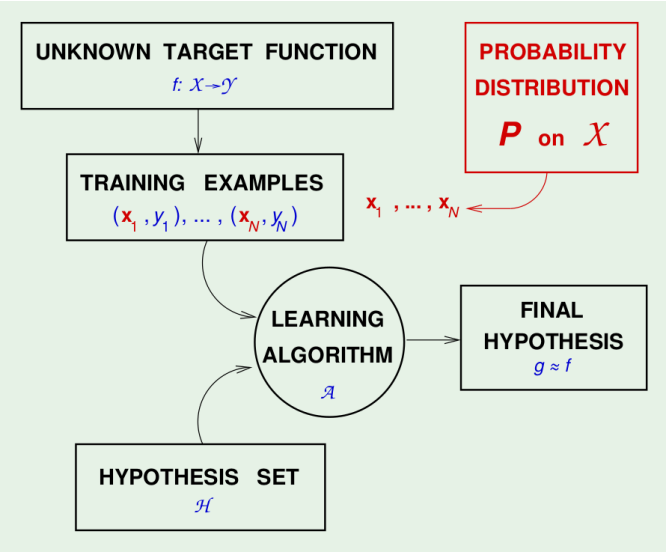
# Lista 2 - Introdução ao Aprendizado de Máquina (MAC5832)

Gustavo Quintero NUSP: 11350395 gdavid@ime.usp.br

16 de maio de 2022

1. Comente sobre o diagrama abaixo. O que o diagrama como um todo ilustra e o que cada componente representa?




**Solução**

O diagrama como um todo ilustra o problema de aprendizado, o qual começa com um conjunto de dados gerados independentemente por uma certa distribuição de probabilidade e termina com uma função hipótese $g$ que aproxima a função objetivo $f$.

O primeiro componente do diagrama é a função objetivo desconhecida $f : \mathcal{X} \to \mathcal{Y}$ a ser aproximada, onde $\mathcal{X}$ é o espaço de entrada e $\mathcal{Y}$ é o espaço de saída.

O segundo componente *Training Examples* é um conjunto de dados de entrada e saída com os quais o treinamento será realizado no processo de aprendizagem, os quais são escolhidos independentemente de acordo com uma creta distribuição de probabilidade $P$.

O componente *Learning Algorithm* consiste de um algoritmo de aprendizado que usa o conjunto de dados de treinamento para escolher uma certa formula $g : \mathcal{X} \to \mathcal{Y}$ que aproxima a função objetivo $f$.

O componente *Hypothesis Set* como o próprio nome indica, consiste em um conjunto de fórmulas candidatas para aproximar a função objetivo $f$.

Finalmente, o ultimo componente consiste do resultado de todo o processo de aprendizado.

2. O que é $E_{\text{in}}$ e $E_{\text{out}}$?

**Solução**

O valor $E_{\text{in}}$ é definido como a taxa de erro na amostra e o valor $E_{\text{out}}$ é definido como a taxa de erro fora da amostra, em função de uma certa hipótese $h$.

3. Quando consideramos a formulação teórica de aprendizado de máquina, uma das possibilidades é investigar o valor $|E_{\text{in}} - E_{\text{out}}|$. O que esse valor expressa e por que nos interessa investigar ele?

**Solução**

Pelas definições de $E_{\text{in}}$ e $E_{\text{out}}$, $|E_{\text{in}} - E_{\text{out}}|$ expressa a diferença entre a taxa de erro dentro e fora da amostra. É interessante investigar este valor pois se a probabilidade de $|E_{\text{in}} - E_{\text{out}}| > \epsilon$ estiver limitada por um certo *bound* pequeno, então poderemos ter uma noção do valor de $E_{\text{out}}$ uma vez que sabemos calcular o valor $E_{\text{in}}$.

4. Porque apenas garantir $|E_{\text{in}} - E_{\text{out}}| < \epsilon$ pode não ser suficiente?

**Solução**

Não é suficiente porque esta diferencia pode ser arbitrariamente pequena mas isto não que dizer que os valores $E_{\text{in}}$ e $E_{\text{out}}$ sejam pequenos.

5. A desigualdade de Hoeffding, no contexto de aprendizado de máquina, com respeito a uma certa hipótese h, é dada por:

$$P\Big(|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\Big) \leq 2e^{-2\epsilon^2 N}.$$

Explique o significado dessa desigualdade.

**Solução**

Esta desigualdade fornece um *bound* ($N, \epsilon$-dependente) que limita probabilidade da variação entre $E_{\text{in}}$ e $E_{\text{out}}$. É interessante notar que este limite não depende de $E_{\text{out}}$, e é valida para todo $N \in \mathbb{N}$ e todo $\epsilon > 0$. Observe que se definimos uma certa tolerância pequena $\epsilon$, então a desigualdade nos diz que neste caso precisaremos um valor de $N$ suficientemente grande, o qual faz muito sentido, pois se desejamos mais precisão no nosso aprendizagem precisamos também fornecer uma grande quantidade de dados.

6. A desigualdade de Hoeffding, no contexto de aprendizado de máquina, quando selecionamos uma hipótese de um espaço com $M$ hipóteses é dada por:

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

Comente sobre a diferença entre essa desigualdade e a do item anterior.

**Solução**

Como no caso anterior, esta desigualdade no da um *bound* para todo $\epsilon > 0$, todo $N \in \mathbb{N}$ e para todas as $M$ hipóteses candidatas. Porém, para qualquer hipótese em particular que consideremos, o novo *bound* é pior do que o anterior, pois quanto maior seja $M$ pior é o limitante da desigualdade.

7. O *bound* $2Me^{-2\epsilon^2 N}$ no item anterior foi obtido aplicando-se o *union-bound*. O que é *union-bound*?

**Solução**

O *union bound* é uma propriedade que satisfaz qualquer probabilidade (desde que toda probabilidade é uma *medida*) chamada de *subaditividade*, a qual nos diz que a probabilidade da união de uma quantidade finita de eventos (não necessariamente disjuntos) é menor ou igual à soma das probabilidades de cada um dos eventos. Formalmente: Dados $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_M$ eventos arbitrários, então

$$\mathbb{P}[\mathcal{B}_1 \textbf{ or } \mathcal{B}_2 \textbf{ or } \cdots \textbf{ or } \mathcal{B}_M] = \mathbb{P}\left(\bigcup_{i=1}^{M} \mathcal{B}_i\right) \leq \sum_{i=1}^{M} \mathbb{P}[\mathcal{B}_i].$$

8. O que são dicotomias ? O que é *growth-function*? O que é *break point*? Qual a relação entre eles?

**Solução**

Dados $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}$ uma dicotomia consiste de uma $N$-tupla $h(\mathbf{x}_1), \ldots, h(\mathbf{x}_N)$ de classes (por exemplo $\{-1, 1\}$) que divide as observações em dois grupos disjuntos (por exemplo os grupos tais que $h(\mathbf{x}_i) = -1$ e $h(\mathbf{x}_j) = +1$, com $i \neq j$), para uma certa hipótese $h \in \mathcal{H}$.

Dado um conjunto de hipóteses $\mathcal{H}$, a *growth-function* é definida como o máximo numero de dicotomias que podem ser geradas por $\mathcal{H}$ em quaisquer $N$ pontos. Formalmente: Dados $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}$ e $\mathcal{H}$ um conjunto de hipóteses, a *growth-function* é

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

Sabemos que o numero máximo de $N$-tuplas de duas classes (+1's e -1's) que podem ser produzidas por $N$ pontos é $2^N$. Logo, dizemos que $k$ é um *break point* se $m_{\mathcal{H}}(k) < 2^k$, isto é, $k$ é um *break point* se nenhum conjunto de hipóteses $\mathcal{H}$ consegue gerar todas as dicotomias possíveis para uma amostra de $k$ pontos $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathcal{X}$.

9. Escreva o que você entendeu sobre o processo envolvido na troca do $M$ em $2Me^{-2\epsilon^2 N}$ pelo *growth-function* $m_{\mathcal{H}}(N)$. Qual é o novo *bound* na troca?

**Solução**

Se interpretamos as dicotomias como uma função que associa a cada elemento de $h \in \mathcal{H}$ uma $N$-tupla $h(\mathbf{x}_1), \ldots, h(\mathbf{x}_N)$, então esta função é não injetora, isto quer dizer que varias hipóteses podem gerar a mesma dicotomia. Assim, podemos obter um *bound* menos pessimista desde que as probabilidades no *union bound* se sobrepõem dando origem ao fenômeno conhecido como *overlap*. Com isso em mente, o novo *bound* é dado por

$$4\, m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\epsilon^2 N},$$

o qual é um limitante mais optimista desde que não contém o fator $M$ e neste caso a *growth function* $m_{\mathcal{H}}(2N)$ é limitada por um polinômio quando possui um *break point*.

10. O que é VC dimension? Como chegamos ao VC generalization bound (Teorema 2.5 do livro) a partir da desigualdade de Hoeffding?

**Solução**

A dimensão VC de um conjunto de hipóteses $\mathcal{H}$ é o maior valor de $N$ tal que $m_{\mathcal{H}}(N) = 2^N$. Isto quer dizer que se $d_{VC}$ é a dimensão VC de um certo conjunto de hipóteses, então $k = d_{VC} + 1$ é um *break point*. Logo pelo Teorema 2.4 do livro temos que

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i} = \sum_{i=0}^{d_{VC}} \binom{N}{i}.$$

Por outro lado, se $\delta = 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$, então

$$\begin{aligned} \epsilon^{\frac{1}{8}\epsilon^2 N} &= \frac{4m_{\mathcal{H}}(2N)}{\delta} \\ \Rightarrow \epsilon^2 &= \frac{8}{N} \ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right) \\ \Rightarrow \epsilon &= \sqrt{\frac{8}{N} \ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right)}, \end{aligned}$$

em que a solução negativa é descartada desde que $\epsilon > 0$. Note que

$$\mathbb{P}[|E_{\text{out}} - E_{\text{in}}| > \epsilon] = 1 - \mathbb{P}[|E_{\text{out}} - E_{\text{in}}| \leq \epsilon].$$

Portanto, $\mathbb{P}[|E_{\text{out}} - E_{\text{in}}| > \epsilon] \leq \delta$ é equivalente a

$$1 - \mathbb{P}[|E_{\text{out}} - E_{\text{in}}| \leq \epsilon] \leq \delta,$$

ou seja, $\mathbb{P}[|E_{\text{out}} - E_{\text{in}}|] \geq 1 - \delta$, e é claro que $|E_{\text{out}} - E_{\text{in}}| \leq \epsilon$ implica

$$E_{\text{out}} - E_{\text{in}} \leq \sqrt{\frac{8}{N} \ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right)}.$$

Consequentemente, para qualquer tolerância $\delta > 0$ obtemos a equação (2.12) do Teorema 2.5 do livro com probabilidade maior ou igual a $1 - \delta$.

11. A equação (2.13) do livro-texto é a seguinte

$$N \geq \frac{8}{\epsilon^2} \ln\left(\frac{4((2N)^{d_{VC}} + 1)}{\delta}\right).$$

Comente quais são as implicações práticas dela.

**Solução**

Se decidimos que a tolerância $\epsilon$ e o parâmetro $\delta$ (que determina com que frequência a tolerância $\epsilon$ é violada) sejam valores muito pequenos, a desigualdade nos diz que precisamos um valor $N$ suficientemente grande para obter a precisão que desejamos.

12. Quais são as similaridades e diferenças entre o *VC analysis* e o *Bias-variance analysis*?

**Solução**

A similaridade que existe é que tanto o *VC analysis* quanto o *Bias-variance analysis* decompõem o valor $E_{\text{out}}$ como a soma de duas quantidades. Quanto às diferencias com o *VC analysis* obtemos um *bound* para $E_{\text{out}}$ da forma

$$E_{\text{out}} \leq E_{\text{in}} + \Omega,$$

enquanto com o *Bias-variance analysis* obtemos

$$E_{\text{out}} = \text{bias} + \text{var}$$

uma identidade que depende de uma expressão dos valores esperados na hipótese escolhida no processo de aprendizado. No caso do *VC analysis* quando maior seja $N$ melhor será o *bound*, mas com o *Bias-variance analysis* devemos tomar cuidado pois valores muito altos para $N$ podem desequilibrar o segundo membro da identidade anterior obtendo valores muito discrepantes de bias e var.