

## Tutorial 8:

# Current Best Learning, Decision Trees, Naive Bayes

### Question 1

Note:

“<nil> <=> Yes” is the most specific hypothesis possible, because it matches no patterns at all

a) Size of the hypothesis space.

We are only considering single hypotheses like the following:

<\*, Warm, High, Strong, Cool, Change> <=> Yes

So the number of possible hypotheses of this type

= # possible Sky values (Sunny, Cloudy, Rain, \*) \* same for each other attribute or output.

=  $4 \cdot 3^5 = 972$ .

Remember: each has to match Yes on the output (here: EnjoySport).

There is also the <nil> <=> Yes hypothesis, which always produces a No (because no test pattern can match <nil>).

So the total number is 973.

With a little thought, we can see that these single hypotheses are by no means the only possible hypotheses on these kinds of space.

E.g.

(<Sunny, Warm, High, Strong, Cool, Change> or

<Cloudy, Warm, High, Strong, Cool, Change>) <=> Yes

has not been considered. But we don't consider this further in COMP3702/7702.

b) Ranking the hypotheses by specificity from most specific to least.

Here, specificity is based on the number of possible patterns which will match the hypothesis.

# of matching patterns:

A= < \*, \*, High, \*, \*, \* > <=> Yes.  $\rightarrow 3 \cdot 2^4 = 48$

B= < \*, \*, Normal, Weak, \*, Change > <=> Yes.  $\rightarrow 3 \cdot 2 \cdot 2 = 12$

C= < Rain, Cold, High, Strong, Warm, Change > <=> Yes.  $\rightarrow 1$

D= < nil > <=> Yes.  $\rightarrow 0$

E= < \*, Cold, Normal, Strong, Cool, Same > <=> Yes.  $\rightarrow 3$

So: ranked hypotheses: D C E B A

c) Current best hypothesis learning:

H1 = <\*, Cold, High, \*, \*, \*> <=> Yes  
 Example 2: false negative -> generalise  
 H2 = <\*, \*, High, \*, \*, \*> <=> Yes  
 Example 3: false positive -> specialise  
 H3 = <Sunny, \*, High, \*, \*, \*> <=> Yes  
 Example 4: classified correctly  
 H4 = no change (H3 again).

## Question 2

Notes:

- When calculating the information gain of an attribute, you need to figure out how much information the data set as a whole contains, and then you figure out the amount of information that is left *after* classifying based on that attribute. In other words, the remainder is the quantity of information that that attribute failed to get rid of.
- When choosing an attribute, we want the one with largest gain. The first term in the gain equation is the same for all the attributes, so we actually want the attribute with the smallest remainder.
- At every node in the tree, ID3 makes a decision to divide based on the maximum quantity of entropy (information) it can get rid of in its current data set.
- Whenever you make a decision at a node in ID3, you're reducing the size of the dataset that you're considering (for any further branching).

a) Finding the attribute with greatest information gain (ie: smallest remainder).  
 There are 4 options: Height, Weight, Age, Gender.

Remainder(A) = sum (i over all possible values of attribute A)  $(p_{i+n_i}) / (p+n)$   
 $I((p_i / (p_{i+n_i}), (n_i / (p_{i+n_i})))$

Worth remembering:  $I(0,1) = 0 = I(1,0)$  and that  $I(0.5,0.5) = 1$ .

Height:

$p_{Tall} = 10$  ;  $n_{Tall} = 0$   
 $p_{Medium} = 1$  ;  $n_{Medium} = 4$   
 $p_{Short} = 0$  ;  $n_{Short} = 5$

Remainder(Height) =  $10/20 I(10/10,0/10) + 5/20 I(1/5,4/5) + 5/20 I(0/5,5/5)$   
 $= 0 + .25 * (-(1/5) * \log_2(1/5) - (4/5) * \log_2(4/5)) + 0$   
 $= 0.180$

Weight:

$p_{Fat} = 3$  ;  $n_{Fat} = 3$   
 $p_{Medium} = 1$  ;  $n_{Medium} = 4$   
 $p_{Thin} = 7$  ;  $n_{Thin} = 2$

Remainder(Weight) =  $6/20 I(3/6,3/6) + 5/20 I(1/5,4/5) + 9/20 I(7/9,2/9)$   
 $= (6/20) * 1 + 0.18 + (9/20) * (-(7/9) * \log_2(7/9) - (2/9) * \log_2(2/9))$   
 $= 0.824$

Age:

$$p\_Young = 7 ; n\_Young = 3$$

$$p\_Middleage = 2 ; n\_Middleage = 2$$

$$p\_Old = 2 ; n\_Old = 4$$

$$\begin{aligned} \text{Remainder(Age)} &= 10/20 I(7/10,3/10) + 4/20 I(2/4,2/4) + 6/20 I(2/6,4/6) \\ &= 0.5*(-(7/10)*\log_2(7/10) - (3/10)*\log_2(3/10)) + 0.2*1 + 0.3*(-(1/3)*\log_2(1/3)- \\ &\quad (2/3)*\log_2(2/3)) \\ &= 0.916 \end{aligned}$$

Gender:

$$p\_Female = 7 ; n\_Female = 4$$

$$p\_Male = 4 ; n\_Male = 5$$

$$\begin{aligned} \text{Remainder(Gender)} &= 11/20 I(7/11,4/11) + 9/20 I(4/9,5/9) \\ &= (11/20)*(-(7/11)*\log_2(7/11) - (4/11)*\log_2(4/11)) + (9/20)*(-(4/9)*\log_2(4/9)- \\ &\quad (5/9)*\log_2(5/9)) \\ &= 0.966 \end{aligned}$$

So Height has the smallest remainder and so the largest information gain and so is the best attribute for the top of the decision tree.

b) Finding the greatest information gain.

$$\text{Gain(A)} = I(p/(p+n),n/(p+n)) - \text{Remainder(A)}.$$

First need to find information content of entire data set

$$\begin{aligned} I(p/(p+n),n/(p+n)) &= I(11/20,9/20) \\ &= - (11/20) * \log_2(11/20) - (9/20) * \log_2(9/20) \\ &= 0.993 \text{ bits per response (Sick = Yes or No).} \end{aligned}$$

Greatest information gain is for height, which had the smallest remainder

$$\text{Gain(Height)} = 0.993 - 0.180 = 0.813$$

c) Creating the rest of the decision tree.

Height=Tall branch:

- all examples have Sick=Yes, so just label that branch Yes.

Height=Medium branch:

#	Weight	Age	Gender	Sick ?
4	Med	Old	F	N
5	Fat	Young	M	N
8	Fat	Young	F	N
15	Thin	Old	F	Y
19	Fat	Old	F	N

Without doing any calculations, do any of these three attributes (Weight, Age, Gender) offer perfect discrimination on this set of examples?

Answer: yes: Weight does. (Note: Age and Gender don't).  
So no need to calculate anything: branch here on Weight.

->

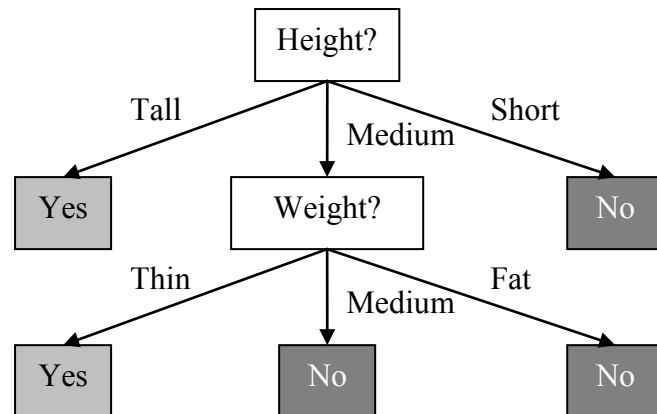
Weight= Med -> Sick = No

Weight= Fat -> Sick = No

Weight= Thin -> Sick = Yes

Height=Short branch

None of these people got sick, so just label that branch No.



d)

This decision tree is possibly not sufficient to warn everyone, as with only 20 examples, all possible combinations of the variables have not been experienced.

Also, there may be problems with dividing age, height, and weight into discrete categories.

### Question 3

For this tutorial, we can just construct the model and its prediction as needed in response to each test pattern. I.e. there is no need to build the full Naive Bayes model for every possible test pattern. However, if we had a lot of data and wanted to use the trained system extensively, it would be worth estimating each of the possible  $P(\text{attribute}=\text{value}|\text{class})$  combinations before using the model on test cases.

The true value of each test case is given in brackets. Remember: we estimate all of the required probabilities from the training data unless given better evidence of their value from somewhere else. We do not use any of the test data to estimate the probabilities.

$P(\text{Sick} = \text{Yes}) = 11/20$ .

$P(\text{Sick} = \text{No}) = 1 - P(\text{Sick} = \text{Yes}) = 9/20$

1<sup>st</sup> Test data: T1 = Tall Medium Young Female (Yes)

Naive Bayes model:

$$\begin{aligned} & P(\text{Sick}=\text{Yes}|\text{T1 (without Sick value !)}) \\ &= P(\text{Sick}=\text{Yes}|\text{Height}=\text{Tall},\text{Weight}=\text{Medium},\text{Age}=\text{Young},\text{Gender}=\text{Female}) \\ &= P(\text{Height}=\text{Tall},\text{Weight}=\text{Medium},\text{Age}=\text{Young},\text{Gender}=\text{Female}|\text{Sick}=\text{Yes}) P(\text{Sick}=\text{Yes})/C \end{aligned}$$

C, above is a constant =

$$P(\text{Height}=\text{Tall},\text{Weight}=\text{Medium},\text{Age}=\text{Young},\text{Gender}=\text{Female}) .$$

We could estimate this from the data using Naive Bayes assumptions (decomposing it into a product of 4 terms for each condition), but there is no need - it will disappear in the renormalisation of the probabilities so that they sum to 1, so we can ignore the constant.

Continuing the equation above using the Naive Bayes model:

$$\begin{aligned} &= P(\text{Height}=\text{Tall},\text{Weight}=\text{Medium},\text{Age}=\text{Young},\text{Gender}=\text{Female}|\text{Sick}=\text{Yes}) P(\text{Sick}=\text{Yes})/C \\ &= P(\text{Height}=\text{Tall}|\text{Sick}=\text{Yes}) P(\text{Weight}=\text{Medium}|\text{Sick}=\text{Yes}) P(\text{Age}=\text{Young}|\text{Sick}=\text{Yes}) \\ & P(\text{Gender}=\text{Female}|\text{Sick}=\text{Yes}) P(\text{Sick}=\text{Yes})/C \end{aligned}$$

All of the above should now be estimated from the training data.

$$= (10/11) * (1/11) * (7/11) * (7/11) * (11/20)/C = 0.018407/C$$

The alternative classification would be Sick=No .

We now find the probability of this under the Naive Bayes model.

$$\begin{aligned} & P(\text{Sick}=\text{No}|\text{Height}=\text{Tall},\text{Weight}=\text{Medium},\text{Age}=\text{Young},\text{Gender}=\text{Female}) \\ &= P(\text{Height}=\text{Tall}|\text{Sick}=\text{No}) P(\text{Weight}=\text{Medium}|\text{Sick}=\text{No}) P(\text{Age}=\text{Young}|\text{Sick}=\text{No}) \\ & P(\text{Gender}=\text{Female}|\text{Sick}=\text{No}) P(\text{Sick}=\text{No})/C \end{aligned}$$

All of the above should now be estimated from the training data.

$$\begin{aligned} &= (0/9) * (?/9) * (?/9) * (?/9) * (9/20)/C \\ &= 0 \end{aligned}$$

No need to work out the ?'s once we see a zero.

From the above, we can see that

$P(\text{Sick}=\text{Yes}|\text{T1})$  is larger than  $P(\text{Sick}=\text{No}|\text{T1})$  , and so the prediction should be Yes.

Renormalising the values so that they sum to 1:

$$P(\text{Sick}=\text{Yes}|\text{T1}) = 1$$

$$P(\text{Sick}=\text{No}|\text{T1}) = 0$$

2<sup>nd</sup> Test data: T2 = Medium Medium Middleage Male (Yes)

Naive Bayes model:

$$P(\text{Sick}=\text{Yes}|\text{T2})$$

$$\begin{aligned}
&= P(\text{Sick}=\text{Yes}|\text{Height}=\text{Medium},\text{Weight}=\text{Medium},\text{Age}=\text{Middleage},\text{Gender}=\text{Male}) \\
&= P(\text{Height}=\text{Medium}|\text{Sick}=\text{Yes}) P(\text{Weight}=\text{Medium}|\text{Sick}=\text{Yes}) \\
&P(\text{Age}=\text{Middleage}|\text{Sick}=\text{Yes}) \\
&P(\text{Gender}=\text{Male}|\text{Sick}=\text{Yes}) P(\text{Sick}=\text{Yes})/C
\end{aligned}$$

All of the above should now be estimated from the training data.  
 $= (1/11) * (1/11) * (2/11) * (4/11) * (11/20)/C = 0.0005464107/C$

The alternative classification would be Sick=No .  
We now find the probability of this under the Naive Bayes model.

$$\begin{aligned}
&P(\text{Sick}=\text{No}|\text{Height}=\text{Medium},\text{Weight}=\text{Medium},\text{Age}=\text{Middleage},\text{Gender}=\text{Male}) \\
&= P(\text{Height}=\text{Medium}|\text{Sick}=\text{No}) P(\text{Weight}=\text{Medium}|\text{Sick}=\text{No}) \\
&P(\text{Age}=\text{Middleage}|\text{Sick}=\text{No}) \\
&P(\text{Gender}=\text{Male}|\text{Sick}=\text{No}) P(\text{Sick}=\text{No})/C
\end{aligned}$$

All of the above should now be estimated from the training data.  
 $= (4/9) * (4/9) * (2/9) * (5/9) * (9/20)/C$   
 $= 0.0109739$

From the above, we can see that  
 $P(\text{Sick}=\text{No}|T2)$  is larger than  $P(\text{Sick}=\text{Yes}|T2)$  , and so the prediction should be No.  
That happens to disagree with the supplied label, but you expect to see some errors on test data.

Renormalising the values so that they sum to 1:

$$\begin{aligned}
P(\text{Sick}=\text{Yes}|T2) &= 0.0005464107/(0.0005464107+0.0109739) = 0.04743021 \\
P(\text{Sick}=\text{No}|T1) &= 0.0109739/(0.0005464107+0.0109739) = 0.95256979
\end{aligned}$$