AI Course Fall HT 2/2015:
Exercise 4: CBR, Decision Tree Learning, Bayes’ Networks

Task 16

Image, somebody gave you the task is to construct a recommender system for books . What are the cases? What are solutions to the cases? What information you would use to describe a case? How a similarity function could look like?

First we need to define a case language: what characterizes a book and how can we compare those features?

Feature	Domain	Comparison
Author	Names, Strings	Pure quality, or a author concept tree with information about the author – age, sex, …
Title	Sequence of Strings	Similar sequences of strings
No of pages	Number	
Category	Concept Tree	In the same subtree… (Children Book, Youth book, Fantasy, Horror…)
Age group	Interval of numbers	
Price	Number	
Publisher	String	Similar to author

Similarity function could be a weighted sum of local similarities; weights could be adapted to the individual user (or with some social mechanism)

Task 17 Decision Tree Learning

Use ID3 to generate a Decision Tree from the following set of examples to generate knowledge classification. This is an example of concept learning, in which the description of a particular concept (here a rules that describe classes of animals)

Name	Hair	Feathers	Eggs	Milk	backbone	fins	legs	tails	Class
frog	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	4	FALSE	amphibian
newt	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	4	TRUE	amphibian
catfish	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	0	TRUE	fish
herring	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	0	TRUE	fish
piranha	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	0	TRUE	fish
tuna	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	0	TRUE	fish
honeybee	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	6	FALSE	insect
ladybird	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	6	FALSE	insect
bear	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	4	FALSE	mammal
pony	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	4	TRUE	mammal
porpoise	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	0	TRUE	mammal
reindeer	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	4	TRUE	mammal
seal	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	0	FALSE	mammal
sealion	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	2	TRUE	mammal
squirrel	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	2	TRUE	mammal
vampire	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	2	TRUE	mammal
pitviper	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	0	TRUE	reptile
seasnake	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0	TRUE	reptile
tortoise	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	4	TRUE	reptile
tuatara	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	4	TRUE	reptile

How can we describe a mammal?

First of all, we decide to ignore the name of the species – ID3 would otherwise select it and we have a tree with 20 branches all leading to a particular species – this does not abstract and this is nothing that we want.

→ we use the examples to create a decision tree about deciding which class an animal belongs to.

1) Start with the full table. Which is the attributed giving the highest information gain (without name)

$$H(class) = \sum_{c \in \{amphi,fish,insect,mammal,reptile\}} - p_c \log_2 p_c$$

$$H(class) = -\frac{2}{20}\log_2\frac{2}{20} - \frac{4}{20}\log_2\frac{4}{20} - \frac{2}{20}\log_2\frac{2}{20} - \frac{8}{20}\log_2\frac{8}{20} - \frac{4}{20}\log_2\frac{4}{20} = 2{,}121928$$

For each attribute in {Hair, Feathers, Eggs, Milk, Backbone, Fins, Legs, Tails} we calculate the conditional entropy – that means the entropy that is left when sorting according to the attribute:

Hair:

$$H(class|Hair) = \frac{hair = true}{all} H(S|hair = true) + \frac{hair = false}{all} H(S|hair = false)$$

So, H(S|Hair=false): There are 12 cases with Hair=false

	Class value	How many?	px	-px*log(px)
	amphi	2	0,166667	0,430827
	fish	4	0,333333	0,528321
	insects	1	0,083333	0,298747
	mammal	1	0,083333	0,298747
	reptile	4	0,333333	0,528321

H(S|hair=false) 2,084963

H(S|Hair=true): There are 8 cases with Hair = true

	Class value	How many?	px	-px*log(px)
	amphi	0	0	0
	fish	0	0	0
	insects	1	0,125	0,375
	mammal	7	0,875	0,168564
	reptile	0	0	0

H(S|hair=true) 0,543564

$$H(class|Hair) = \frac{8}{20}0{,}543564 + \frac{12}{20}2{,}084963 = 1{,}468403$$

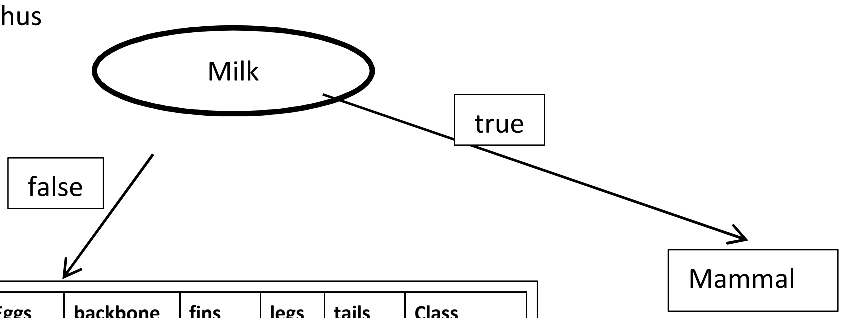
→ Gain (Hair) = H(class) – H(class|hair) = 0,653525

Bei Feathers – alle Werte = false → Gain(Feathers) = 0

We do all these calculations for all relevant attributes:

Hair	0,653525
Feathers	0
Eggs	0,830519
Milk	0,970951
Backbone	0,468996
Fins	0,575061
Legs	0,909128
Tails	0,386767

So, Milk is the best attribute and thus our root of the tree:

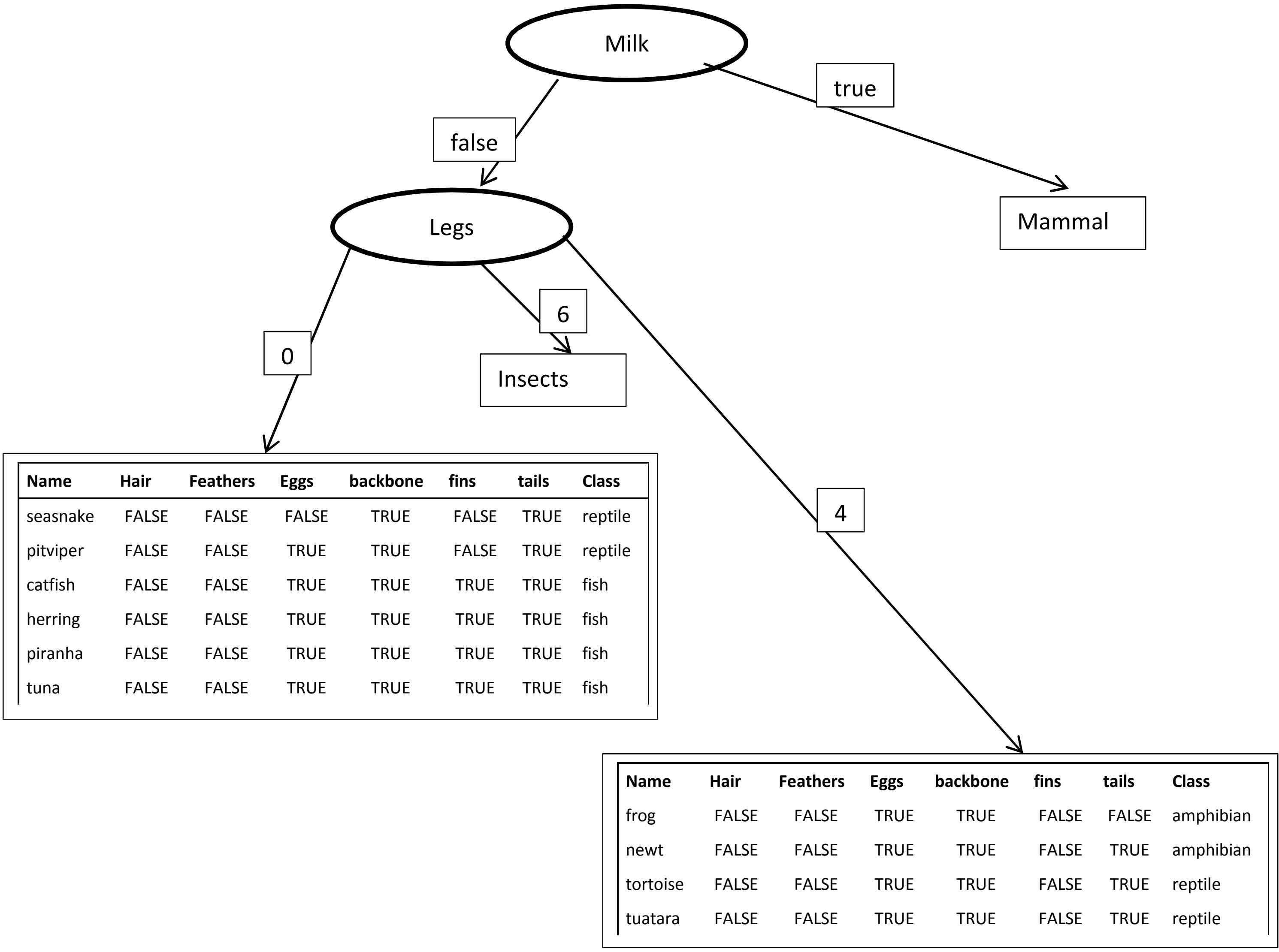


Name	Hair	Feathers	Eggs	backbone	fins	legs	tails	Class
seasnake	FALSE	FALSE	FALSE	TRUE	FALSE	0	TRUE	reptile
frog	FALSE	FALSE	TRUE	TRUE	FALSE	4	FALSE	amphibian
newt	FALSE	FALSE	TRUE	TRUE	FALSE	4	TRUE	amphibian
catfish	FALSE	FALSE	TRUE	TRUE	TRUE	0	TRUE	fish
herring	FALSE	FALSE	TRUE	TRUE	TRUE	0	TRUE	fish
piranha	FALSE	FALSE	TRUE	TRUE	TRUE	0	TRUE	fish
tuna	FALSE	FALSE	TRUE	TRUE	TRUE	0	TRUE	fish
honeybee	TRUE	FALSE	TRUE	FALSE	FALSE	6	FALSE	insect
ladybird	FALSE	FALSE	TRUE	FALSE	FALSE	6	FALSE	insect
pitviper	FALSE	FALSE	TRUE	TRUE	FALSE	0	TRUE	reptile
tortoise	FALSE	FALSE	TRUE	TRUE	FALSE	4	TRUE	reptile
tuatara	FALSE	FALSE	TRUE	TRUE	FALSE	4	TRUE	reptile

This divides the overall table into two parts: a table in which we have all cases with Milk=false → this table has mixed classes, so see next steps and a table in which there are all classes with milk=true. That is a homogeneous group with just mammals → that is the answer to our question – mammals are characterized by milk-feeding of offsprings. But what about the others? → lets continue with the table of all cases in which Milk is false

With the same formulas as above, just without mammals: H(Class) = 1,918296

Hair	0,24715
Feathers	0
Eggs	0,143391
Backbone	0,650022
Fins	0,918296
Legs	1,125815
Tails	0,644611



Name	Hair	Feathers	Eggs	backbone	fins	tails	Class
seasnake	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	reptile
pitviper	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	reptile
catfish	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	fish
herring	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	fish
piranha	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	fish
tuna	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	fish

Name	Hair	Feathers	Eggs	backbone	fins	tails	Class
frog	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	amphibian
newt	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	amphibian
tortoise	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	reptile
tuatara	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	reptile

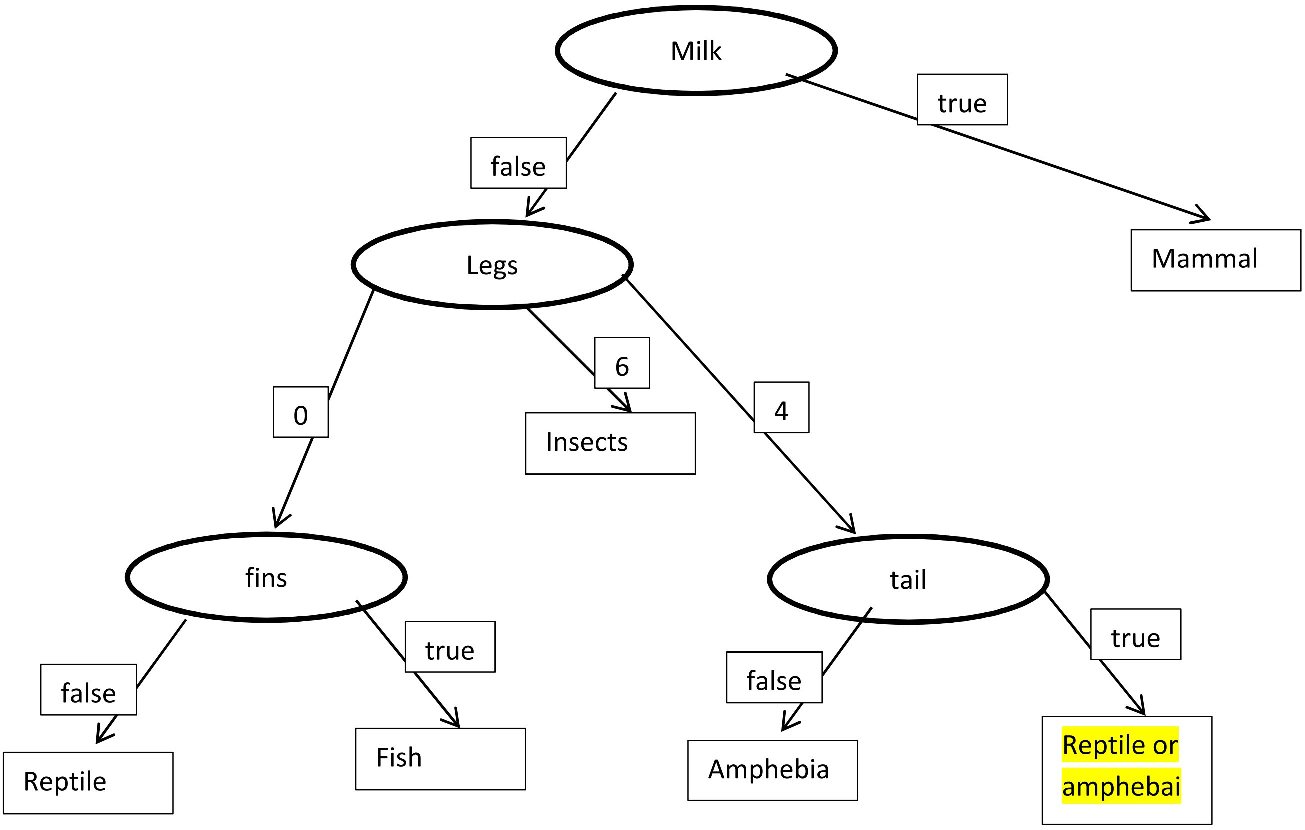
ID3 is a “greedy” procedure! It takes the local optimum without considering future steps. If we would have used the attribute fins, the fishes would be ruled out and the rest could be divided in one step checking backbone. But as we choose the legs (→ id3 biasses attributes with more values!)

The next level for the left partial tree:

Name	Hair	Feathers	Eggs	backbone	fins	tails	Class
Seasnake	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	reptile
Pitviper	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	reptile
Catfish	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	fish
Herring	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	fish
Piranha	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	fish
Tuna	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	fish

H(class) = 0,918296

But, there are just two attributes left: egg and fins, the rest has an gain=0. Fins clearly is better.

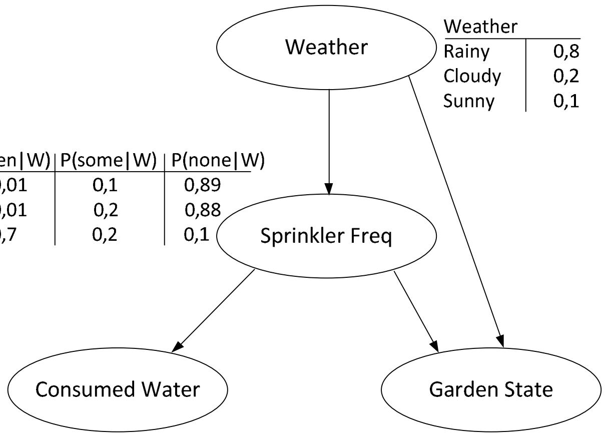


Task 18 Bayesian Networks

You have been on vacation for some weeks and your neighbor took care of your garden. Construct a Bayesian Network using the following variables, inventing some reasonable conditional probability table values.

- Current state of your garden
- Amount of consumed water
- Frequency of usage of the sprinkler
- Weather conditions during your vacation

Formulate queries for your Bayesian Network with relation to the general weather conditions as conditional probabilities that could be calculated. What you might want to know being on vacation and thinking about your garden?



Weather	
Rainy	0,8
Cloudy	0,2
Sunny	0,1

W	P(often|W)	P(some|W)	P(none|W)
Rainy	0,01	0,1	0,89
Cloudy	0,01	0,2	0,88
Sunny	0,7	0,2	0,1

W	SpFreq	P(good|W,SpFrq)	P(bad|W,SpFreq)
Rainy	often	0,2	0,8
Rainy	some	0,5	0,5
Rainy	none	0,9	0,1
Cloudy	often	0,4	0,6
Cloudy	some	0,8	0,2
Cloudy	none	0,7	0,3
Sunny	often	0,9	0,1
Sunny	some	0,5	0,5
Sunny	none	0,2	0,8

SpFreq	P(huge|W)	P(some|W)	P(none|W)
often	0,7	0,3	0
some	0,2	0,8	0
none	0	0,05	0,95

What might we want to know:

P(Garden State = good)

P(Consumed Water = huge)

Diagnostic
P(Springer Freq = high |Garden State = bad) -

P(Weather = Sunny | consumed water = huge, Garden State = bad)

Causal:
P(Garden State = good | Sprinker Freq = often, weather = rainy)