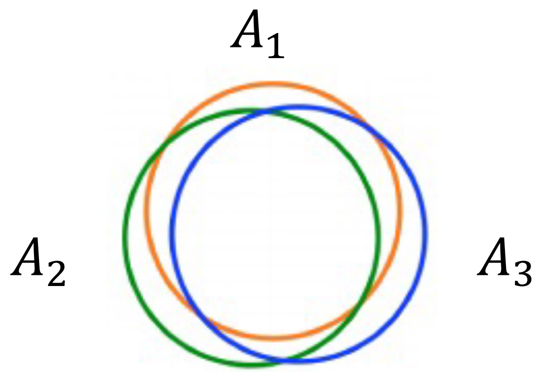
- **Probabilidade da união**

Seja $A_m = |E_{in}(h_m) - E_{out}(h_m)| > \epsilon$,

$$P(A_1 \cup A_2 \ldots \cup A_M) = \left(\sum_{i=1}^{M} P(A_i)\right) - \boldsymbol{expr}$$

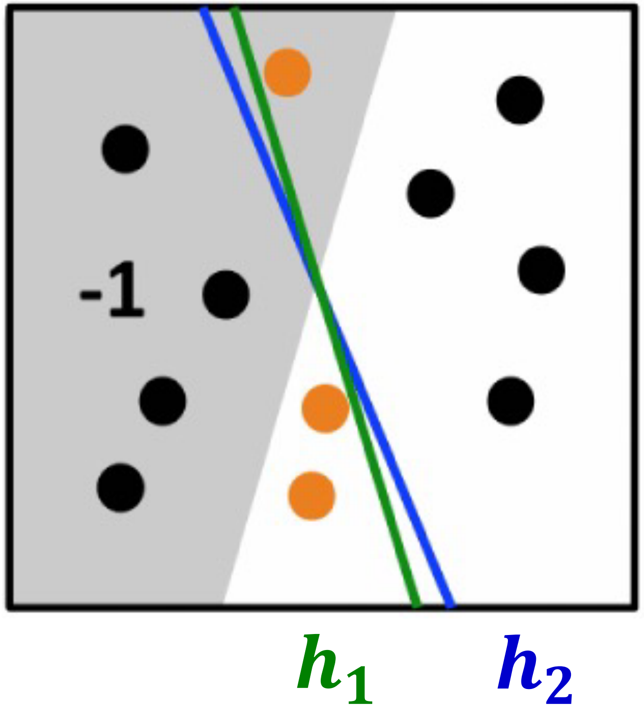


Eventos sem sobreposição



Eventos com muita sobreposição

Sobreposição dos eventos A_m

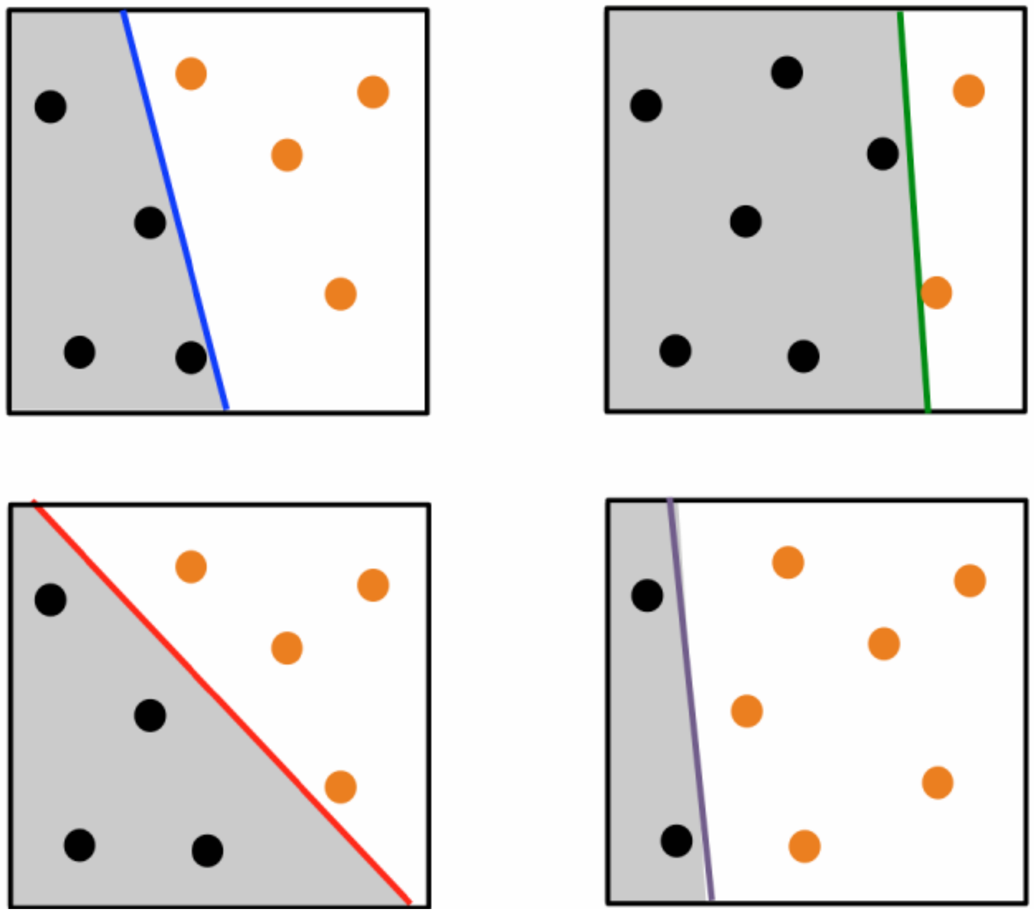


Há enorme sobreposição nos eventos $|E_{in}(\boldsymbol{h_1}) - E_{out}(\boldsymbol{h_1})| > \epsilon$ e $|E_{in}(\boldsymbol{h_2}) - E_{out}(\boldsymbol{h_2})| > \epsilon$.

Podemos reduzir o valor de M?

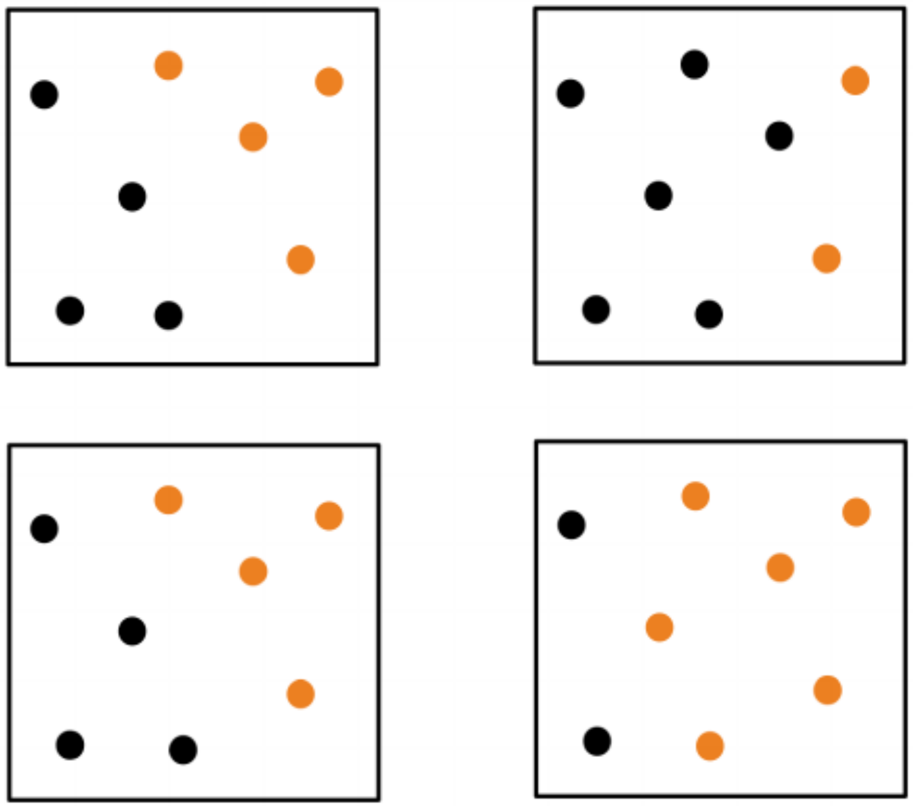
- Considere um conjunto finito de pontos.

Pontos classificados da mesma forma, indicando que há sobreposição de eventos A_m.



Podemos substituir M por um valor finito?

- De quantas maneiras podemos colorir o conjunto de dados abaixo?



Dicotomias: mini-hipóteses

- Uma **hipótese** é uma função $h: X \to \{-1, +1\}$.
- Uma **dicotomia** é uma função
  $$h: \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_N}\} \to \{-1, +1\}.$$
- Número de hipóteses $|H|$ pode ser infinito.
- Número de dicotomias $|H(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_N})|$ é no máximo $\boldsymbol{2^N}$.
- Candidato para substituir $\boldsymbol{M}$!

Função de crescimento

Seja o conjunto de dicotomias

$$H(\boldsymbol{x_1}, \dots, \boldsymbol{x_N}) = \left\{ \left(h(\boldsymbol{x_1}), \dots, h(\boldsymbol{x_n})\right) \middle| \, h \in H \right\}.$$

A função de crescimento representa o número **máximo** de dicotomias em **quaisquer *N*** pontos de X:

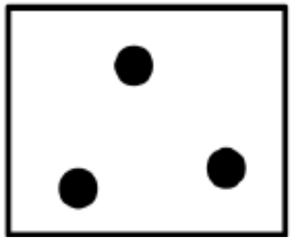
$$m_H(N) = max_{\boldsymbol{x}_1, \dots, \boldsymbol{x_n} \in X} \left| H(\boldsymbol{x_1}, \dots, \boldsymbol{x_n}) \right|$$

A função de crescimento satisfaz:

$$\boldsymbol{m_H(N) \leq 2^N}$$

$m_H(N)$ para classificadores lineares

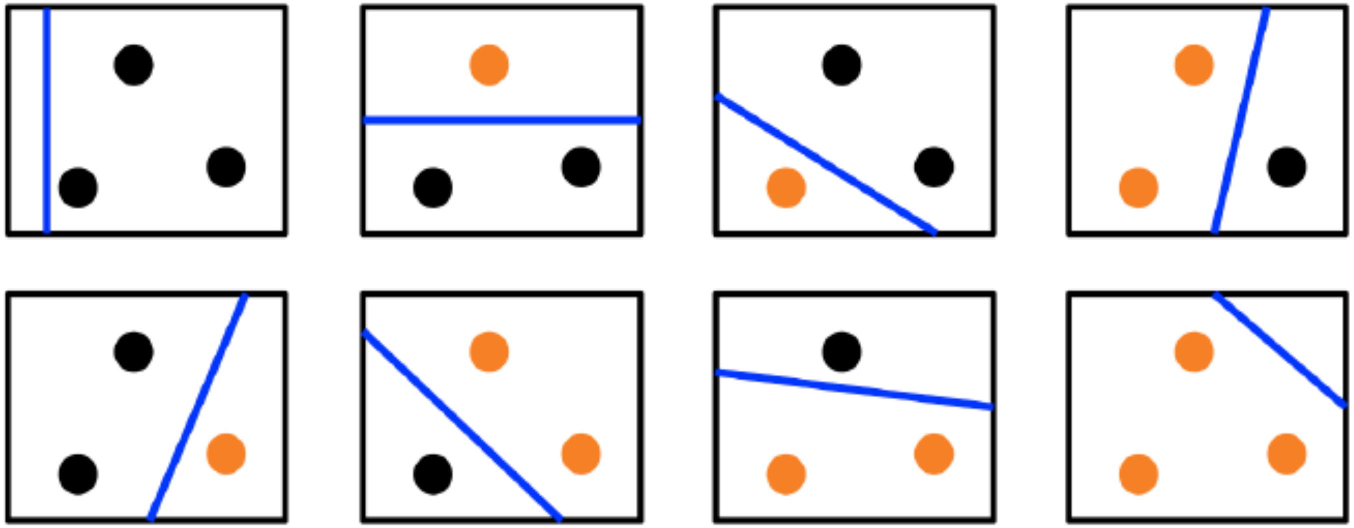
- Considere o conjunto de hipóteses H de um perceptron $2D$ e conjunto de pontos a seguir:



- Qual seria o valor de $\boldsymbol{m_H(3)}$?

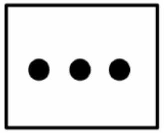
$m_H(N)$ para classificadores lineares

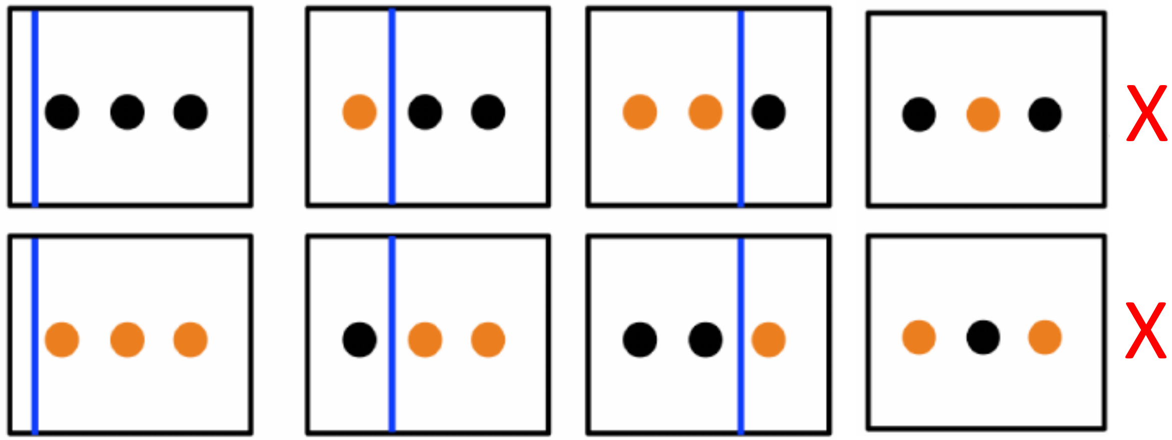
- H = Perceptron 2D



- $m_H(3) = 8$

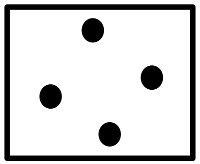
$m_H(N)$ para classificadores lineares

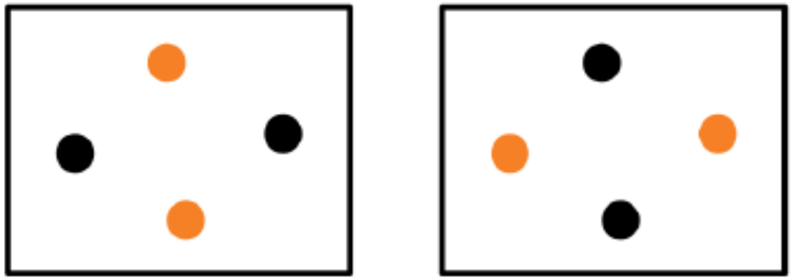
- H = Perceptron 2D
- Dados: 



- $\mathbf{m_H(3) = 8}$ pois o que importa é o número máximo de dicotomias considerando **qualquer amostra** de 3 pontos.

$m_H(N)$ para classificadores lineares

- H = Perceptron 2D
- Dados: 



- $\boldsymbol{m_H(4) \leq 14}$
- Limitante mais forte que 2^N

Pausa para reflexão

A partir de Hoeffding, temos que:

$$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2Me^{-2\epsilon^2 N}$$

$$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq \frac{2M}{e^{2\epsilon^2 N}}$$

O que acontece se substituirmos M por $m_H(N)$?

$m_H(N)$ precisa ser polinomial

Break point de um conjunto $\boldsymbol{H}$

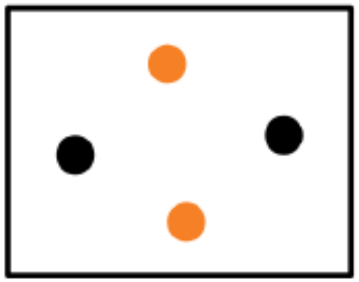
Definição

Se **nenhum** conjunto de dados (amostra) de tamanho $\boldsymbol{k}$ pode ser separado completamente por $\boldsymbol{H}$, então $\boldsymbol{k}$ é um break point para $\boldsymbol{H}$.

De outro modo,

$$\boldsymbol{m_H(k) < 2^k}.$$

Exemplo:



Para perceptrons $2D$, $\boldsymbol{k = 4}$ é um break point.

Break point de um conjunto $\boldsymbol{H}$

- **Resultado**
  - Caso não exista break point, $\boldsymbol{m_H(N) = 2^N}$.
  - Caso exista, $m_H(N)$ é **polinomial** em N.

Limitante para a função de crescimento

Teorema

Se $m_H(k) < 2^k$, para algum valor de k, então

$$m_H(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}.$$

Implicações

Como $\sum_{i=0}^{k} \binom{N}{i}$ é um polinômio em N de grau $k-1$, temos a garantia de que $m_H(N)$ é **polinomial**.

Dimensão Vapnik-Chervonenkis (VC)

- A dimensão VC de um conjunto de hipóteses H, denotada por $\boldsymbol{d_{VC}(H)}$, é o maior valor de N para o qual $m_H(N) = 2^N$.

- Em outras palavras, é o **número máximo de pontos** que pode ser separado de todas as formas por um conjunto de hipóteses H.

- Se d_{vc} é a dimensão VC de H, então $\boldsymbol{k = d_{vc} + 1}$ é um break point para $\boldsymbol{H}$.

Dimensão Vapnik-Chervonenkis (VC)

- Em termos de um break point $\boldsymbol{k}$:

$$m_H(N) \leq \sum_{i=0}^{\boldsymbol{k-1}} \binom{N}{i}$$

- Em termos da dimensão VC $\boldsymbol{d_{vc}}$:

$$m_H(N) \leq \sum_{i=0}^{d_{vc}} \binom{N}{i} \leq \boldsymbol{N^{d_{vc}} + 1}$$

Dimensão Vapnik-Chervonenkis (VC)

- **Considerações**
  1. Existe um conjunto de N pontos que pode ser separado de todas as formas por H.
     - **Conclusão: $\boldsymbol{d_{vc} \geq N}$.**

  2. Qualquer conjunto de N pontos pode ser separado de todas as formas por H.
     - **Conclusão: $\boldsymbol{d_{vc} \geq N}$.**

Dimensão Vapnik-Chervonenkis (VC)

- **Considerações**

3. Existe um conjunto de N pontos que **não** pode ser separado de todas as formas por H.
   - **Conclusão: Nenhuma.**

4. Nenhum conjunto de N pontos pode ser separado de todas as formas por H.
   - **Conclusão: $\boldsymbol{d_{vc} < N}$.**

Limitante de generalização VC

Lembremos que, com probabilidade $\geq 1 - \delta$,

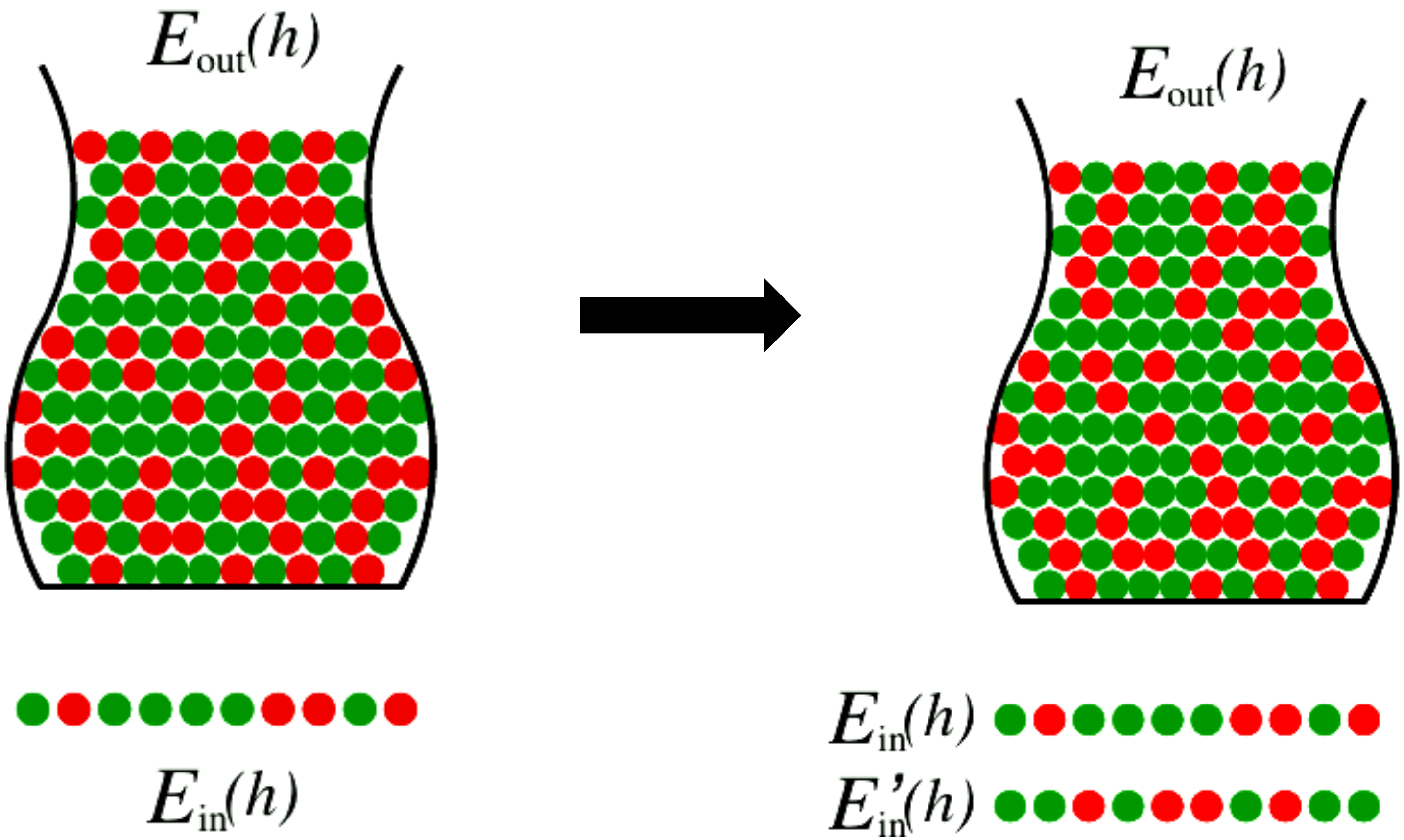
$$E_{out}(g) \leq \; E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2\boldsymbol{M}}{\delta}}.$$

Substituindo M por $m_H(N)$, obtemos:

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

Podemos mesmo substituir M por $m_H(N)$?

Limitante de generalização VC



Limitante de generalização VC

Teorema

Seja $\delta > 0$ uma métrica de tolerância, com probabilidade $\geq 1 - \delta$,

$$\boldsymbol{E'_{in}(g)} \leq E_{in}(g) + \sqrt{\frac{\mathbf{8}}{\boldsymbol{N}} \ln \frac{\mathbf{4} m_H(\mathbf{2}\boldsymbol{N})}{\delta}}$$

Limitante de generalização VC

Teorema

Seja $\delta > 0$ uma métrica de tolerância, com probabilidade $\geq 1 - \delta$,

$$\boldsymbol{E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4m_H(2N)}{\delta}}}$$

Resultado mais importante da teoria da aprendizagem!

Tamanho mínimo da amostra

Exemplo: Suponha que temos um modelo de aprendizagem com $d_{vc} = 3$ e desejamos um erro de generalização de no máximo 0.1 com confiança de 90% ($\epsilon = 0.1$ e $\delta = 0.1$). Qual deve ser o tamanho mínimo da amostra?

Do limitante de generalização VC, temos que

$$\epsilon = \sqrt{\frac{8}{N} ln \frac{4m_H(2N)}{0.1}} \leq 0.1$$

$$\boldsymbol{m_H(2N) \leq (2N)^{d_{vc}} + 1}$$

Assim,

$$N \geq \frac{8}{0.1^2} \ln\left(\frac{4(2N)^3 + 4}{0.1}\right).$$

Tamanho mínimo da amostra

Continuação

Fazendo $N = 1000$ no lado direito da inequação, obtemos

$$N \geq \frac{8}{0.1^2} \ln\left(\frac{4(2\times 1000)^3 + 4}{0.1}\right) \cong 21193\,.$$

Ao atribuir 21193 a N e continuar com o processo iterativo, converge-se para $\boldsymbol{N \cong 30000}$.

Limitante teórico “frouxo”

- O limitante de generalização é uma estimativa **“frouxa”** para estimar E_{out} com base em E_{in}.

- Para **não depender** de uma ***dataset* específico**, o limitante apoia-se em $m_H(N)$, que é calculado a partir da amostra de N exemplos que permite o **maior número** de dicotomias possível.

Conjunto de teste e Hoeffding

- Quando se usa um conjunto de dados de teste, a partir de uma hipótese g determinada a priori, o **fator $\boldsymbol{M}$ pode ser eliminado**.

- Como resultado, são necessários menos exemplos no conjunto de teste para obter-se boas estimativas para $\boldsymbol{E_{test}}$.

$\boldsymbol{d_{vc}}$ de importantes modelos de aprendizagem

- **Perceptron**
  - $\boldsymbol{d_{vc} = d + 1}$, onde d é a quantidade de parâmetros (variáveis).

- **Regressão linear**
  - $\boldsymbol{d_{vc} \cong d}$

- **Redes Neurais**
  - $\boldsymbol{d_{vc} \cong |W|}$, onde W é o conjunto de todos os pesos da rede.

Regras de ouro

- Dada a **dimensão VC** de um modelo de aprendizagem de máquina qualquer, para garantir a generalização é necessário que:

$$\boxed{\boldsymbol{N \geq 10d_{vc}}}$$

- Para regressão linear múltipla,

$$\boxed{\boldsymbol{N \geq 50 + 8d} \text{ ou } \boldsymbol{N \geq 104 + d}}$$

Referências bibliográficas

- Abu-Moustafa, Y.S.; Magdon-Ismail, M.; Lin, H-S. “*Learning from data*”. AMLBook, 2012.

- Faceli, K.; Lorena, A.C.; Gama, J.; Carvalho, A.C.P.L.F. “*Inteligência Artificial Uma Abordagem de Aprendizado de Máquina*”. LTC, 2011.

- Notas de aula do prof. Abu-Moustafa.