1.31 Lecture 3

Knowledge: Testable Predictions 可证伪性

Confirmation Bias (unwillingness to try negative example)

- 所得data可能biased
- 可得example满足很多rule

hypothesis 可能包含于 rule

基于个人的hypothesis，尝试与hypo相违背的example，才有可能找到rule.

Types of Data.

- Records
  tuples/vector e.g. (name, age) (Bob, 4)
- Graphs
  nodes, edges (directed/undirected)
  Adjacency list/matrix
- Image
  由pixels组成 (red, green, blue)
- Text
- Strings

- Unsupervised Learning
  将data分类 (clustering)
- Supervised Learning

Data Processing

↓ · can / can't be used? missing? inconsistent?

Explore Data → Extract Features → Create Model

Distance

compare data points的方法

dissimilarity function

$d(i,j) = \begin{cases} \text{值大，不相似} \\ \text{小，相似} \end{cases}$

Minkowski Distance

$x, y \in \mathbb{R}^d$, $p \geq 1$

$L_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$

当 $p=2$ 时为 Euclidean distance.

· Cosine Similarity

算2个 vectors 间 cos(θ)

因 dissimilarity function 大时不相似，function 小为 $k - sl(x,y)$