机器学习
Machine Learning

（二）概念学习

何劲松
中国科学技术大学

Machine Learning

10-701/15-781, Spring 2011

Carnegie Mellon University

Tom Mitchell

MACHINE LEARNING DEPARTMENT

Home | People | Lectures | Recitations | Homeworks | Project | Previous material

Date	Lecture	Topics	Readings and useful links	Handouts
Jan 11	Intro to ML Decision Trees Slides video	• Machine learning examples • Well defined machine learning problem • Decision tree learning	Mitchell: Ch 3 Bishop: Ch 14.4 The Discipline of Machine Learning	
Jan 13	Decision Tree learning Review of Probability Annotated slides video	• The big picture • Overfitting • Random variables, probabilities	Andrew Moore's Basic Probability Tutorial Bishop: Ch. 1 thru 1.2.3 Bishop: Ch 2 thru 2.2	HW1 out Jan 14
Jan 18	Probability and Estimation Annotated slides video	• Bayes rule • MLE • MAP	Andrew Moore's Basic Probability Tutorial Bishop: Ch. 1 thru 1.2.3 Bishop: Ch 2 thru 2.2	
Jan 20	Naive Bayes Annotated slides video	• Conditional independence • Multinomial Naive Bayes	Mitchell: Naive Bayes and Logistic Regression	

	video			
Jan 20	Naive Bayes Annotated slides video	• Conditional independence • Multinomial Naive Bayes	Mitchell: Naive Bayes and Logistic Regression	
Jan 25	Gaussian Naive Bayes Slides Annotated Slides video	• Gaussian Bayes classifiers • Document classification • Brain image classification • Form of decision surfaces	Mitchell: Naive Bayes and Logistic Regression	HW1 due HW2 out
Jan 27	Logistic Regression Slides Annotated slides video	• Naive Bayes - the big picture • Logistic Regression: Maximizing conditional likelihood • Gradient ascent as a general learning/optimization method	Mitchell: Naive Bayes and Logistic Regression Ng & Jordan: On Discriminative and Generative Classifiers, NIPS, 2001.	
Feb 1	Linear Regression Slides Annotated slides video	• Generative/Discriminative models • minimizing squared error and maximizing data likelihood • bias-variance decomposition • regularization		
Feb 3	Practical Issues	• Feature selection • Overfitting • Bias-Variance tradeoff		
Feb 8	Graphical models 1 Annotated slides video	• Bayes nets • representing joint distributions with conditional independence assumptions	Bishop: Ch 8, through 8.2	HW3 out
Feb 15	Graphical models 2 slides video	• D-separation and Conditional Independence • Inference • Learning from fully observed data • Learning from partially observed data		
Feb 17	Graphical models 3 annotated slides video	• EM	EM and HMM tutorial J. Bilmes	

Feb 22	Graphical models 4 annotated slides video	• Mixture of Gaussians clustering • Learning Bayes Net structure - Chow Liu	Intro. to Graphical Models, K. Murphy Graphical Models tutorial, M. Jordan	HW3 due HW4 out
Feb 24	Computational Learning Theory annotated slides video	• PAC Learning	Mitchell: Ch. 7	
Mar 1	Midterm Review PAC learning slides midterm review slides video			HW4 due
Mar 3	Midterm Exam	• in class • open notes, open book, no internet		Midterm Solution
Mar 15	Computational Learning Theory annotated slides video	• Mistake bounds • Weighted Majority Algorithm	Mitchell: Ch. 7	
Mar 17	Semi-Supervised Learning slides: CoTraining NELL video	• CoTraining / Multi-view Learning • Never ending learning (NELL)	• Cotraining: Blum & Mitchell • NELL: Carlson et al., 2010	
Mar 22	Hidden Markov Models annotated slides	• Markov models • HMM's and Bayes Nets • Other probabilistic time series models	Bishop Ch. 13	
Mar 24	Neural Networks slides video	• Non-linear regression • Backpropagation and Gradient descent • Learning hidden layer representations	Mitchell Ch. 4 Bishop Ch. 5	Project proposals due
Mar 29	Learning Representations I slides video	• Artificial neural networks • PCA	Bishop Ch. 12 through 12.1 A Tutorial on PCA, J. Schlens SVD and PCA, Wall et al.	

Mar 31	Learning Representations II slides video	• Deep belief networks • ICA • CCA	Deep Belief Nets paper, Hinton & Salakhutdinov CCA Tutorial, M. Borga	
Apr 5	Learning Representations III slides video	• Fisher Linear Discriminant • Latent Dirichlet Allocation • Intro to Kernel Functions	Bishop Ch. 6.1 (required) Bishop Ch. 6.2, 6.3 (optional)	
Apr 7	Kernel Methods and SVM's slides video	• Regression: Primal and Dual forms • Kernels and Kernel Regression • SVMs	Bishop Ch. 6.1 Bishop Ch. 7, through 7.1.2	
Apr 12	SVM's II slides video	• Maximizing the margin • Noise and soft margin SVM's • PAC learning and SVM's • Hinge loss, log loss, 0-1 loss	Bishop Ch. 7, through 7.1.2	Project midway report due
Apr 14		No CMU classes today		
Apr 19	Active Learning slides video	Guest lecture: Dr. Burr Settles • Uncertainty sampling • Query by committee	Settles: Active learning survey	
Apr 21	ML in Computational Biology slides video	Guest lecture: Prof. Ziv Bar-Joseph		
Apr 26	Reinforcement Learning I slides video	• Markov Decision Processes • Value Iteration • Q learning	Kaelbling et al.: Reinforcement Learning: A Survey	
Apr 28	Reinforcement Learning 2 RL slides Final study guide video	• Q learning in non-deterministic domains • RL as model for learning in animals • Final exam review		

几个需要体会的名词：

* 框架 (Frame)
* 模型 (Model)
* 算法 (Algorithm)

1.关于归纳学习

- 归纳学习（Inductive Learning）旨在从大量的经验数据中归纳抽取出一般的判定规则和模式。它是机器学习（符号智能）最核心最成熟的分支。
- 归纳学习依赖于经验数据，因此又叫经验学习（Empirical Learning）。
- 由于归纳学习依赖于数据间的相似性，所以也叫基于相似性学习（Similarity-based Learning）。
- 归纳学习的三个基本论题：1.计算复杂性，2.概率逼近的正确性，3.可理解性。

2.归纳学习的分支

- 归纳学习
  - 有导师学习 Supervised Learning
    又称“示例学习”
    “概念学习”
    - 覆盖算法 Covering Algorithms　e.g., AQ11, AQ15
    - 分治算法 Devided and Comquer　e.g., CART, ID3/C4.5
  - 无导师学习 UnSupervised Learning
    - 概念聚类 Conceptual Clustering
    - 机器发现 Machine Discovery　e.g., KDD

（Knowledge Discovery in Database, KDD）是所谓“数据挖掘”的一种更广义的说法。

3.学习数据

Iris数据集是常用的分类实验数据集，由Fisher收集整理。Iris也称鸢尾花卉数据集，是一种多重变量分析的数据集。数据集包含150个数据样本，分为3类，每类50个数据，每个数据包含4个属性。可通过花萼长度，花萼宽度，花瓣长度，花瓣宽度4个属性预测鸢尾花卉属于（Setosa，Versicolour，Virginica）三个种类中的哪一类。

☞Title: Iris Plants Database

☞ Sources:
(a) Creator: R.A. Fisher
(b) Donor: Michael Marshall
(c) Date: July, 1988

☞ Attribute Information:
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
   -- Iris Setosa
   -- Iris Versicolour
   -- Iris Virginica

1	5.1,3.5,1.4,0.2,Iris-setosa
2	4.9,3.0,1.4,0.2,Iris-setosa
3	4.7,3.2,1.3,0.2,Iris-setosa
4	4.6,3.1,1.5,0.2,Iris-setosa
5	5.0,3.6,1.4,0.2,Iris-setosa
6	5.4,3.9,1.7,0.4,Iris-setosa
7	4.6,3.4,1.4,0.3,Iris-setosa
8	5.0,3.4,1.5,0.2,Iris-setosa
9	4.4,2.9,1.4,0.2,Iris-setosa
10	4.9,3.1,1.5,0.1,Iris-setosa
11	5.4,3.7,1.5,0.2,Iris-setosa
12	4.8,3.4,1.6,0.2,Iris-setosa
13	4.8,3.0,1.4,0.1,Iris-setosa
14	4.3,3.0,1.1,0.1,Iris-setosa
15	5.8,4.0,1.2,0.2,Iris-setosa
16	5.7,4.4,1.5,0.4,Iris-setosa
17	5.4,3.9,1.3,0.4,Iris-setosa
18	5.1,3.5,1.4,0.3,Iris-setosa
19	5.7,3.8,1.7,0.3,Iris-setosa
20	5.1,3.8,1.5,0.3,Iris-setosa
21	5.4,3.4,1.7,0.2,Iris-setosa
22	5.1,3.7,1.5,0.4,Iris-setosa
23	4.6,3.6,1.0,0.2,Iris-setosa
24	5.1,3.3,1.7,0.5,Iris-setosa
25	4.8,3.4,1.9,0.2,Iris-setosa
26	5.0,3.0,1.6,0.2,Iris-setosa
27	5.0,3.4,1.6,0.4,Iris-setosa
28	5.2,3.5,1.5,0.2,Iris-setosa
29	5.2,3.4,1.4,0.2,Iris-setosa
30	4.7,3.2,1.6,0.2,Iris-setosa
31	4.8,3.1,1.6,0.2,Iris-setosa
32	5.4,3.4,1.5,0.4,Iris-setosa
33	5.2,4.1,1.5,0.1,Iris-setosa
34	5.5,4.2,1.4,0.2,Iris-setosa
35	4.9,3.1,1.5,0.1,Iris-setosa
36	5.0,3.2,1.2,0.2,Iris-setosa
37	5.5,3.5,1.3,0.2,Iris-setosa
38	4.9,3.1,1.5,0.1,Iris-setosa
39	4.4,3.0,1.3,0.2,Iris-setosa
40	5.1,3.4,1.5,0.2,Iris-setosa
41	5.0,3.5,1.3,0.3,Iris-setosa
42	4.5,2.3,1.3,0.3,Iris-setosa
43	4.4,3.2,1.3,0.2,Iris-setosa
44	5.0,3.5,1.6,0.6,Iris-setosa
45	5.1,3.8,1.9,0.4,Iris-setosa
46	4.8,3.0,1.4,0.3,Iris-setosa
47	5.1,3.8,1.6,0.2,Iris-setosa
48	4.6,3.2,1.4,0.2,Iris-setosa
49	5.3,3.7,1.5,0.2,Iris-setosa
50	5.0,3.3,1.4,0.2,Iris-setosa
51	7.0,3.2,4.7,1.4,Iris-versicolor
52	6.4,3.2,4.5,1.5,Iris-versicolor
53	6.9,3.1,4.9,1.5,Iris-versicolor
54	5.5,2.3,4.0,1.3,Iris-versicolor
55	6.5,2.8,4.6,1.5,Iris-versicolor
56	5.7,2.8,4.5,1.3,Iris-versicolor
57	6.3,3.3,4.7,1.6,Iris-versicolor
58	4.9,2.4,3.3,1.0,Iris-versicolor
59	6.6,2.9,4.6,1.3,Iris-versicolor
60	5.2,2.7,3.9,1.4,Iris-versicolor
61	5.0,2.0,3.5,1.0,Iris-versicolor
62	5.9,3.0,4.2,1.5,Iris-versicolor
63	6.0,2.2,4.0,1.0,Iris-versicolor
64	6.1,2.9,4.7,1.4,Iris-versicolor
65	5.6,2.9,3.6,1.3,Iris-versicolor
66	6.7,3.1,4.4,1.4,Iris-versicolor
67	5.6,3.0,4.5,1.5,Iris-versicolor
68	5.8,2.7,4.1,1.0,Iris-versicolor
69	6.2,2.2,4.5,1.5,Iris-versicolor
70	5.6,2.5,3.9,1.1,Iris-versicolor
71	5.9,3.2,4.8,1.8,Iris-versicolor
72	6.1,2.8,4.0,1.3,Iris-versicolor
73	6.3,2.5,4.9,1.5,Iris-versicolor
74	6.1,2.8,4.7,1.2,Iris-versicolor
75	6.4,2.9,4.3,1.3,Iris-versicolor
76	6.6,3.0,4.4,1.4,Iris-versicolor
77	6.8,2.8,4.8,1.4,Iris-versicolor
78	6.7,3.0,5.0,1.7,Iris-versicolor
79	6.0,2.9,4.5,1.5,Iris-versicolor
80	5.7,2.6,3.5,1.0,Iris-versicolor
81	5.5,2.4,3.8,1.1,Iris-versicolor
82	5.5,2.4,3.7,1.0,Iris-versicolor
83	5.8,2.7,3.9,1.2,Iris-versicolor
84	6.0,2.7,5.1,1.6,Iris-versicolor
85	5.4,3.0,4.5,1.5,Iris-versicolor
86	6.0,3.4,4.5,1.6,Iris-versicolor
87	6.7,3.1,4.7,1.5,Iris-versicolor
88	6.3,2.3,4.4,1.3,Iris-versicolor
89	5.6,3.0,4.1,1.3,Iris-versicolor
90	5.5,2.5,4.0,1.3,Iris-versicolor
91	5.5,2.6,4.4,1.2,Iris-versicolor
92	6.1,3.0,4.6,1.4,Iris-versicolor
93	5.8,2.6,4.0,1.2,Iris-versicolor
94	5.0,2.3,3.3,1.0,Iris-versicolor
95	5.6,2.7,4.2,1.3,Iris-versicolor
96	5.7,3.0,4.2,1.2,Iris-versicolor
97	5.7,2.9,4.2,1.3,Iris-versicolor
98	6.2,2.9,4.3,1.3,Iris-versicolor
99	5.1,2.5,3.0,1.1,Iris-versicolor
100	5.7,2.8,4.1,1.3,Iris-versicolor
101	6.3,3.3,6.0,2.5,Iris-virginica
102	5.8,2.7,5.1,1.9,Iris-virginica
103	7.1,3.0,5.9,2.1,Iris-virginica
104	6.3,2.9,5.6,1.8,Iris-virginica
105	6.5,3.0,5.8,2.2,Iris-virginica
106	7.6,3.0,6.6,2.1,Iris-virginica
107	4.9,2.5,4.5,1.7,Iris-virginica
108	7.3,2.9,6.3,1.8,Iris-virginica
109	6.7,2.5,5.8,1.8,Iris-virginica
110	7.2,3.6,6.1,2.5,Iris-virginica
111	6.5,3.2,5.1,2.0,Iris-virginica
112	6.4,2.7,5.3,1.9,Iris-virginica
113	6.8,3.0,5.5,2.1,Iris-virginica
114	5.7,2.5,5.0,2.0,Iris-virginica
115	5.8,2.8,5.1,2.4,Iris-virginica
116	6.4,3.2,5.3,2.3,Iris-virginica
117	6.5,3.0,5.5,1.8,Iris-virginica
118	7.7,3.8,6.7,2.2,Iris-virginica
119	7.7,2.6,6.9,2.3,Iris-virginica
120	6.0,2.2,5.0,1.5,Iris-virginica
121	6.9,3.2,5.7,2.3,Iris-virginica
122	5.6,2.8,4.9,2.0,Iris-virginica
123	7.7,2.8,6.7,2.0,Iris-virginica
124	6.3,2.7,4.9,1.8,Iris-virginica
125	6.7,3.3,5.7,2.1,Iris-virginica
126	7.2,3.2,6.0,1.8,Iris-virginica
127	6.2,2.8,4.8,1.8,Iris-virginica
128	6.1,3.0,4.9,1.8,Iris-virginica
129	6.4,2.8,5.6,2.1,Iris-virginica
130	7.2,3.0,5.8,1.6,Iris-virginica
131	7.4,2.8,6.1,1.9,Iris-virginica
132	7.9,3.8,6.4,2.0,Iris-virginica
133	6.4,2.8,5.6,2.2,Iris-virginica
134	6.3,2.8,5.1,1.5,Iris-virginica
135	6.1,2.6,5.6,1.4,Iris-virginica
136	7.7,3.0,6.1,2.3,Iris-virginica
137	6.3,3.4,5.6,2.4,Iris-virginica
138	6.4,3.1,5.5,1.8,Iris-virginica
139	6.0,3.0,4.8,1.8,Iris-virginica
140	6.9,3.1,5.4,2.1,Iris-virginica
141	6.7,3.1,5.6,2.4,Iris-virginica
142	6.9,3.1,5.1,2.3,Iris-virginica
143	5.8,2.7,5.1,1.9,Iris-virginica
144	6.8,3.2,5.9,2.3,Iris-virginica
145	6.7,3.3,5.7,2.5,Iris-virginica
146	6.7,3.0,5.2,2.3,Iris-virginica
147	6.3,2.5,5.0,1.9,Iris-virginica
148	6.5,3.0,5.2,2.0,Iris-virginica
149	6.2,3.4,5.4,2.3,Iris-virginica
150	5.9,3.0,5.1,1.8,Iris-virginica

22	5.1,3.7,1.5,0.4,Iris-setosa
23	4.6,3.6,1.0,0.2,Iris-setosa
24	5.1,3.3,1.7,0.5,Iris-setosa
25	4.8,3.4,1.9,0.2,Iris-setosa

3.学习数据（符号）

Balloon databases

Attribute Information: (Classes Inflated T or F)

Color	yellow, purple
size	large, small
act	stretch, dip
age	adult, child
inflated	T, F

YELLOW,SMALL,STRETCH,ADULT,T
YELLOW,SMALL,STRETCH,CHILD,T
YELLOW,SMALL,DIP,ADULT,T
YELLOW,SMALL,DIP,CHILD,T
YELLOW,SMALL,STRETCH,ADULT,T
YELLOW,SMALL,STRETCH,CHILD,T
YELLOW,SMALL,DIP,ADULT,T
YELLOW,SMALL,DIP,CHILD,T
YELLOW,LARGE,STRETCH,ADULT,F
YELLOW,LARGE,STRETCH,CHILD,F
YELLOW,LARGE,DIP,ADULT,F
YELLOW,LARGE,DIP,CHILD,F
PURPLE,SMALL,STRETCH,ADULT,F
PURPLE,SMALL,STRETCH,CHILD,F
PURPLE,SMALL,DIP,ADULT,F
PURPLE,SMALL,DIP,CHILD,F
PURPLE,LARGE,STRETCH,ADULT,F
PURPLE,LARGE,STRETCH,CHILD,F
PURPLE,LARGE,DIP,ADULT,F
PURPLE,LARGE,DIP,CHILD,F

YELLOW,SMALL,STRETCH,ADULT,T
YELLOW,SMALL,STRETCH,CHILD,T
YELLOW,SMALL,DIP,ADULT,T
YELLOW,SMALL,DIP,CHILD,T
YELLOW,LARGE,STRETCH,ADULT,T
YELLOW,LARGE,STRETCH,CHILD,F
YELLOW,LARGE,DIP,ADULT,F
YELLOW,LARGE,DIP,CHILD,F
PURPLE,SMALL,STRETCH,ADULT,T
PURPLE,SMALL,STRETCH,CHILD,F
PURPLE,SMALL,DIP,ADULT,F
PURPLE,SMALL,DIP,CHILD,F
PURPLE,LARGE,STRETCH,ADULT,T
PURPLE,LARGE,STRETCH,CHILD,F
PURPLE,LARGE,DIP,ADULT,F
PURPLE,LARGE,DIP,CHILD,F

YELLOW,SMALL,STRETCH,ADULT,T
YELLOW,SMALL,STRETCH,ADULT,T
YELLOW,SMALL,STRETCH,CHILD,F
YELLOW,SMALL,DIP,ADULT,F
YELLOW,SMALL,DIP,CHILD,F
YELLOW,LARGE,STRETCH,ADULT,T
YELLOW,LARGE,STRETCH,ADULT,T
YELLOW,LARGE,STRETCH,CHILD,F
YELLOW,LARGE,DIP,ADULT,F
YELLOW,LARGE,DIP,CHILD,F
PURPLE,SMALL,STRETCH,ADULT,T
PURPLE,SMALL,STRETCH,ADULT,T
PURPLE,SMALL,STRETCH,CHILD,F
PURPLE,SMALL,DIP,ADULT,F
PURPLE,SMALL,DIP,CHILD,F
PURPLE,LARGE,STRETCH,ADULT,T
PURPLE,LARGE,STRETCH,ADULT,T
PURPLE,LARGE,STRETCH,CHILD,F
PURPLE,LARGE,DIP,ADULT,F
PURPLE,LARGE,DIP,CHILD,F

YELLOW,SMALL,STRETCH,ADULT,T
YELLOW,SMALL,STRETCH,CHILD,T
YELLOW,SMALL,DIP,ADULT,T
YELLOW,SMALL,DIP,CHILD,F
YELLOW,SMALL,DIP,CHILD,F
YELLOW,LARGE,STRETCH,ADULT,T
YELLOW,LARGE,STRETCH,CHILD,T
YELLOW,LARGE,DIP,ADULT,T
YELLOW,LARGE,DIP,CHILD,F
YELLOW,LARGE,DIP,CHILD,F
PURPLE,SMALL,STRETCH,ADULT,T
PURPLE,SMALL,STRETCH,CHILD,T
PURPLE,SMALL,DIP,ADULT,T
PURPLE,SMALL,DIP,CHILD,F
PURPLE,SMALL,DIP,CHILD,F
PURPLE,LARGE,STRETCH,ADULT,T
PURPLE,LARGE,STRETCH,CHILD,T
PURPLE,LARGE,DIP,ADULT,T
PURPLE,LARGE,DIP,CHILD,F
PURPLE,LARGE,DIP,CHILD,F

3.学习数据

是否适合
打高尔夫球问题：

怎样做
归纳？

利用这14条例子（样本），找出合适的规律（或规则）！

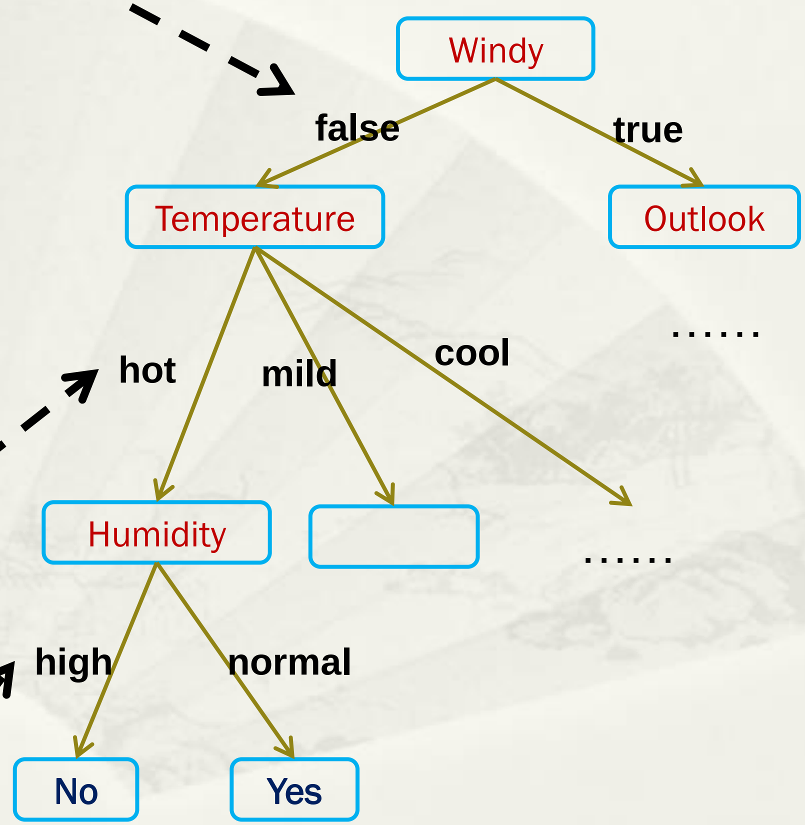
属性	属性	属性	属性	类别
Outlook	Temperature	Humidity	Windy	Play ?
Sunny	85（hot）	85（high）	False	Don’t Play
Sunny	80（hot）	90（high）	True	Don’t Play
Overcast	83（hot）	78（normal）	False	Play
Rain	70（mild）	96（high）	False	Play
Rain	68（cool）	80（normal）	False	Play
Rain	65（cool）	70（normal）	True	Don’t Play
Overcast	64（cool）	65（normal）	True	Play
Sunny	72（mild）	95（high）	False	Don’t Play
Sunny	69（cool）	70（normal）	False	Play
Rain	75（mild）	81（high）	False	Play
Sunny	75（mild）	70（normal）	True	Play
Overcast	72（mild）	90（high）	True	Play
Overcast	81（hot）	75（normal）	False	Play
Rain	71（mild）	81（high）	True	Don’t Play

4. 决策树及最优决策树

Outlook	Temperature	Humidity	Windy	Play ?
Sunny	hot	high	False	No
Overcast	hot	normal	False	Yes
Rain	mild	high	False	Yes
Rain	cool	normal	False	Yes
Sunny	mild	high	False	No
Sunny	cool	normal	False	Yes
Rain	mild	high	False	Yes
Overcast	hot	normal	False	Yes

Outlook	Temperature	Humidity	Windy	Play ?
Sunny	hot	high	False	No
Overcast	hot	normal	False	Yes
Overcast	hot	normal	False	Yes

Outlook	Temperature	Humidity	Windy	Play ?
Sunny	hot	high	False	No



5. ID3/C4.5决策树算法

J. Ross Quinlan

@符号主义智能： Symbolism　@归纳学习：Inductive Learning

（1） 由类别获得信息.　$\mathrm{Info}(Y,N) = -\left(\frac{9}{14}\log\frac{9}{14} + \frac{5}{14}\log\frac{5}{14}\right) = 0.940$

（2）在当前窗口下查看并计算每个属性的熵。

$$\begin{aligned}\mathrm{E(Outlook)} &= \frac{5}{14}\mathrm{Info(sunny)} + \frac{4}{14}\mathrm{Info(overcast)} + \frac{5}{14}\mathrm{Info(rain)} \\ &= \frac{5}{14}\times 0.971 + \frac{4}{14}\times 0 + \frac{5}{14}\times 0.971 = {\color{red}0.694}\end{aligned}$$

$$\begin{aligned}\mathrm{E(Temperature)} &= \frac{4}{14}\mathrm{Info(hot)} + \frac{6}{14}\mathrm{Info(mild)} + \frac{4}{14}\mathrm{Info(cool)} \\ &= \frac{4}{14}\mathrm{Info}(2,2) + \frac{6}{14}\mathrm{Info}(2,4) + \frac{4}{14}\mathrm{Info}(1,3) = 0.911\end{aligned}$$

$$\mathrm{E(Humidity)} = \frac{6}{14}\mathrm{Info(high)} + \frac{8}{14}\mathrm{Info(normal)} = 0.789$$

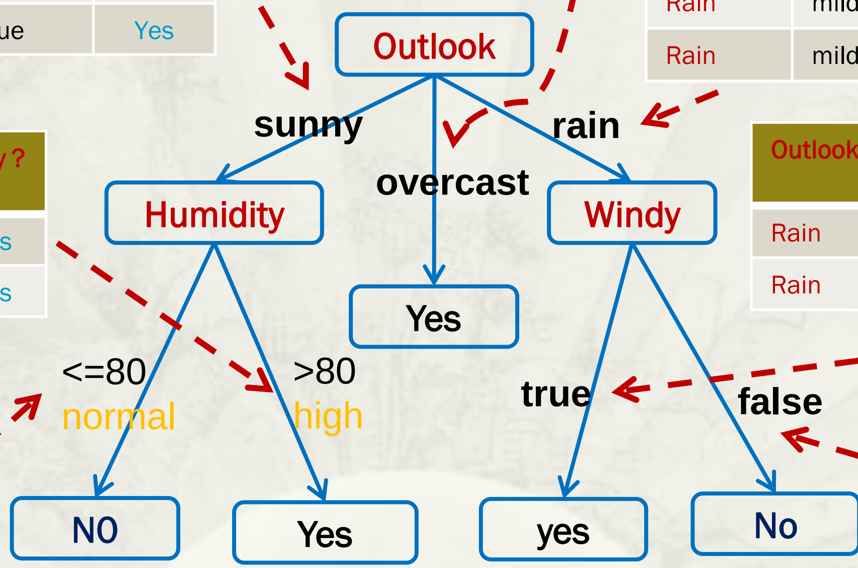
$$\mathrm{E(Windy)} = \frac{8}{14}\mathrm{Info(false)} + \frac{6}{14}\mathrm{Info(true)} = 0.892$$

（3）对每个数据子集递归调用（2），直到每个窗口下的数据都为同一类别，结束。

Outlook	Temper.	Humidity	Windy	Play ?
Overcast	hot	normal	False	Yes
Overcast	cool	normal	True	Yes
Overcast	mild	high	True	Yes
Overcast	hot	normal	False	Yes

Outlook	Temper.	Humidity	Windy	Play ?
Sunny	hot	high	False	No
Sunny	hot	high	True	No
Sunny	mild	high	False	No
Sunny	cool	normal	False	Yes
Sunny	mild	normal	True	Yes

Outlook	Temper.	Humidity	Windy	Play ?
Rain	mild	high	False	Yes
Rain	cool	normal	False	Yes
Rain	cool	normal	True	No
Rain	mild	high	False	Yes
Rain	mild	high	True	No



Outlook	Temper.	Humidity	Windy	Play ?
Sunny	cool	normal	False	Yes
Sunny	mild	normal	True	Yes

Outlook	Temper.	Humidity	Windy	Play ?
Rain	cool	normal	True	No
Rain	mild	high	True	No

Outlook	Temper.	Humidity	Windy	Play ?
Sunny	hot	high	False	No
Sunny	hot	high	True	No
Sunny	mild	high	False	No

Outlook	Temper.	Humidity	Windy	Play ?
Rain	mild	high	False	Yes
Rain	cool	normal	False	Yes
Rain	mild	high	False	Yes

例如规则：

If <Outlook=sunny> and <Humidity=high> then <Yes>

（1）以上为ID3算法采用的决策树属性选择策略，即，熵最小准则。

（2）第2种属性选择策略：熵增益最大。例如，

$$\text{Gain(Outlook)} = \text{I(Y, N)} - \text{E(Outlook)} = 0.940 - 0.694 = 0.246;$$
$$\text{Gain(Temperature)} = 0.029;$$
$$\text{Gain(Humidity)} = 0.151;$$
$$\text{Gain(Windy)} = 0.048$$

（3）第3种属性选择策略：熵增益率最大。即，

$$\text{GainRatio(D, T)} = \frac{Gain(D,T)}{SplitInfo(D,T)}$$

例如，

$$\text{SplitInfo(Outlook, T)} = -\frac{5}{14}log_2\left(\frac{5}{14}\right) - \frac{4}{14}log_2\left(\frac{4}{14}\right) - \frac{5}{14}log_2\left(\frac{5}{14}\right) = 1.577$$

$$\text{GainRatio(Outlook, T)} = \frac{0.246}{1.577} = 0.156$$

6. 关于ID3/C4.5的扩展议题（讨论）

- 当T_i的样本数目不多（很少）时，熵度量不可靠，可使用其它的概率方法度量。
- 最小决策树准则是不可靠的。（Occam’s Razor是可疑的）。这一结论初见于1992年，此后有了Boosting, Bagging, SVM等方法。
- 决策树与规则集，树的修剪。
- 连续数值的归纳问题，混合数据学习问题。
- 增量学习问题（Incremental Learning, e.g., ID4, ID5）。

7.动态聚类与概念聚类

- [动态聚类]
- Step 1.已知观察事件（例子）E={e_1，e_2，e_3,……,e_r}，K是指定类别数。
- Step 2. 从E中选K个种子，$\widetilde{e_1}$,……, $\widetilde{e_K}$。
- Step 3. 对E中所有其它事件e_i求j使dis(e_i , $\widetilde{e_j}$)=minimum ,其中dis为欧氏距离，并把e_i归入第j类。
- Step 4. 重新选择K个种子$\widetilde{e_1'}$, … …, $\widetilde{e_K'}$使$\widetilde{e_i}$距离第j类的几何中心最近。
- Step 5. 若条件满足，则结束；否则转Step 3。

动态聚类示意（K=2）

- [概念聚类]
- 修改[动态聚类]第3步。将欧氏距离计算修改为求概念$R_i = cover(e_i | S - \{\tilde{e_i}\})$。
- 当然，这里面用到了覆盖算法AQ。也可以用决策树完成，因为决策树也是规则集。