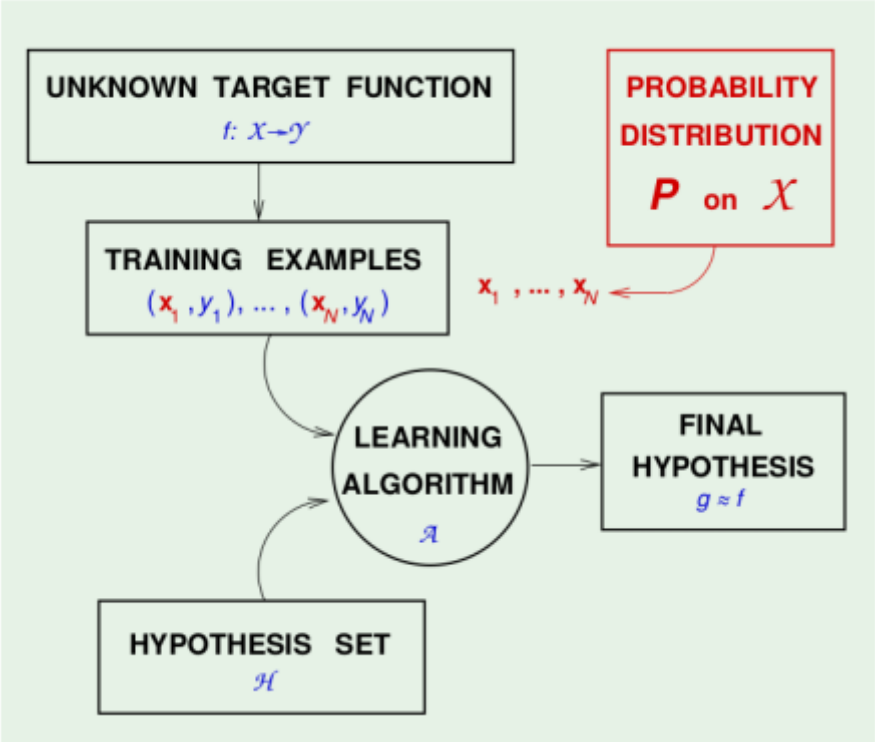
MAC0460

Lista de exercícios 2

Caio Túlio de Deus Andrade - 9797232

---

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



Resposta: O diagrama como um todo ilustra o **learning problem** (ou problema do aprendizado). Cada componente é um componente do aprendizado: Dado um conjunto de treino (**training examples**, existe uma função alvo desconhecida f que mapeia perfeitamente x_i com y_i. Cada ponto x_i é amostrado de um conjunto universo seguindo uma distribuição de probabilidade P. O algoritmo de aprendizado 'A' seleciona funções possíveis dentro do conjunto de hipótese e produz funções de hipóteses. Esse processo é realizado até que o algoritmo de aprendizado encontra uma função hipótese que aproxime apropriadamente a função desconhecida f.

2. O que é o E_{in} e o E_{out}?

   Resposta: E_{in} é o *In sample error*, ou seja, é a taxa de erro de classificação sobre os pontos amostrados. E_{out} é o *Out of sample error*, em outras palavras, é a probabilidade de um ponto qualquer não ser corretamente classificado.

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

Resposta: O valor expressa a diferença entre o erro de predição de uma função hipótese (h, por exemplo) com dados de treino e o erro de predição com dados quaisqueres, dos quais os dados de treino são amostrados. É interessante analisar este valor porque, se queremos que nosso algoritmo **aprenda** a prever valores a partir dos dados, ao invés de **decorar** respostas, é interessante que o valor em questão seja pequeno.

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

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

Explique o significado dessa desigualdade.

Resposta: Fixando uma função hipótese h e uma precisão ϵ , a desigualdade de Hoefflding afirma que conforme o tamanho da amostra N cresce, é exponencialmente menos provável que o erro da predição de h na amostra se distancie do erro de predição de h out of sample por mais do que um fator de tolerância ϵ. Em outras palavras, esperamos ,que conforme aumentamos o tamanho da amostra, o erro de predição da função h seja menos influenciado pelo processo de amostragem e se aproxime da probabilidade real do erro de predição da função h;

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

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

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

Resposta: No item anterior, a função hipótese é fixada. Neste, a função em questão (g), é a função **final** escolhida pelo algoritmo de aprendizado, dentre as M possíveis funções em H. Dessa forma, a Desigualdade de Hoefflding se torna uma maneira de caracterizar o erro de generalização do algoritmo de aprendizado.

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

Resposta: Union bound é uma inequalidade em teoria de probabilidades que afirma que a Probabilidade da união de eventos é menor ou igual à soma das probabilidades de cada evento individual.

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

Resposta: Dicotomias são n-tuplas, onde cada tupla assume valores -1 ou 1. Growth-function é o número máximo de dicotomias que pode ser gerado em um conjunto de hipótese H tomando N pontos. Break point é um limite superior para a growth-function: um k é um break point para o conjunto de hipóteses H se não é possivel cobrir (shatter) um conjunto de pontos de tamanho k. Dizemos então que se k é um break point, então $m_h(N) < 2^k$.

A relação entre growth-function e dicotomias já foi explicada. Resta explicar a relação entre growth-function e break point: estudamos a growth function para limitar, de maneira mais restrita, o erro de generalização. Como é muito difícil calcular exatamente todas dicotomias (e, portanto, encontrar $m_H(n)$), faz mais sentido encontrar um limite superior: se encontrarmos o break point, conseguimos limitar superiormente a growth function pois $m_h(N) < 2^k$. Assim, os três conceitos estão relacionados com o objetivo de limitar de maneira mais rigorosa o valor do erro de generalização, já que o valor M (número de funções hipótese) é geralmente algo que tende ao infinito.

8. O que você entendeu sobre o processo envolvido na troca do M em $2Me^{2\epsilon^2 N}$ pelo *growth-function* $m_h(N)$? Qual o interesse em se fazer essa troca? Qual é o novo *bound* obtido após a troca?

Resposta: Entendi que é um processo que traz um limite mais rigoroso: como o número de funções hipóteses (M) tende ao infinito, não faz sentido usar este número como um limitante, ainda mais porque a diferença entre duas funções hipóteses pode ser mínimo (um exemplo é um ajuste mínimo nos pesos do Perceptron). O interesse então é ter um limite mais rígido considerando as possíveis "intersecções" entre as funções de hipótese. O novo bound obtido é:

$$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2m_h(n)e^{-2\epsilon^2 N} \leq 2(\sum_{i=0}^{k-1} \binom{N}{i})e^{-2\epsilon^2 N}$$

Onde k é o break-point para o espaço de hipóteses H. Essa troca ainda não caracteriza a hipótese de que um aumento em N implica em uma maior genrealização do modelo, isso é obtido na formulação escrita na questão 10.

9. Dissemos que a VC *dimension* relaciona-se com a expressividade do espaço de hipóteses. Comente sobre isso.

   Resposta: O VC dimension está diretamente relacionado com o número de parâmetros que um modelo tem. Quanto mais parâmetros, maior a cardinalidade do espaço de hipótese. Sendo assim, podemos dizer que um modelo com *VC dimension* alto apresenta alta expressividade devido ao número de funções hipótese que ele pode expressar. Alternativamente, um modelo com VC dimension baixo apresenta baixa expressividade por tratar de poucas funções de hipótese comparativamente.

10. Como o VC *bound* é expresso em termos da VC *dimension*?
    Resposta:

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} ln(\frac{4((2N)^{d_{vc}}+1)}{\delta})}$$

11. Baseado no VC bound, explique como podemos calcular o número de amostras necessárias para se garantir uma certa precisão ϵ, com probabilidade $1 - \delta$, supondo que o espaço de
    hipóteses considerado tem dimensão VC igual a d_{VC} ?

    Resposta: Pela equação dada no exercício 10, para garantir uma precisão ϵ basta fazer:

$$\sqrt{\frac{8}{N} ln(\frac{4((2N)^{d_{vc}}+1)}{\delta})} \leq \epsilon$$

    Elevando ambos lados ao quadrado, segue que:

$$N \geq \frac{8}{\epsilon^2} ln(\frac{4((2N)^{d_{vc}}+1)}{\delta})$$

    Então basta escolher um N que respeite a inequação acima para garantir precisão ϵ com probabilidade $1 - \delta$ num espaço de dimensão d_{vc}.

12. Por que apenas garantir $|E_{in}(h) - E_{out}(h)| < \epsilon$ não é o suficiente?

    Resposta: porque essa é uma análise fixa pra uma hipótese h. Queremos que nosso algoritmo de aprendizado tenha uma hipótese final g que generalize bem o seu aprendizado. Garantir a condição acima para uma das funções possíveis a serem escolhidas no espaço de hipótese não é o bastante.

13. Quais as similaridades e diferenças entre o VC analysis e o *Bias-Variance analysis*?

    Resposta:

    - A VC analysis é baseada no conjunto de hipóteses h. Já a *Bias-variance* analysis, se baseia no conjunto de hipóteses H e no algoritmo de aprendizado A
    - A analise Bias-Variance é mais conceitual: como ela exige conhecimento da função alvo f e da distribuição de probabilidade de amostragem de X, ambos desconhecidos, não é possível calcula-la. Já a VC Analysis é útil para encontrar um tamanho de amostra que nos garante uma certe precisão com uma dada probabilidade num espaço de hipoteses com dimensão VC.
    - Ambas análises são medidas que analisam se o modelo está generalizando bem seu conhecimento ou não.

14. Escreva a sua opinião sobre quão úteis são os conteúdos cobertos nas *lectures* mencionadas para o entendimento sobre *Machine Learning*.

Resposta: Achei bem útil. Já tinha uma familiriadade com Machine Learning na prática e sabia implementar alguns modelos e entendia alguns conceitos, mas o curso está descendo em discussões bem *Meta* sobre machine Learning. O capítulo 2 inteiro é dedicado a como podemos garantir que modelos de ML de fato aprendem, isso está ajudando muito a enriquecer meu conhecimento.