Lista 1 - MAC0460

Nome: André Akira Hayashi NUSP: 9293011

1

O modelo representa o modelo do apredizado de máquina. Sendo que suas componentes são: A entrada $\mathbf{x}$, a função destino desconhecida $f : X \to Y$, onde X é o espaço de entrada (conjunto de todas as entradas possíveis $\mathbf{x}$) , e Y é o espaço de saída (conjunto de todas as saídas possíveis). Há um conjunto de dados D de exemplos de entrada e saída $(x_1, y_1), \cdots, (x_N, y_N)$, onde $y_n = f(x_n)$ para n = 1, ..., N. Finalmente, existe o algoritmo de aprendizado que usa o conjunto de dados D para escolher uma fórmula $g : X \to Y$ que se aproxima de f. O algoritmo escolhe g de um conjunto de fórmulas candidatas, que é chamado de conjunto de hipóteses H.

2

E_{in} é a taxa de erro que acontece na amostra (os pontos que não deveriam estar na amostra), o que corresponde a ν no modelo de compartimento (bin), será chamada de "*in-sample error*" (ou E_{in}). E_{out} é a taxa de erro que acontece fora da amostra (os pontos que não deveriam estar fora da amostra), o que corresponde a μ no modelo de compartimento (bin).

3

Existem casos em que o E_{in}, não precisará ser aproximadamente zero (ou atingir seu mínimo). Um exemplo disso ocorre em previsões financeiras em que a imprevisibilidade do mercado torna impossível de se obter uma previsão com erro próximo de zero. Tudo o que é esperado é uma previsão que acerte com uma maior frequência, se isso for possível, as apostas vencerão a longo prazo. Isso significa que uma hipótese que tem $E_{in}(g)$ um pouco abaixo de 0,5 funcionará, desde que $E_{out}(g)$ esteja próximo o suficiente de $E_{in}(g)$.

4

O valor $E_{in} - E_{out}$, representa uma generalização do erro, ou seja, é a medida do quão preciso um algoritmo é capaz de prever valores de resultados para dados que não foram vistos anteriormente. Esse erro pode ser minimizado, para

que seja evitando ajustes excessivos no algoritmo de aprendizado, pois com o seu excesso o modelo pode ser muito mais afetado pelas diversas aproximações feitas.

5

Existem duas hipóteses comentada pelo prof. Abu-Mostafa, o h e o H, sendo que elas representam:

- h (hipótese): uma única hipótese, ou seja é um modelo candidato que relaciona as entradas com as saídas e pode ser avaliado e usado para fazer previsões.
- H (conjunto de hipóteses): um conjunto de hipóteses possíveis para o relacionamento de entradas com as saídas que podem ser utilizadas.

6

A desigualdade de Hoeffding é uma probabilidade garantida de que $E_{in}(h)$ não se afasta muito de $E_{out}(h)$, sendo que ϵ é um valor pequeno que é usado para medir o desvio de $E_{in}(h)$ em relação à $E_{out}(h)$. Com isso pode-se concluir que a probabilidade de $E_{in}(h)$ estar mais próximo de ϵ do que $E_{out}(h)$ é menor ou igual do que algum limite que diminui de acordo com o ϵ e/ou com o aumento do tamanho da amostra. Ou seja, quanto maior o tamanho da amostra e maior a margem de erro, menor a probabilidade de você ultrapassar essa margem de erro.

7

Matematicamente, esta é uma versão "uniforme" da equação do exercício 6, onde aproxima-se simultaneamente todos os $E_{out}(h_m)'s$ pelos $E_{in}(h_m)'s$ correspondentes. Isso permite que o algoritmo de aprendizado escolha qualquer hipótese baseada em E_{in} e espere que o E_{out} correspondente siga uniformemente o exemplo, independentemente da hipótese escolhida. A desvantagem para estimativas uniformes é que a probabilidade vinculada $2Me^{-2\epsilon^2}$ é um fator de M mais flexível do que o limite de uma única hipótese e só será significativo se M for finito.

8

Seja os conjuntos B1, B2, ..., BM estão fortemente sobrepostos, o union bound fica fraco, já que as áreas de diferentes conjuntos correspondem às suas probabilidades. O union bound diz que a área total coberta por B1, B2, ..., BM é menor que a soma das áreas individuais, porém isso é uma uma superestimação bruta de quando as áreas se sobrepõem fortemente.

9

A definição de growth function é baseada no número de hipóteses diferentes que H pode implementar, mas apenas em uma amostra finita de pontos, e não em todo o espaço de entrada X. Se $h \in H$ for aplicado a uma amostra finita x_1, ... , $x_N \in X$, obtemos uma N-tupla $h(x_1), ..., h(x_N)$ de 1's. Essa N-tupla é chamada de dicotomia, pois divide $x_1, ..., x_N$ em dois grupos: aqueles pontos para os quais h é - 1 e para aqueles que o h é + 1. Essa tupla N é chamada dicotomia, pois divide x1, · ·, XN em dois grupos: aqueles pontos para os quais h é - 1 e aqueles para os quais h é + 1.

As dicotomias geradas por um espaço de hipóteses H(x_1, ..., x_N) são um conjunto de hipóteses exatamente como H, exceto que as hipóteses são vistas apenas pelos olhos de N pontos. Um H maior (x_1, ..., X_N) significa que H é mais "diversificado", gerando mais dicotomias em x_1, ..., X_N.

10

Uma growth-function é baseada no número de hipóteses diferentes que H (conjunto de hipóteses) pode implementar, mas apenas em uma amostra finita de pontos, e não em todo o espaço de entrada X.

11

12

Considerando que $m_H(N) = 2^N$ para todo N, se substituir M por ele na equação:

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

, o limite $\sqrt{\frac{1}{2N} ln \frac{2M}{\delta}}$ na generalização do erro ela não irá para zero, independentemente do número de exemplos treinos N que tivermos. Entretanto, se $m_H(N)$, pode ser limitado por um polinômio, a generalização do erro irá para zero, desde que $N \to \infty$, ou seja, a generalização será boa se tivermos um número suficiente de exemplos.

13

Se $d_{VC}(H) = \infty$, a substituição de M por $m_H(N)$ irá falhar, pois a growth function neste caso é exponencial em N. Para qualquer valor finito de d_{vc}, o erro irá convergir para zero a uma velocidade determinada por d_{vc}, pois ele é a ordem do polinômio, quanto menor for o d_{vc}, mais rápida será a convergência para zero.

14

Para qualquer H, desde que $H(x_1, ..., x_N) \subseteq \{-1, +1\}^N$ (o conjunto de todoas as possiveis dicotomias em qualquer N pontos), o valor de $m_H(N)$ é o máximo $|\{-1, +1\}^N|$, portanto $m_H(N) \leq 2^N$. Se H é capaz de gerar todas as possíveis dicotomias em $x_1, ..., x_N$, então $H(x_1, ..., x_N) = \{-1, +1\}^N$ e é possível dizer que H pode ser "quebrar" $x_1, ..., x_N$.

15

A VC dimension de um dimensão d Perceptron é $d+1$. Com isso temos também que, $d_{VC} \geq d+1$, portanto para provar a d_{VC}, basta mostrar que $d_{VC} \leq d+1$, ou que não é possível "quebrar" nenhum conjunto de $d+2$ pontos.

Para qualquer $d+2$ pontos, $x_1, ..., x_{d+2}$, e como temos mais pontos do que dimensões, temos que $x_j = \sum_{i \neq j} a_i x_i$, onde nem todos os a_i's são zeros. Considerando a dicotomia de que, x_i's com não zeros a_i, tem $y_i = \text{sign}(a_i)$ e x_j tem $y_j = -1$, e como nenhum Perceptron consegue implementar tal dicotomia, não existe $d_{VC} = d + 2$.

16

Quanto maior for a quantidade de parâmetros que um modelo possui, mais diverso será o seu conjunto de hipóteses, o que se reflete em um valor maior da growth function $m_H(N)$. A VC dimension mede a efetividade desses parâmetros que permite o modelo expressar um conjunto diversificado de hipóteses, essa diversidade não é necessariamente uma coisa boa no contexto da generalização.

17

O bound inicial $2Me^{-2\epsilon^2 N}$, foi substituído pelo nível de tolerância δ. Para qualquer talerância fixa ϵ, o bound em E_{out} estará arbitrariamente perto de E_{in} para um N suficientemente grande.

18

19

Como em "bons modelos", possuem um d_{VC} finito e um N suficientemente grande, E_{in} será próximo de E_{out}, para "bons modelos", a performace do in-sample generaliza para o out-sample, portanto $|E_{in} - E_{out}|$ será próximo de zero neste caso. Para "maus modelos", o d_{VC} é infinito, não importando o tamanho do N, não é possível fazer conclusões sobre a generalização de E_{in} para E_{out}, portanto $|E_{in} - E_{out}|$ tenderá ao infinito.

20

Provar que $|E_{in} - E_{out}| < \epsilon$ não é suficiente, pois a razão de que $m_H(2N)$ aparece no VC Bound ao invés de $m_H(N)$, decorre do evento $|E_{in} - E_{out}| > \epsilon$, que não depende apenas de D, mas também de todo o X, já que $E_{out}(h)$ é baseado em X. Isso quebra a premissa principal de agrupar $h's$ com base em seu comportamento em D, já que os aspectos de cada h fora de D afetam $|E_{in} - E_{out}| > \epsilon$. Para solucionar isso, considera-se o evento artificial $|E_{in} - E_{out}| > \epsilon$, em que E_{in} e E'_{in} são baseadas em duas amostras D e D' cada uma de tamanho N, esse é o motivo do $2N$.

21

Não.

22

A VC theory, tenta explicar o processo de aprendizagem utilizando estratégias estatísticas. Uma de suas principais aplicações na aprendizagem de máquina é fornecer generalizações para os algoritmos de aprendizagem. Desse ponto de vista, a VC theory está relacionada à estabilidade do algoritmo.