COMP90051 Statistical Machine Learning

Semester 2, 2015

Lecturer: Ben Rubinstein

COMP90049 Revision

Covered Knowledge

- Supervised vs Unsupervised Learning
- Unsupervised learning
  - association rule mining
  - k-means clustering ***
- Supervised learning
  - naïve Bayes ***
  - instance-based learning (IB1)
  - decision stump/tree induction (0R, 1R, ID5)
    - probability theory; entropy
  - feature selection (mutual information)
  - Evaluation
    - basic sampling (hold-out, cross-validation)
    - metrics: precision/recall/F, ROC

Supervised vs Unsupervised Learning

- Training data: used to construct models

	Training data	Model used for
Supervised learning	Labelled	Predict labels on new instances
Unsupervised learning	Unlabelled	Cluster related instances; Understand attribute relationships

Association Rules: Definitions

- An association rule is an implication $A \to B$, where A and B are disjoint itemsets
  - ⋆ $A =$ **antecedent**
  - ⋆ $B =$ **consequent**

- N.B. in association mining parlance:
  - ⋆ **item** $=$ attribute–value pair ($I =$ set of items)
  - ⋆ **itemset** $=$ set of attribute–value pairs
  - ⋆ k-**itemset** $=$ set of k attribute–value pairs
  - ⋆ **transaction** $=$ exemplar ($T =$ set of transactions)

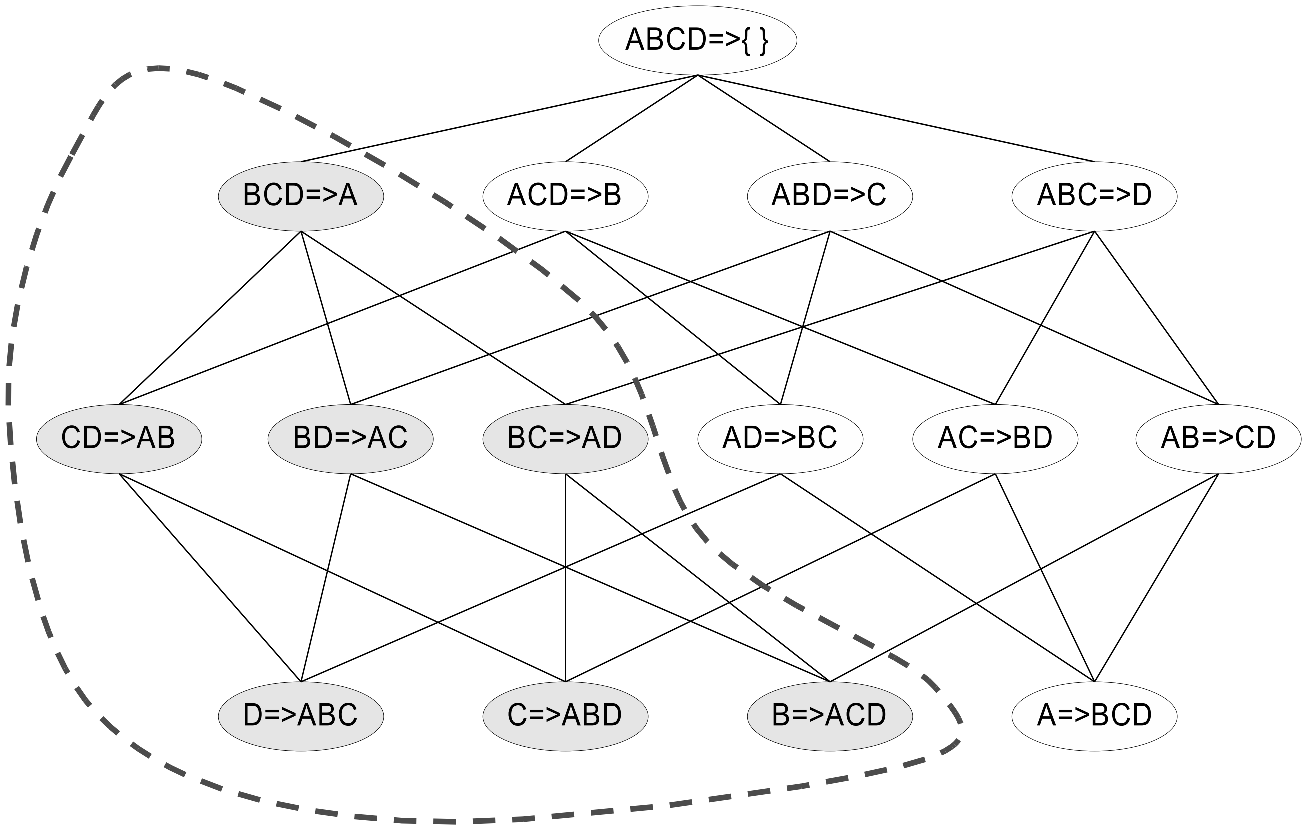
Association Rules: Examples

- An example (transaction) database

Transaction ID	Items
T1	milk, bread, cereal
T2	milk, bread, sugar, eggs
T3	butter, eggs

- Itemset $I = \{milk, bread\}$
- Rule $r = \{milk\} \rightarrow \{bread\}$
- How good is association rule r?
  * support(r) = 2/3
  * confidence (r) = supp($\{milk, bread\}$) / supp($\{milk\}$) = 2/2

APriori Algorithm (Rule Generation)

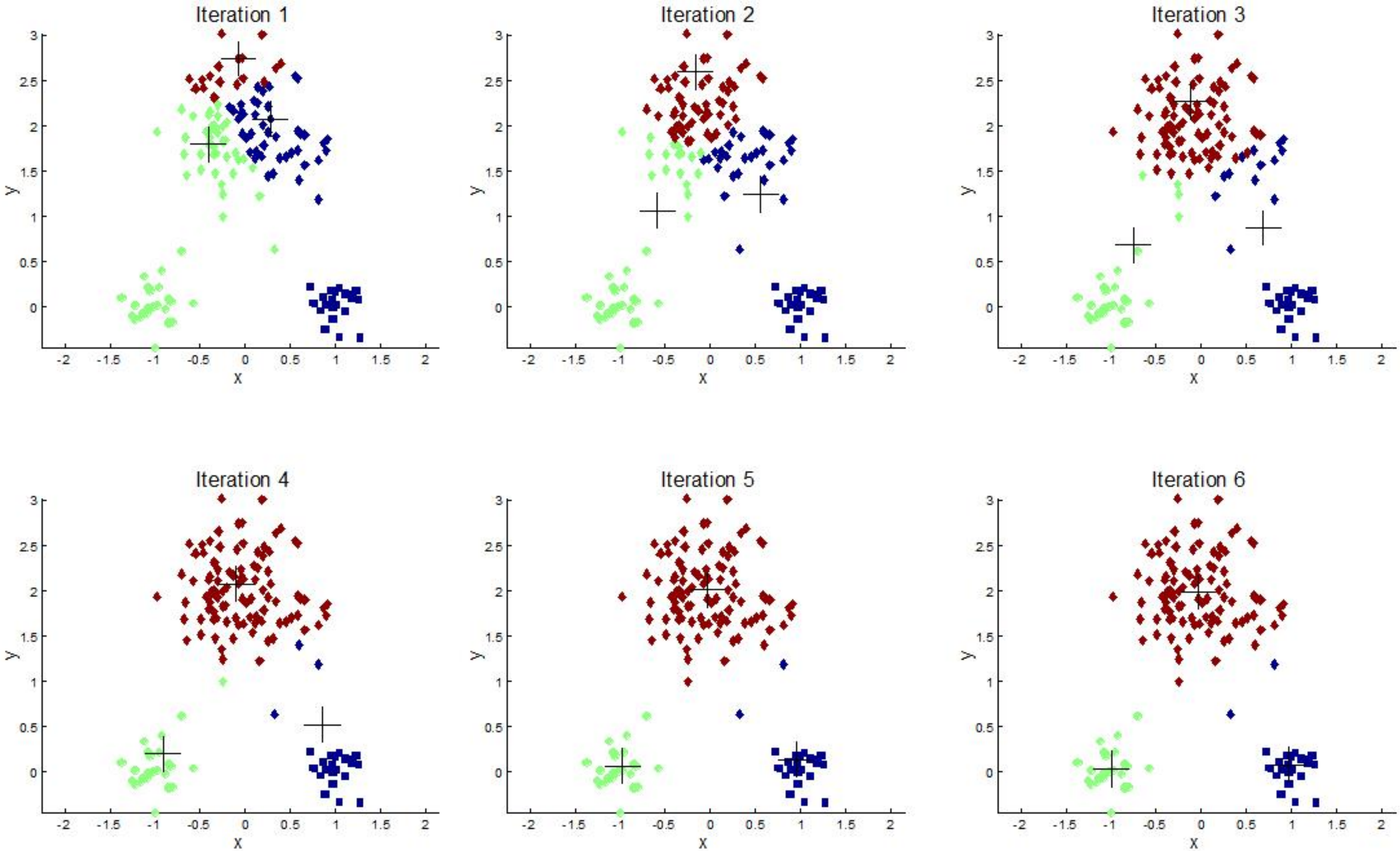


k-means Clustering

- Given k, the k-means algorithm is implemented in four steps:
  1. Select k points at random to act as seed clusters
  2. Compute seed points as the centroids of the clusters of the current partition (the **centroid** is the centre, i.e., mean point, of the cluster)
  3. Assign each instance to the cluster with the nearest centroid
  4. Go back to 2, stop when no reassignments

- Exclusive, deterministic, partitioning, batch clustering method

k-means Clustering: Example



Source: http://www.practicaldb.com/data-visualization-consulting/cluster-analysis/

Naive Bayes (NB) Classifiers

- Classify instance $D = \langle x_1, x_2, ..., x_n \rangle$ as class $c_j \in C$

$$\begin{aligned} c &= \arg\max_{c_j \in C} P(c_j | x_1, x_2, ..., x_n) \\ &= \arg\max_{c_j \in C} \frac{P(x_1, x_2, ..., x_n | c_j) P(c_j)}{P(x_1, x_2, ..., x_n)} \\ &= \arg\max_{c_j \in C} P(x_1, x_2, ..., x_n | c_j) P(c_j) \\ &= \arg\max_{c_j \in C} P(c_j) \prod_{i=1}^{n} P(x_i | c_j) \end{aligned}$$

- Model trained using frequencies

Bayesian Rule

- $P(H|E) = \frac{P(E|H)\times P(H)}{P(E)}$
  - H = hypothesis; $E = evidence$
  - In plain text: $posterior = \frac{likelihood \times prior}{evidence}$
- $P(E)$ will disappear after normalization
  - $P(H|E) \propto P(E|H) \times P(H)$
- The not so naïve assumption of independence
  - $E = \{x_{1,\dots,}x_n\}$
  - $P(E|H) = \prod P(x_i|H)$

Training data

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Model

Training

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

Source: Witten, Ian H. et al.. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

Training data

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Model

Training

$x_1 =$ Outlook			$x_2 =$ Temperature			$x_3 =$ Humidity			$x_4 =$ Windy			$H =$ Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

$P(x_1|H)$ $P(x_2|H)$ $P(x_3|H)$ $P(x_4|H)$ $P(H)$

parameters

Source: Witten, Ian H. et al.. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

Training data

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Model

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

Source: Witten, Ian H. et al.. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

Model

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

A test instance

Outlook	Temperature	Humidity	Windy	Play
Sunny	cool	high	true	?

$$P(yes|E) \propto P(x_1 = sunny|yes)P(x_2 = cool|yes) \times P(x_3 = high|yes) \times P(x_4 = true|yes) \times P(yes) = 0.0053$$

$$P(no|E) \propto P(x_1 = sunny|no)P(x_2 = cool|no) \times P(x_3 = high|no) \times P(x_4 = true|no) \times P(no) = 0.0206$$

$$P(yes|E) = \frac{0.0053}{(0.0053 + 0.0206)} = 0.205$$

Source: Witten, Ian H. et al.. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

Handling Numeric Attributes

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	*std. dev.*	6.2	7.9	*std. dev.*	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

Outlook	Temperature	Humidity	Windy	Play
Sunny	66	90	true	?

Use probability density function: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Given $x = 66, \mu = 73, \sigma = 6.2$:

$$f(temperature = 66|yes) = \frac{1}{\sqrt{2\pi}6.2} e^{-\frac{(66-73)^2}{2\times6.2^2}} = 0.034$$

Source: Witten, Ian H. et al.. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

Quiz

- Is Naïve Bayes a parametric or a non-parametric model?

Nearest Neighbour Classification

- Combining training–test instance scores to form an overall categorisation function:
- **Method 1:** index all training documents, and query the training document set with each test document; classify the test document according to the class of the top-ranked training document **[1-NN]**
- **Method 2:** index all training documents, and query the training document set with each test document; classify the test document according to the **majority class** within the k top-ranked training documents **[k-NN]**

Similarity/Distance Metrics

- Cosine similarity:

$$sim(x,y) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

- Relative entropy:

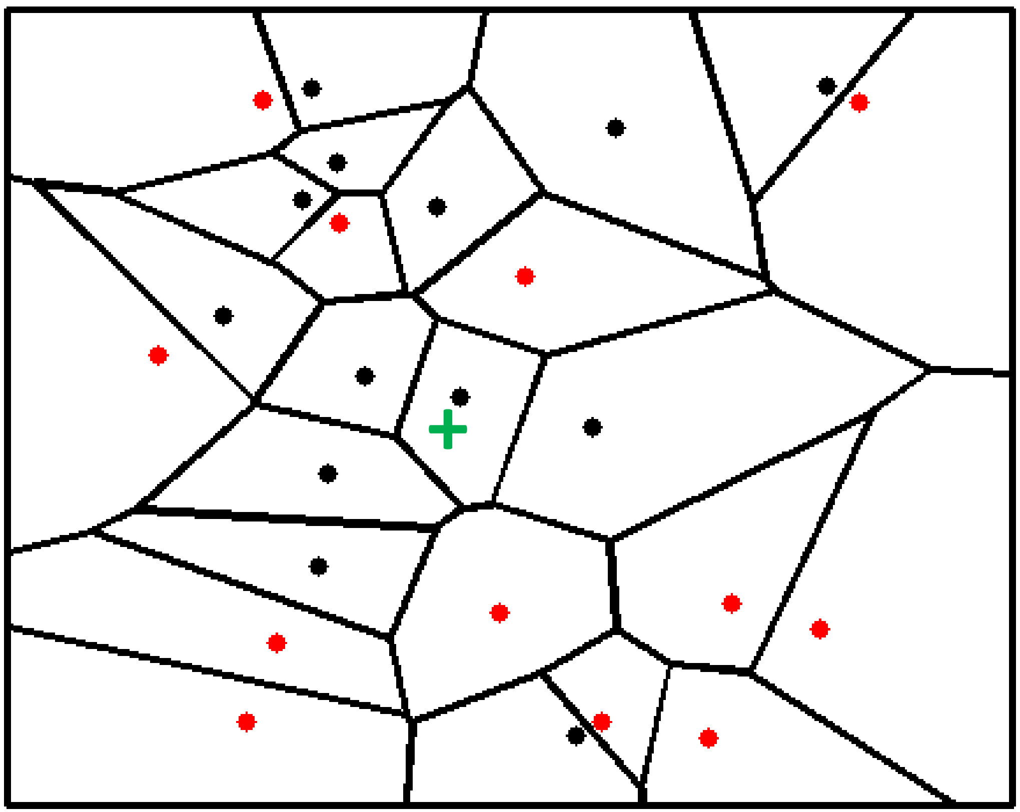
$$D(x \,||\, y) = \sum_i x_i(\log_2 x_i - \log_2 y_i)$$

or alternatively **skew divergence**:

$$s_\alpha(x,y) = D(x \,||\, \alpha y + (1-\alpha)x)$$

1-NN: Example

+ : test instance



Source: http://pmtk3.googlecode.com/svn/trunk/docs/demoOutput/bookDemos/(1)-Introduction/knnVoronoi.html

Constructing Decision Trees: ID3

- **Basic method:** construct decision trees in recursive divide-and-conquer fashion

  FUNCTION ID3 (Root)

  IF all instances at root have same class

  THEN stop

  ELSE Select a new attribute to use in partitioning root node instances

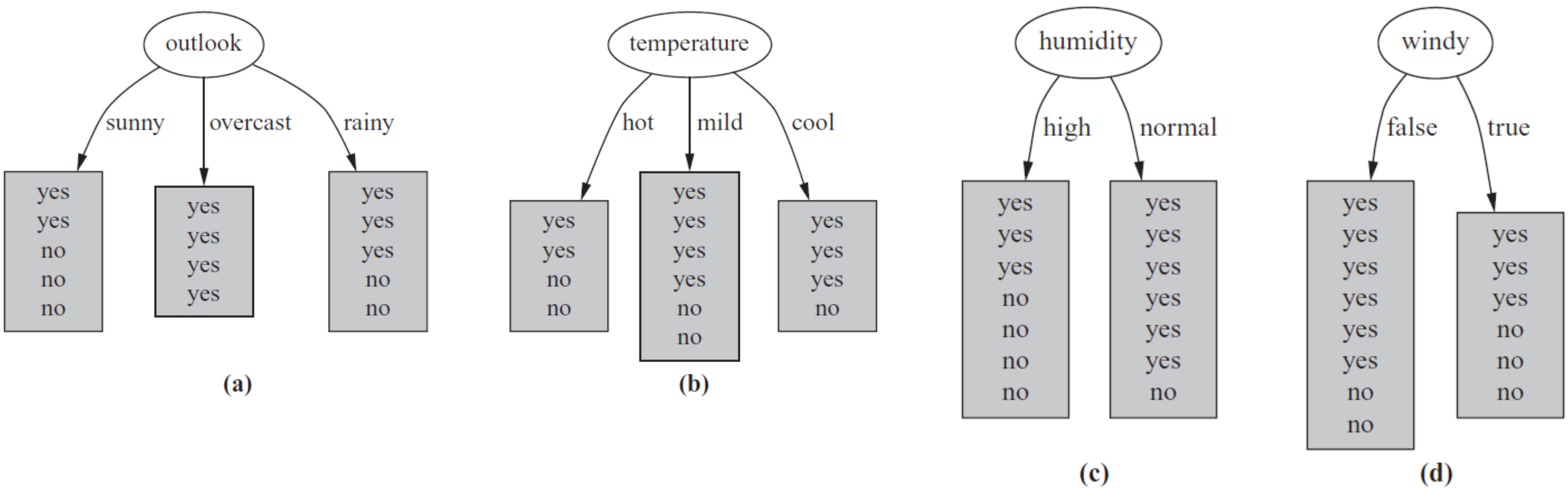
  Create a branch for each attribute value and partition up root node instances according to each value

  Call ID3(LEAF$_i$) for each leaf node LEAF$_i$

- Note: we may not end up with pure leaves

Training data

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no



Which one is the best choice?

Source: Witten, Ian H. et al.. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

Entropy

- The entropy of a discrete random event x with possible states $1, ..n$ is:

$$H(x) = -\sum_{i=1}^{n} P(i) \log_2 P(i)$$

where $0 \log_2 0 =^{def} 0$

Split Criteria

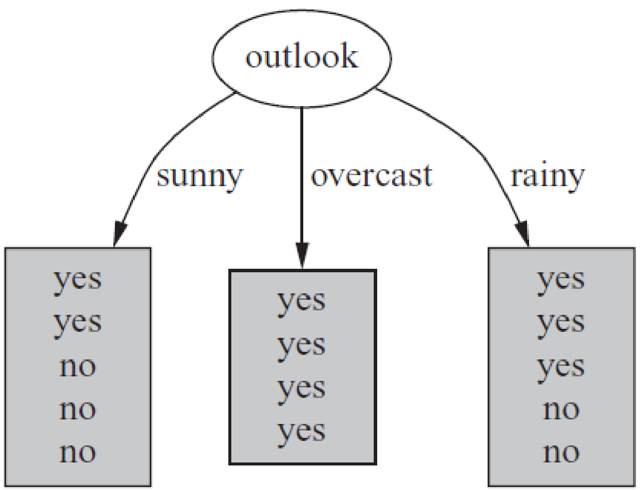
- The **information gain** for attribute R_A (with values $x_1, ...x_m$) at a given root node R is:

$$IG(R_A|R) = H(R) - \sum_{i=1}^{m} P(x_i)H(x_i)$$

- The corresponding **gain ratio** is:

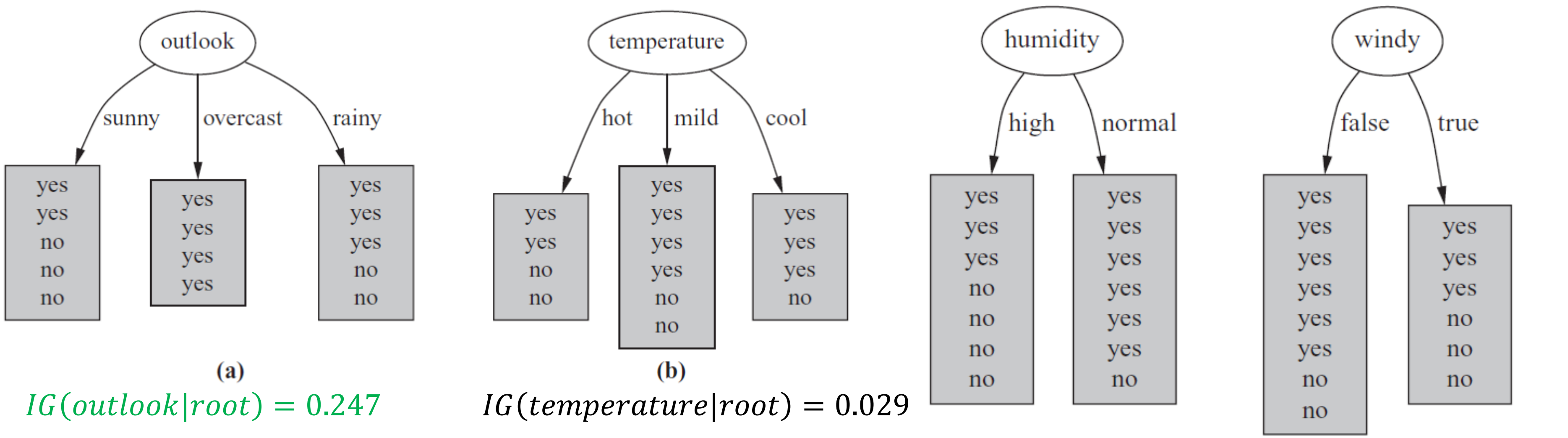
$$\begin{aligned} GR(R_A|R) &= \frac{IG(R_A|R)}{H(R_A)} \\ &= \frac{H(R) - \sum_{i=1}^{m} P(x_i)H(x_i)}{-\sum_{i=1}^{m} P(x_i)\log_2 P(x_i)} \end{aligned}$$

- Before any node was created: $root$: $\#yes = 9$ and $\#no = 5$
  - $Info([9,5]) = Entropy(\frac{9}{14}, \frac{5}{14}) = -\frac{9}{14} log^{\frac{9}{14}} - \frac{5}{14} log^{\frac{5}{14}} = 0.94 bits$



- Entropy for each branch
  - $Info([2,3]) = 0.971 bits$
  - $Info([4,0]) = 0\ bits$
  - $Info([3,2]) = 0.971 bits$
- Average entropy:
  - $Info([2,3],[4,0],[3,2]) = \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 = 0.693\ bits$
- Information gain:
  - $IG(outlook|root) = 0.94 - 0.693 = 0.247\ bits$

Source: Witten, Ian H. et al.. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.



(a) (b) (c) (d)

$IG(outlook|root) = 0.247$

$IG(temperature|root) = 0.029$

$IG(humidity|root) = 0.152$

$IG(hwindy|root) = 0.048$

Source: Witten, Ian H. et al.. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

Quiz

- Is ID3 decision tree a parametric or a non-parametric model?

Feature Selection

- **Mutual information**:

$$MI(T;C) = \sum_{t \in \{0,1\}} \sum_{c} P(t,c) \log_2 \frac{P(t,c)}{P(t)P(c)}$$

Evaluation

- confusion matrix of two-class prediction:

		Predicted Class	
		yes	*no*
Actual Class	*yes*	true positive	false negative
	no	false positive	true negative

Source: Witten, Ian H. et al.. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

Evaluation

- **Classification accuracy**: is the proportion of

$$\mathrm{ACC} = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Error rate**:

$$\mathrm{ER} = \frac{FP + FN}{TP + FP + FN + TN}$$

- **Error rate reduction**:

$$\mathrm{ERR} = \frac{\mathrm{ER}_0 - \mathrm{ER}}{\mathrm{ER}_0}$$

- **Precision**:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F-score**:

$$\text{F-score} = (1 + \beta^2)\frac{PR}{R + \beta^2 P}$$

Sampling

- **Holdout** $=$ train a classifier over a fixed training dataset, and evaluate it over a fixed held-out test dataset
- **Random Subsampling** $=$ perform holdout over multiple iterations, randomly selecting the training and test data (maintaining a fixed size for each dataset) on each iteration
- **Cross Validation** $=$ partition data into N folds, and use $N-1$ as training data and 1 as test $\times N$ iterations
- **Stratified Cross Validation** $=$ partition the data so as to maintain the overall class distribution within individual partitions

ROC Curves

1. Sort the test instances in ascending order of "rating" $t_1, t_2, ..., t_k$

2. Initialise $TP_{k+1} = FP_{k+1} = 0$, and set FN_{k+1} and TN_{k+1} to the number of positive and negative instances in the dataset, resp.

3. For each $i = k, ..., 2, 1$

   i. update TP_i, FP_i, FN_i and TN_i assuming positive classification of instance i, based on the actual class of t_i and TP_{i+1}, FP_{i+1}, FN_{i+1} and TN_{i+1}

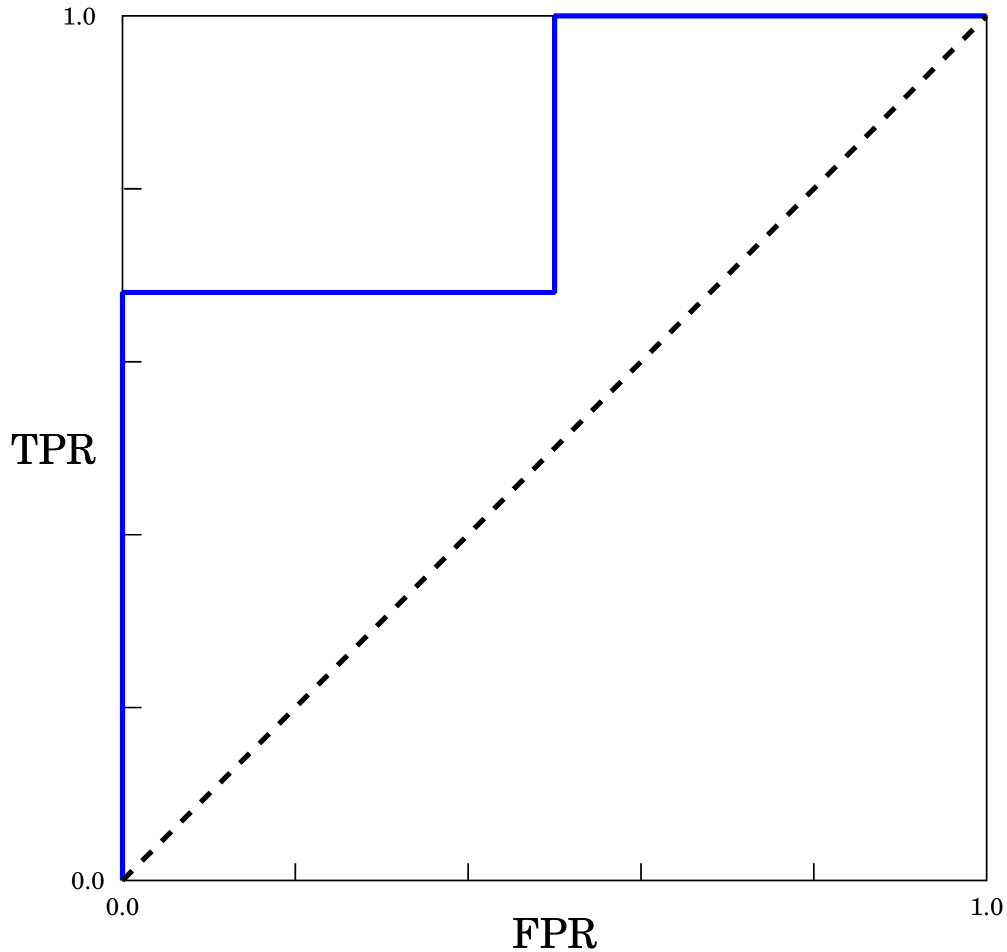
   ii. calculate TPR and FPR at t_i

4. Plot the TPR and FPR values for each t_i

Generating ROC Curves: Example

Class	$-$	$+$	$-$	$+$	$+$	
Score	0.85	0.85	0.87	0.93	0.95	
i	1	2	3	4	5	6
TP	3	3	2	2	1	0
FP	2	1	1	0	0	0
FN	0	0	1	1	2	3
TN	0	1	1	2	2	2
TPR	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	0
FPR	$\frac{2}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0

Generating ROC Curves: Example



What other topics in ML have you seen?