Artificial Intelligence

Decision Trees

Nilsson - briefly mentioned in Chapter 3

(Russell and Norvig - Chapter 18)

---

credit card application (1)

information from your application:

- how long have you lived at your current address?
- what is your salary?
- do you have a savings account?
- how old are you?

information from the credit bureau:

- have you ever defaulted on a loan? when?
- how many other credit cards do you have already?
- have you ever declared bankruptcy?

---

credit card application (2)

information from your application:

F1: how long have you lived at your current address?	5 years
F2: what is your salary?	$25,000
F3: do you have a savings account?	no
F4: how old are you?	28

information from the credit bureau:

F5: have you ever defaulted on a loan? when?	yes, 8 years ago
F6: how many other credit cards do you already have?	none
F7: have you ever declared bankruptcy?	no

Do we issue a credit card to this person?
If yes, a regular or a gold card?

classification problem

---

credit card application (3)
inductive learning problem

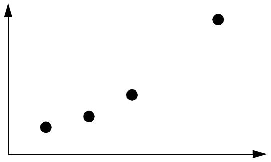
features					class
F1	F2	F3	F4	...	was it a good idea to issue a credit card to this person?
5	$100,000	yes	52	...	yes
3	$50,000	yes	40	...	no
2	$12,000	no	20	...	yes
9	$60,000	yes	31	...	yes
5	$25,000	no	28	...	???
a	b	c	d		f(a,b,c,d)

training examples

find a function that is consistent with all training examples
and that we believe will make the fewest mistakes
on the examples with unknown classes
note: we need bias!

credit card application (4)

bias

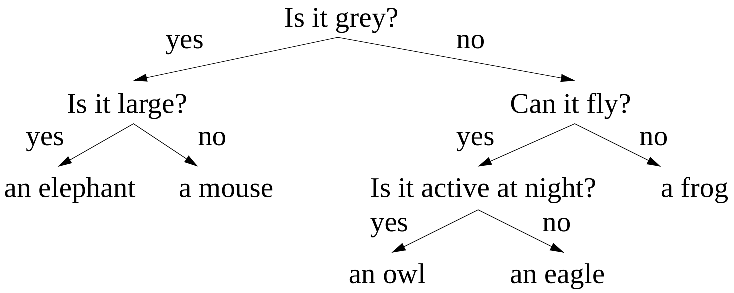


---

decision trees (1)

= one particular inductive learning method

learning takes time, classification is fast



Is it grey?
yes
no
Is it large?
yes
no
an elephant
a mouse
Can it fly?
yes
no
Is it active at night?
yes
no
an owl
an eagle
a frog

in the following: binary values only
(not multi-valued, not continuous, not missing)

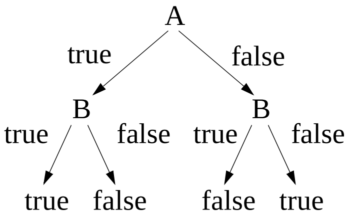
---

decision trees (2)

every prop formula corresponds to a decision tree with binary variable
every decision tree with binary variables corresponds to a prop formula

convert into disjunctive normal form

(A AND B) OR (NOT A AND NOT B)



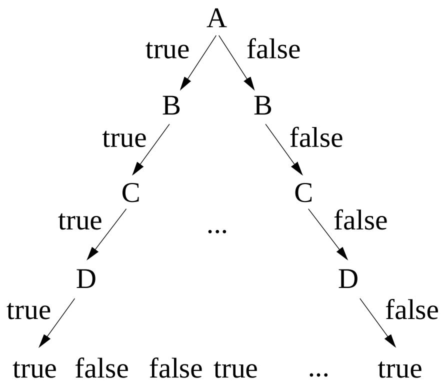
A
true
false
B
true
false
B
true
false
true
false
false
true

---

decision trees (3)

decision trees are often much more compact than tables

however, some simple concepts have large decision trees
for example: even parity



A
true
false
B
B
true
false
C
...
C
true
false
D
D
true
false
true
false
false
true
...
true

inductive bias of decision trees

small is beautiful (number of tests, depth)

Occams Razor

there are fewer small decision trees than large ones
thus, there is only a small chance
that ANY small decision tree
that is completely incorrect
will be consistent with all training examples

problem:
finding the smallest decision tree
that is consistent will all training examples
is NP hard

---

example

put the most discriminating feature first (= at the root)

	F1	F2	F3	F4	class
E1:	true	true	false	true	true
E2:	true	false	false	false	true
E3:	true	true	true	true	false
E4:	true	true	true	false	false

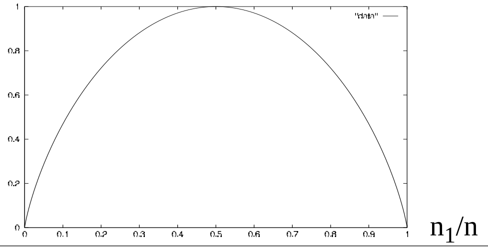
---

entropy

assume that there are n examples total
n_i examples have class i

$$-\Sigma_i (n_i/n) \log_2 (n_i/n)$$

assume two classes

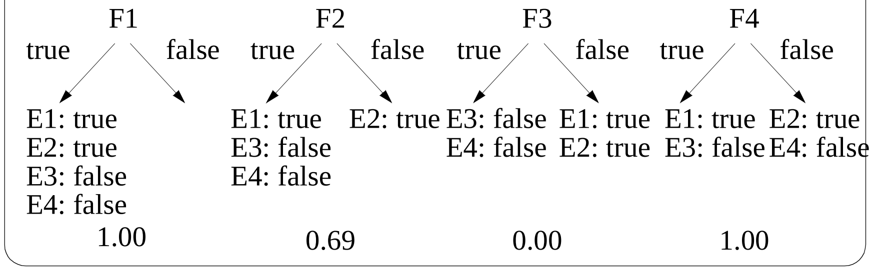


---

example (1)

put the feature first that results in the smallest average entropy

	F1	F2	F3	F4	class
E1:	true	true	false	true	true
E2:	true	false	false	false	true
E3:	true	true	true	true	false
E4:	true	true	true	false	false



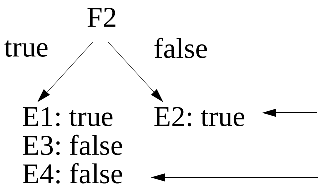
F1
true: E1: true, E2: true, E3: false, E4: false
false:
1.00

F2
true: E1: true, E3: false, E4: false
false: E2: true
0.69

F3
true: E3: false, E4: false
false: E1: true, E2: true
0.00

F4
true: E1: true, E3: false
false: E2: true, E4: false
1.00

example (2)



F2

true

false

E1: true
E3: false
E4: false

E2: true

why do we want to minimize these entropies?
because they are a measure of the remaining work and this work must be small for the tree to be small

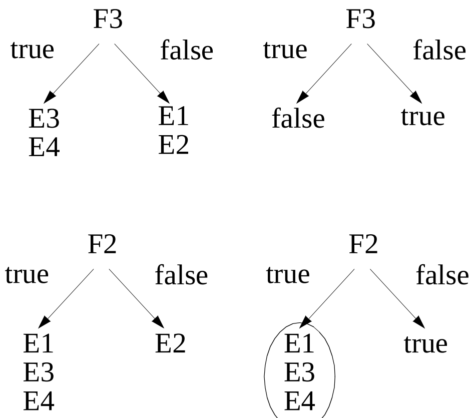
left branch: 3 out of 4 examples
entropy is 0.9182

right branch: 1 out of 4 examples
entropy is 0.0000

average entropy is 3/4 0.9182 + 1/4 0.0000 = 0.6887

---

example (3)



F3

true

false

E3
E4

E1
E2

F3

true

false

false

true

F2

true

false

E1
E3
E4

E2

F2

true

false

E1
E3
E4

true

continue to split
no examples?
no attributes left for splitting?
probabilities or majority?