MAC0460 - Introdução ao aprendizado de máquina
Lista 1

Felipe Castro de Noronha
NUSP: 10737032

Questão 1

O diagrama abaixo mostra os componentes que participam no processo de aprendizado, em termos de aprendizado de maquina. Onde cada um contribui para gerar uma função que aproxima uma *função alvo*. Podemos descrever cada bloco da seguinte maneira:

- Uma *função alvo desconhecida* $f : X \to Y$ que *gera* todos os exemplos de treinamento. É a função que queremos aproximar.
- Um conjunto de *exemplos de treinamento*. São as observações que temos do mundo para podermos treinar nossa hipotese. Essas observações são da forma $(x_i, y_i = f(x_i))$ onde $x_i \in \mathbb{R}^d$.
- Um *espaço de hipoteses* $\mathcal{H}$ que representa uma abstração de todas as possiveis funções que nosso algoritmo pode gerar. Chamamos de hipotese um elemento $h \in \mathcal{H}$.
- Um *algoritmo de aprendizado* $\mathcal{A}$ que, a partir dos exemplos de treinamento, gera uma função que aproxima f. Diversos algoritmos podem ser usados aqui, como o *perceptron*, *regressão linear*, *regressão logistica* entre outros. A escolha do algoritmo a ser utilizado se da de acordo com a necessidade.
- Uma *hipotese final* g que aproxima f. Essa hipotese é o resultado de nosso algoritmo de aprendizado, a partir dela, podemos obter novos resultados de acordo com novas observações do mundo, e esses resultados se aproximam ao esperado caso fossem calculados com a função f.

Questão 2

O E_{in}, *in sample error*, é o erro obtido a partir dos exemplos de treinamento, isto é, é uma medida para dar a ideia entre a diferença dos resultados calculados por f e g. Já E_{out}, *out of sample error*, é o erro calculado a partir de um conjunto de dados que não estava nos exemplos de treinamento, é uma métrica de quão bem nosso treinamento generalizou para um conjunto de dados não visto até então.

Uma situação para explicarmos um exemplo é a seguintes: imagine que somos um banco que oferecemos empréstimos á nossos clientes. Para isso, temos varias informações

sobre o mesmo, como: idade, sexo, movimentação mensal na conta, valor em bens, etc. Todas essas informações formam um x_i. Podemos colocar todas esses x_i e os respectivos valores liberados para empréstimo por gerentes humanos (y_i) nos anos anteriores no nosso conjunto de exemplos de treinamento.

No cenário acima, a diferença entre os valores liberados por gerentes humanos e nossa hipotese g, para os dados de anos anteriores, seria o E_{in}, essa diferença é normalmente calculada pelo *erro quadrático médio*. E o E_{out} poderia ser obtido de maneira semelhante, porém, agora, com dados de novos pedidos de empréstimos, que não foram usados para treinamento.

Questão 3

Pois minimizar o E_{in} não é suficiente para dizermos que nosso algoritmo é bom, em particular, podemos gerar uma hipotese que não é uma boa generalização do mundo, isto é, ela se mostra muito precisa com os dados do conjunto de treinamento, porém, não tem uma performance tão boa com dados do mundo real.

Inclusive, podemos ter o problema do *over fitting*, onde nossa hipotese se ajusta tanto ao conjunto de treinamento que a avaliação de E_{out} se torna ainda pior, pois essa hipotese se tornou ainda menos geral.

Minimizar o E_{in} pode ser suficiente quando a função que queremos explorar é muito simples.

Questão 4

Esse valor nos mostra a diferença de performance da nossa hipotese entre os dados de treinamento e os dados de mundo real. Podemos dizer que, quanto maior este valor, menos precisa/previsivel é nossa hipotese. Alguns chamam esse valor de *erro de generalização*.

Questão 5

Uma hipotese é uma função g que aproxima f (função alvo). A hipotese é gerada por um algoritmo de aprendizado.

Questão 6

Essa desigualdade pode ser lida como a *probabilidade do erro calculado não ser uma boa aproximação para a qualidade da hipotese, dada uma certa tolerância, uma hipotese e o numero de observações no conjunto de treinamento*. Ela nos dá uma ideia da qualidade de generalização da hipotese.

Questão 7

A diferença em que essa desigualdade tem outro fator contribuinte, o M, isto é, o numero de hipoteses de $\mathcal{H}$. Ela nos dá um pouco mais de generalidade na analise do erro, fazendo com que nossa probabilidade diga a respeito de todo o espaço de hipoteses.

Questão 8

Podemos entender *union bound* como um aproximação da probabilidade da disjunção de vários eventos. Com isso, podemos obter que $\mathbb{P}[e_1 \vee e_2 \vee \cdots \vee e_n] \leq \sum_{i=1}^{n} \mathbb{P}[e_i]$.

Questão 9

Para definir uma dicotomia precisamos de um espaço de hipoteses e um conjunto de observações. Assim, dado $h \in \mathcal{H}$, temos que o conjunto $\{h(x_1), h(x_2), ..., h(x_n)\}$ é uma dicotomia, ou seja, uma dicotomia é o conjunto de valores gerados por uma hipotese.

Uma observação importante é que uma hipotese só gera uma dicotomia, mas uma dicotomia pode ser gerada por varias hipoteses. Logo, as dicotomias podem nos dar uma noção do tamanho pratico do espaço de hipoteses.

Questão 10

A *função de crescimento* $m_{\mathcal{H}}(N)$ nos diz o numero máximo de dicotomias que podem ser geradas a partir de um espaço de hipoteses $\mathcal{H}$ usando N pontos/observações. Segue que $m_{\mathcal{H}}(N) \leq 2^N$, porém, esse limite pode ser ainda menor quando usamos a ideia de *break points*.

Questão 11

As dicotomias, e consequentemente, as funções de crescimento, nos dão uma ideia do tamanho do espaço de hipoteses, levando o numero de pontos do meu conjunto de treinamento. Isso é útil para fazermos uma melhor estimativa do erro.

Questão 12

Provando que a função de crescimento é polinomial nos permite fazer a substituição de M por $m_{\mathcal{H}}(N)$ na formula $\sqrt{\frac{1}{2N} \ln(\frac{2m_{\mathcal{H}}(N)}{\delta})}$ que constitui uma avaliação útil do *generalization bound*.

Questão 14

Primeiramente, sabemos que se $m_{\mathcal{H}}(N) < 2^k$ então $m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$. Um de k em que isso acontece é chamado de *break point*.

Além disso, temos que a *VC dimension* de um espaço de hipoteses $\mathcal{H}$ é o valor $d_{vc}(\mathcal{H})$, que representa o maior N tal que $m_{\mathcal{H}}(N) = 2^N$. Se $\mathcal{H}$ não possui um break point então $d_{vc}(\mathcal{H}) = \infty$.

Com isso temos que a *VC dimension* nos permite saber até onde $m_{\mathcal{H}}(N)$ tem crescimento exponencial, e, com isso, conseguimos ter uma outra aproximação, em termos de crescimento, do nosso espaço de hipoteses.

Questão 15

O valor d_{vc} de um perceptron d-dimensional é $d + 1$.

A primeira coisa necessaria para a demonstração é mostrar que um perceptron d-dimensional pode dividir $d + 1$ pontos de todas as maneiras possiveis, ou seja, de 2^{d+1} maneiras diferentes. Logo, d_{vc} do perceptron é pelo menos $d + 1$.

O segundo, e ultimo passo, é que o perceptron falha em gerar todas as possiveis dicotomias para $d+2$ pontos, e logo, aqui, estariamos estabelecendo um limitante superior para d_{vc}.

Questão 16

Pelo fato de que o *VC dimension* nos da informações sobre o *break point* da função de crescimento, podemos dizer que quanto menor o valor d_{vc} menor a ordem da nossa função de crescimento, com isso, nosso *generalization bound* fica menor, o que resulta em uma aproximação/hipotese de melhor qualidade.

Questão 17

Esse é o chamado *generalization bound* e nos define um limite para o valor que E_{out} pode assumir. Com isso, temos a seguinte expressão $E_{out} \leq E_{in} + \sqrt{\frac{8}{N} ln(\frac{4((2N)^{d_{vc}}+1)}{\delta})}$.

Questão 20

Pois podemos ter o cenário onde temos um alto valor de E_{in} e E_{out}, logo, podemos ter criado uma hipotese que não seja boa para o conjunto de treinamento e nem seja geral, porém, se se os erros observados fossem próximos, poderíamos ter um ϵ pequeno. Logo, essa desigualdade não seria muito suficiente para avaliarmos a qualidade de h.

Questão 21

Acredito que o *VC bound* já é um ótimo bound para nossos estudos.

Questão 22

Acredito que a principal ideia a ser tomada é que d_{vc} nos da uma ideia geral de até que ponto nosso espaço de hipoteses é, de fato, exponencial. Com isso, conseguimos ser mais precisos á respeito das estimativas de erros calculadas, além de que, agora, temos mais noção do que esses resultados significam.