Computer Science 477/577

Entropy based Tree Induction

Lecture 6

Reminder: what this class is about

- Extracting knowledge, patterns, useful information from large data sets
- Specific techniques:
  - Classification
    - Constructing a method of classifying new instances using information in a training set
  - Clustering: subsetting large datasets into meaningful grouping.
  - Association Analysis: determining whether elements tend to occur together
    - Paradigm: market baskets
      - Are beer and diapers purchased together?
  - Sequence mining
    - Finding meaningful, recurring sequences of events

Top-Down Induction of Decision Tree - Problem

- TDIDT algorithm is *underspecified*
  - Does not dictate which attribute to choose to split on
- Some alternatives:
  - – ***takefirst*** – for each branch take the attributes in the order in which they appear in the training set, working from left to right, e.g. for the *degrees* training set in the order *SoftEng*, *ARIN*, *HCI*, *CSA* and *Project*.
  - – ***takelast*** – as for *takefirst*, but working from right to left, e.g. for the *degrees* training set in the order *Project*, *CSA*, *HCI*, *ARIN* and *SoftEng*.
  - – ***random*** – make a random selection (with equal probability of each attribute

Application of Policies

Dataset	take first	take last	random					most	least
			1	2	3	4	5		
contact_lenses	42	27	34	38	32	26	35	42	26
lens24	21	9	15	11	15	13	11	21	9
chess	155	56	94	52	107	90	112	155	52
vote	40	79	96	78	116	110	96	116	40
monk1	60	75	82	53	87	89	80	89	53
monk2	142	112	122	127	109	123	121	142	109
monk3	69	69	43	46	62	55	77	77	43

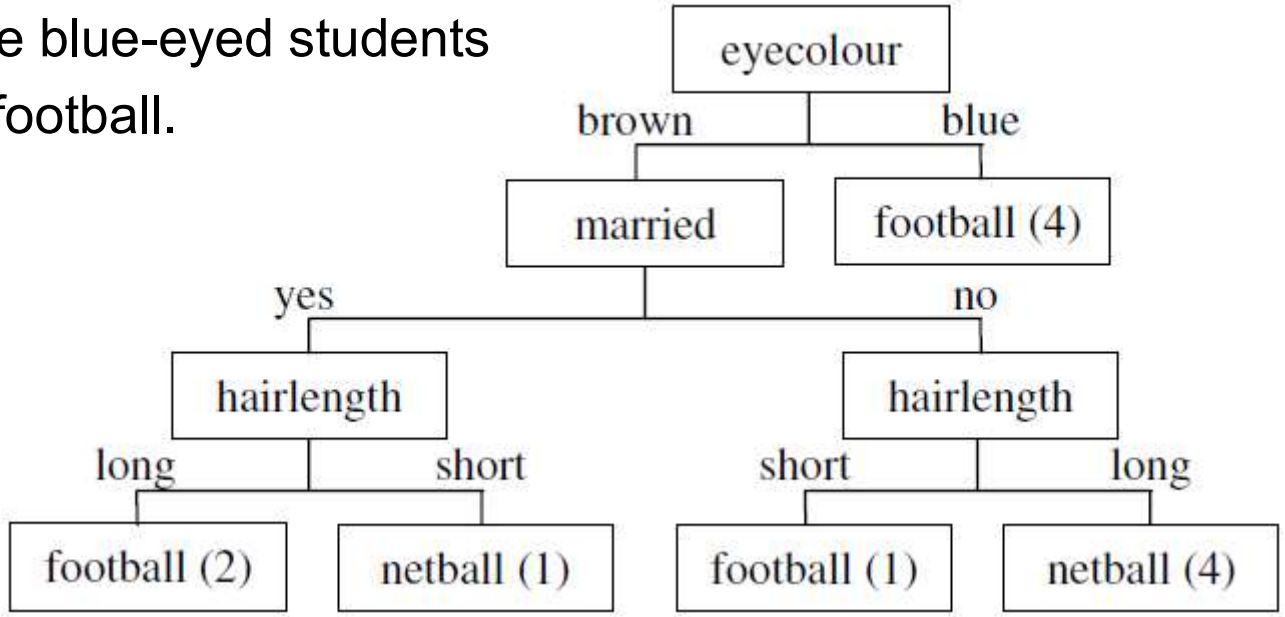
- Number of Branches Generated by TDIDT with Three Attribute Selection Methods

Football / Netball Training set

- University requires its students to enroll in one of its sports clubs
  - Either the Football Club or the Netball Club.
  - Forbidden to join both clubs.
  - Any student joining no club at all will be awarded an automatic failure in their degree (this being considered an important disciplinary offence)
- Training set for twelve students.

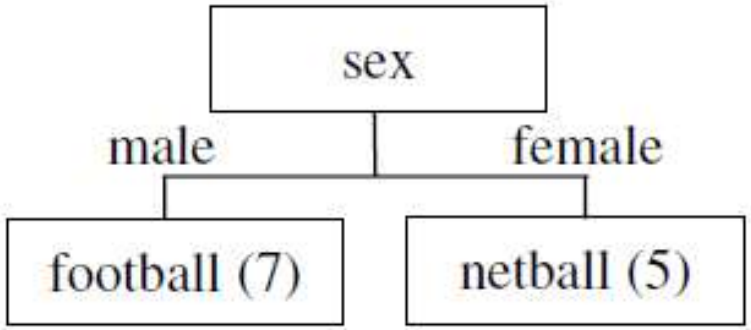
eyecolour	married	sex	hairlength	class
brown	yes	male	long	football
blue	yes	male	short	football
brown	yes	male	long	football
brown	no	female	long	netball
brown	no	female	long	netball
blue	no	male	long	football
brown	no	female	long	netball
brown	no	male	short	football
brown	yes	female	short	netball
brown	no	female	long	netball
blue	no	male	long	football
blue	no	male	short	football

- All the blue-eyed students play football.



- For the brown-eyed students, the critical factor is whether or not they are married.
  - If they are, then the long-haired ones all play football and the short-haired ones all play netball.
  - If they are not married, it is the other way round: the short-haired ones play football and the long-haired ones play netball.

Another Decision Tree



- Both decision trees are compatible with the data from which they were generated.
- Only way to know which one gives better results for unseen data is to use them both and compare the results.

Anonymous Data Set

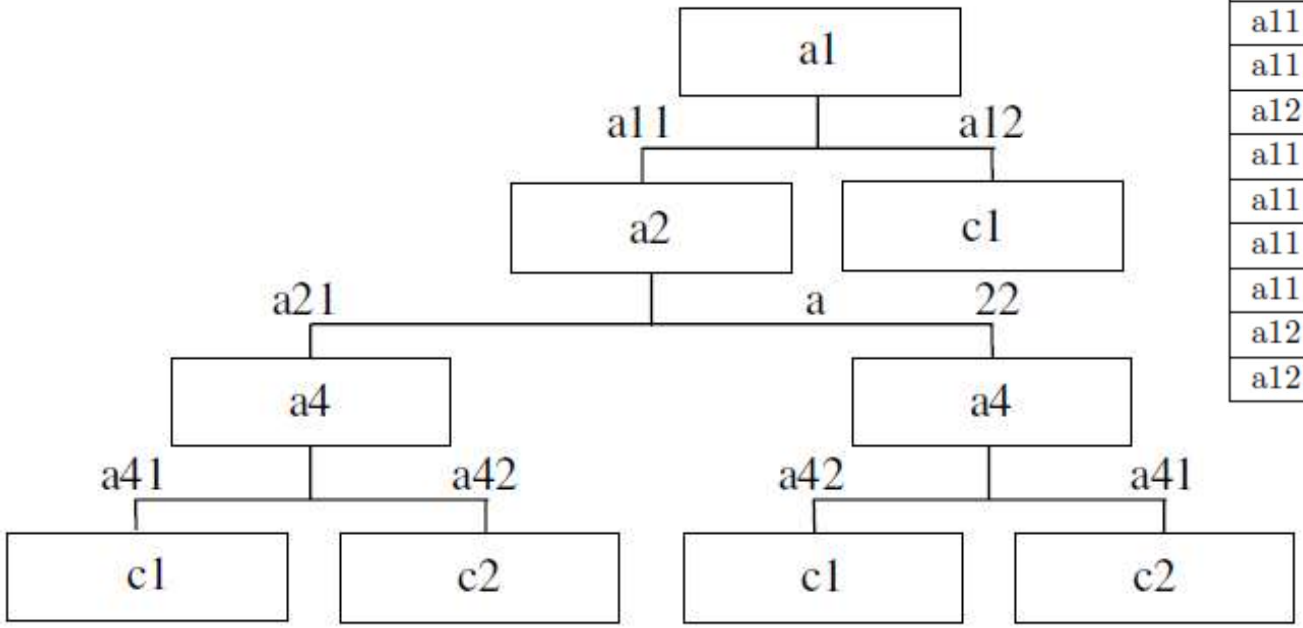
- 12 instances.
- Four attributes, **a1**, **a2**, **a3** and **a4**, with values **a11**, **a12**

  etc.,
- Two classes **c1**

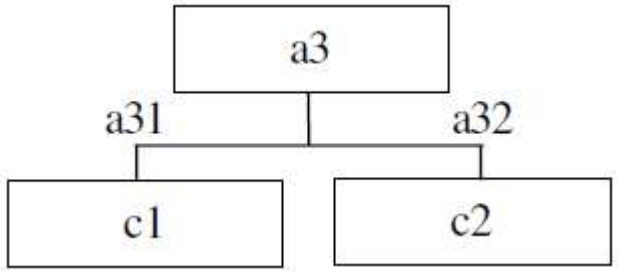
  and **c2**.

a1	a2	a3	a4	class
a11	a21	a31	a41	c1
a12	a21	a31	a42	c1
a11	a21	a31	a41	c1
a11	a22	a32	a41	c2
a11	a22	a32	a41	c2
a12	a22	a31	a41	c1
a11	a22	a32	a41	c2
a11	a22	a31	a42	c1
a11	a21	a32	a42	c2
a11	a22	a32	a41	c2
a12	a22	a31	a41	c1
a12	a22	a31	a42	c1

a1	a2	a3	a4	class
a11	a21	a31	a41	c1
a12	a21	a31	a42	c1
a11	a21	a31	a41	c1
a11	a22	a32	a41	c2
a11	a22	a32	a41	c2
a12	a22	a31	a41	c1
a11	a22	a32	a41	c2
a11	a22	a31	a42	c1
a11	a21	a32	a42	c2
a11	a22	a32	a41	c2
a12	a22	a31	a41	c1
a12	a22	a31	a42	c1



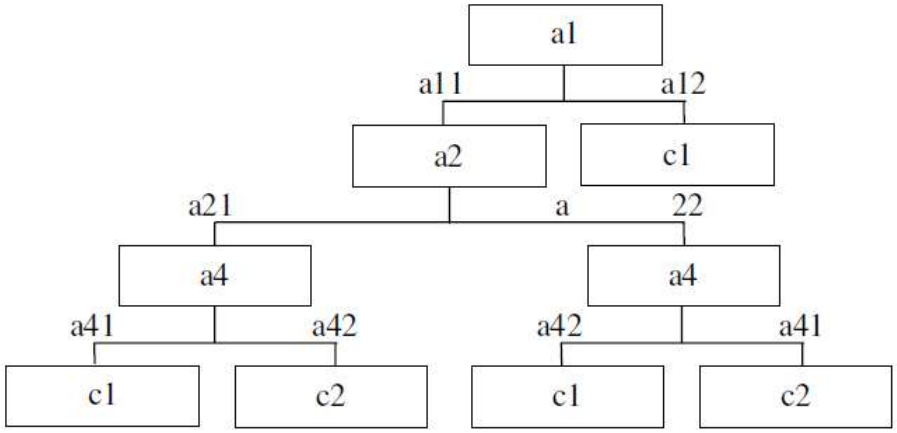
- Decision tree gotten by considering attributes in left-to-right order

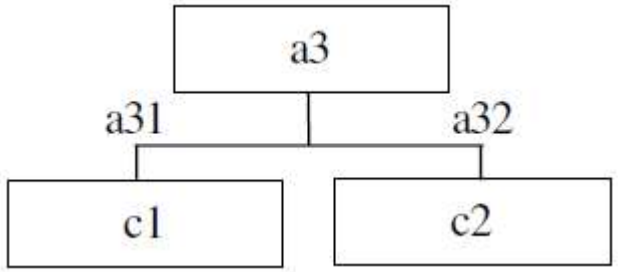


- Also supported by the data

a1	a2	a3	a4	class
a11	a21	a31	a41	c1
a12	a21	a31	a42	c1
a11	a21	a31	a41	c1
a11	a22	a32	a41	c2
a11	a22	a32	a41	c2
a12	a22	a31	a41	c1
a11	a22	a32	a41	c2
a11	a22	a31	a42	c1
a11	a21	a32	a42	c2
a11	a22	a32	a41	c2
a12	a22	a31	a41	c1
a12	a22	a31	a42	c1

Choosing Amongst Possible Trees





- Rightward tree seems preferable because simpler.
- Data mining algorithms generally do not allow the use of any background knowledge.
- What prevents meaningless decision rules from being generated?
- Our next current concern.

Entropy, Information Gain - Motivation

Dataset	excluding entropy		entropy
	most	least	
contact lenses	42	26	**16**
lens24	21	**9**	**9**
chess	155	52	**20**
vote	116	40	**34**
monk1	89	53	**52**
monk2	142	109	**95**
monk3	77	43	**28**

- Different choice of attribute for splitting results in different branching
- Entropy minimizes the complexity, number of branches, in the decision tree.
- No guarantee that using entropy will always lead to a small decision Tree
  - Experience shows that it generally produces trees with fewer branches than other attribute selection criteria.
- Experience shows also that small trees tend to give more accurate predictions than large one

Lens Dataset

Value of attribute				Class
age	specRx	astig	tears	
1	1	1	1	3
1	1	1	2	2
1	1	2	1	3
1	1	2	2	1
1	2	1	1	3
1	2	1	2	2
1	2	2	1	3
1	2	2	2	1
2	1	1	1	3
2	1	1	2	2
2	1	2	1	3
2	1	2	2	1
2	2	1	1	3
2	2	1	2	2
2	2	2	1	3
2	2	2	2	3
3	1	1	1	3
3	1	1	2	3
3	1	2	1	3
3	1	2	2	1
3	2	1	1	3
3	2	1	2	2
3	2	2	1	3
3	2	2	2	3

classes
1: hard contact lenses
2: soft contact lenses
3: no contact lenses

age
1: young
2: pre-presbyopic
3: presbyopic

specRx
(spectacle prescription)
1: myopia
2: high hypermetropia

astig
(whether astigmatic)
1: no
2: yes

tears
(tear production rate)
1: reduced
2: normal

Entropy Defined

- Entropy is a measure of presence of more than one possible classification.
- If there are K classes, we can denote the proportion of instances with classification i by p_i for $i = 1$ to K.
- The value of p_i is the number of instances of class *i* divided by the total number of instances,
  - A number between 0 and 1 inclusive.
- The entropy of the training set defined by the formula

$$E = -\sum_{i=1}^{K} p_i \log_2 p_i$$

- (summed over the non-empty classes only, i.e. classes for which $p_i \neq 0$.)

Entropy Properties

- Value of $-p_i \log_2 p_i$ is positive for values of p_i
- Greater than zero and less than 1.
- When $p_i = 1$ the value of $-p_i \log_2 p_i$ is zero.
- Thus, *E* is positive or zero for all training sets.
- *E* has minimum value (zero) if and only if all the instances have the same classification
- Entropy takes its maximum value when the instances are equally distributed
- Attribute values equally distributed
- Each $p_i = \frac{1}{K}$, so

$$-\sum_{i=1}^{K}(\frac{1}{K})log_2\,(\frac{1}{K})$$

$$= -K(\frac{1}{K})log_2(\frac{1}{K})$$

$$-log_2\left(\frac{1}{K}\right) = log_2(\frac{1}{K})$$

- If there are 2, 3 or 4 classes this maximum value is 1, 1.5850, 2.

Entropy for Lens Dataset

- For the initial *lens24* training set
  - 24 instances
  - 3 classes.
  - There are 4 instances with classification 1
  - 5 instances with classification 2
  - 15 instances with classification 3
  - So $p_1 = 4/24, p_2 = 5/24 \; and \; p_3 = 15/24.$
- We will calculate the entropy of the starting node.
- $-(\frac{4}{24}) \log_2(\frac{4}{24})$
- $-(\frac{5}{24}) \log_2(\frac{5}{24})$
- $-(\frac{15}{24}) \log_2(\frac{15}{24})$
- $= 0.4308 + 0.4715 + 0.4238 = 1.3261$

Class
3
2
3
1
3
2
3
1
3
2
3
1
3
2
3
3
3
3
3
3
1
3
2
3
3

Using Entropy for Attribute Selection

- Using Lens training set and splitting on Age, three subsets E_1, E_2, E_3.

Value of attribute				Class
age	specRx	astig	tears	
1	1	1	1	3
1	1	1	2	2
1	1	2	1	3
1	1	2	2	1
1	2	1	1	3
1	2	1	2	2
1	2	2	1	3
1	2	2	2	1

Value of attribute				Class
age	specRx	astig	tears	
2	1	1	1	3
2	1	1	2	2
2	1	2	1	3
2	1	2	2	1
2	2	1	1	3
2	2	1	2	2
2	2	2	1	3
2	2	2	2	3

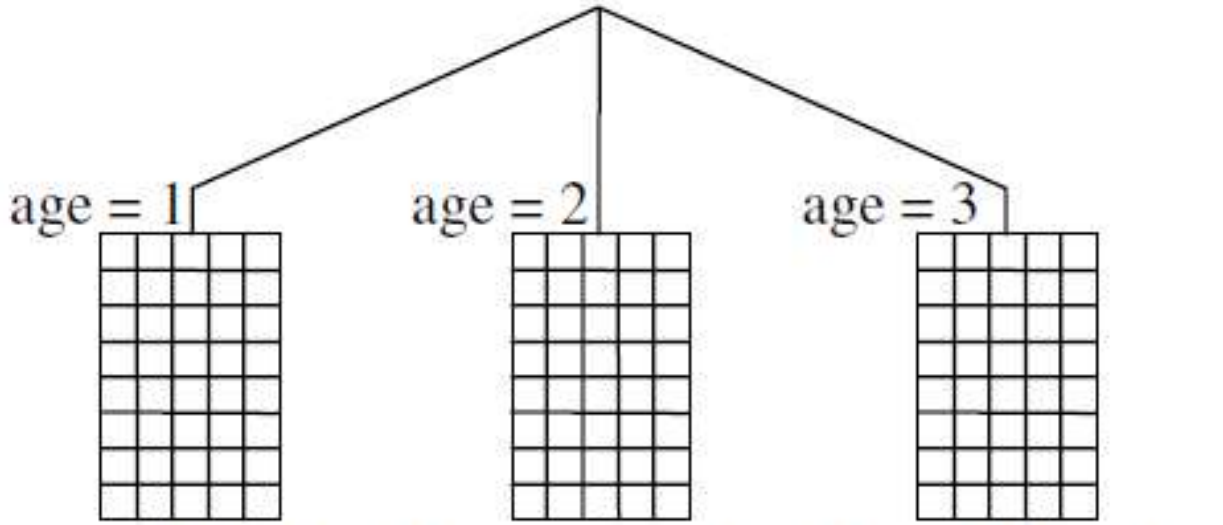
Value of attribute				Class
age	specRx	astig	tears	
3	1	1	1	3
3	1	1	2	3
3	1	2	1	3
3	1	2	2	1
3	2	1	1	3
3	2	1	2	2
3	2	2	1	3
3	2	2	2	3

- **Training set 2 (age = 1)**
- Entropy $E_1 = -(2/8)\log_2(2/8) - (2/8)\log_2(2/8) - (4/8)\log_2(4/8) = 0.5 + 0.5 + 0.5 = 1.5$
- **Training set 2 (age = 2)**
- Entropy $E_2 = -(1/8)\log_2(1/8) - (2/8)\log_2(2/8) - (5/8)\log_2(5/8) = 0.375 + 0.5 + 0.4238 = 1.2988$
- **Training Set 3 (age = 3)**
- Entropy $E_3 = -(1/8)\log_2(1/8) - (1/8)\log_2(1/8) - (6/8)\log_2(6/8) = 0.375 + 0.375 + 0.3113 = 1.0613$

Using Entropy for Attribute Selection

- Average entropy of the three training sets produced by splitting on attribute Age, denoted by E_{new}
- Here, $E_{new} = (\frac{8}{24})E_1 + (\frac{8}{24})E_2 + (\frac{8}{24})E_3 = 1.2867$ bits.
- Define Information Gain = $E_{start} - E_{new}$
- Here, the *information gain* from splitting on attribute *age* is 1.3261 − 1.2867 = 0.0394 bits
- The 'entropy method' of attribute selection: choose to split on the attribute that gives the greatest reduction in (average) entropy
  - That maximizes the value of Information Gain.
  - Equivalent to minimizing the value of E_{new} as E_{start} is fixed.

Maximizing Gain



Initial Entropy = 1.3261 bits
Average Entropy of Subsets = 1.2867 bits
Information Gain = 1.3261–1.2867 = 0.0394 bits

Maximizing Gain

- The values of *Enew* and Information Gain for splitting on each of the four attributes *age*, *specRx*, *astig* and *tears:*
- **Attribute *age***
  - E_{new} = 1.2867
  - Information Gain = 1.3261 − 1.2867 = 0.0394 bits
- **Attribute *specRx***
  - E_{new} = 1.2866
  - Information Gain = 1.3261 − 1.2866 = 0.0395 bits
- **Attribute *astig***
  - E_{new} = 0.9491
  - Information Gain = 1.3261 − 0.9491 = 0.3770 bits
- **Attribute *tears***
  - E_{new} = 0.7773
  - Information Gain = 1.3261 − 0.7773 = 0.5488 bits

Maximizes Gain

SoftEng	ARIN	HCI	CSA	Project	Class
A	B	A	B	B	SECOND
A	B	B	B	A	FIRST
A	A	A	B	B	SECOND
B	A	A	B	B	SECOND
A	A	B	B	A	FIRST
B	A	A	B	B	SECOND
A	B	B	B	B	SECOND
A	B	B	B	B	SECOND
A	A	A	A	A	FIRST
B	A	A	B	B	SECOND
B	A	A	B	B	SECOND
A	B	B	A	B	SECOND
B	B	B	B	A	SECOND
A	A	B	A	B	FIRST
B	B	B	B	A	SECOND
A	A	B	B	B	SECOND
B	B	B	B	B	SECOND
A	A	B	A	A	FIRST
B	B	B	A	A	SECOND
B	B	A	A	B	SECOND
B	B	B	B	A	SECOND
B	A	B	A	B	SECOND
A	B	B	B	A	FIRST
A	B	A	B	B	SECOND
B	A	B	B	B	SECOND
A	B	B	B	B	SECOND

Classes
FIRST, SECOND
SoftEng
A,B
ARIN
A,B
HCI
A,B
CSA
A,B
Project
A,B