Aprendizado e Bias Indutivo

Prof. Ricardo J. G. B. Campello – ICMC/USP

(Seguindo de perto o livro *Machine Learning* de Tom Mitchell)

Em **aprendizado indutivo** (supervisionado ou não) tem-se um conjunto $\mathbf{D} \subset \mathbf{X}$ de exemplos (padrões) de treinamento e um espaço de hipóteses (modelos) H a partir do qual deve-se selecionar uma hipótese (modelo) h particular que possa associar qualquer instância $x_i \in \mathbf{X}$ a um determinado valor de saída $f(x_i)$ da função alvo do aprendizado (classe de x_i, *cluster* de x_i, ação de controle a partir de x_i, etc). A ampla maioria dos problemas de Aprendizado de Máquina (AM) é desta natureza[1].

Pode-se descrever um passo de **inferência indutiva** como $(\mathbf{D} \wedge x_i) \succ f(x_i)$, onde "$\wedge$" é uma conjunção e $a \succ b$ significa que b é inferido a partir de a. Ou seja, a partir do conjunto de exemplos de treinamento **D** e de um novo exemplo x_i é possível obter uma saída $f(x_i)$ utilizando um modelo h aprendido indutivamente a partir de **D**. O conjunto das considerações e restrições necessárias para induzir h é denominado de "bias indutivo".

Formalmente, **bias indutivo** é o conjunto de todas as considerações e restrições adicionais a **D** e x_i necessário para que se possa inferir $f(x_i)$ **dedutivamente** a partir de x_i. Em outras palavras, o bias indutivo é um conjunto **C** de considerações e restrições tal que $(\mathbf{C} \wedge \mathbf{D} \wedge x_i) \vdash f(x_i)$, onde $a \vdash b$ significa que b segue dedutivamente a partir de a.

Pelas próprias definições acima, fica evidente que o bias indutivo é responsável pela **capacidade de generalização** de um método de AM, ou seja, a capacidade de inferir uma saída $f(x_i)$ para um exemplo x_i novo ($x_i \notin \mathbf{D}$). De fato, um algoritmo de aprendizado sem bias indutivo não é capaz de aprender sobre padrões não vistos durante o treinamento. Uma discussão e exemplo interessantes são apresentados em [1] (Seção 2.7.2).

O bias indutivo pode ser de dois tipos fundamentais:

---

[1] Uma contraposição é o **Aprendizado Dedutivo**, e.g., de proposições lógicas a partir de outras proposições. Aprendizados dedutivo e indutivo podem ser combinados. Por exemplo, em **Aprendizado Baseado em Explicações** (*Explanation-Based Learning*), discutido no Capítulo 11 (*Analytical Learning*) do livro de Tom Mitchell [1], tem-se, além de **D** e H, uma "teoria de domínio" B (e.g., regras lógicas fornecidas por um especialista) que consiste essencialmente de conhecimento de domínio prévio introduzido no sistema de aprendizado e que: (i) deve ser respeitado ao se escolher uma hipótese h de H com base em **D**; e/ou (ii) pode ser utilizado para **deduzir** propriedades e informações do problema, não explícitas nos exemplos em **D**, que permitam levar a uma escolha mais apropriada para a hipótese h.

(a) **Bias de Restrição** (ou de **linguagem**): Ocorre na escolha do espaço de hipóteses *H*. Se *H* não contém modelos candidatos capazes de representar qualquer função alvo para o problema de AM em questão, então se diz que *H* não é um espaço completo e, portanto, restringe a representabilidade da solução. Em regressão ou classificação, por exemplo, se escolhemos um modelo linear para um problema não linear, certamente *H* (espaço de todos os modelos lineares) não conterá uma hipótese consistente com os dados do problema em questão. Por esta razão, em geral procura-se evitar este tipo de bias. Algoritmos de classificação baseados em árvores de decisão, como ID3 e C4.5 por exemplo, não possuem esse tipo de bias. Logo, diz-se que o espaço de hipóteses *H* desses algoritmos é um **espaço completo**. De fato, esse espaço é completo porque as árvores de decisão ou as regras correspondentes, dadas por disjunções de conjunções, podem representar perfeitamente qualquer classificação de um conjunto de instâncias que não possua inconsistências[2].

(b) **Bias de Busca** (ou de **preferência**): Ocorre na escolha da estratégia de seleção de uma hipótese *h* específica dentre todas as candidatas em *H*. Em outras palavras, esse tipo de bias é decorrência da estratégia de busca utilizada para encontrar um modelo apropriado aos dados em questão dentro do espaço *H* de possíveis modelos. Diferentes estratégias de busca representam diferentes metodologias de preferir certas hipóteses frente a outras hipóteses. Esse é o tipo de bias presente em algoritmos de classificação como o ID3 e C4.5 por exemplo. Esses algoritmos preferem árvores cujos atributos que possuem maior ganho de informação estão acima na árvore, sem retroagir depois nesta decisão (ou seja, sem *backtracking*, ficando, portanto, sujeitos a encontrar soluções sub-ótimas, apesar do espaço completo dentro do qual realizam a busca). Conforme discutido em [1] (Seção 3.6), esse bias também implica uma tendência para árvores mais curtas / simples do que longas.

Referências

[1] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[2] Instâncias inconsistentes são aquelas com valores diferentes do atributo meta (i.e. de classes diferentes), porém com valores idênticos dos demais atributos.