TP DTB (Decision Trees and Bayes)
Techniques of AI [INFO-H-410]
Correction

v1.0.0

Source files, code templates and corrections related to practical sessions can be found on the UV or on github (`https://github.com/iridia-ulb/INFOH410`).

Representation and Interpretation of Boolean Functions

Symbol	Name
0	FALSE
1	TRUE
$!A$ / $\neg A$	NOT A
$A \wedge B$	A AND B
$A \vee B$	A OR B
$A \oplus B$	A XOR B

Decision Trees

Question 1. Give decision trees to represent the following Boolean functions:

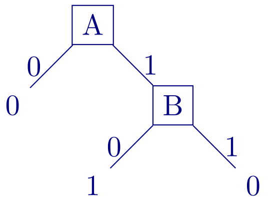
a) $A \wedge \neg B$

b) $A \vee (B \wedge C)$

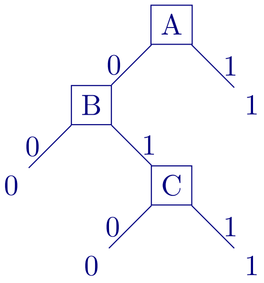
c) $A \oplus B$

d) $(A \wedge B) \vee (C \wedge D)$

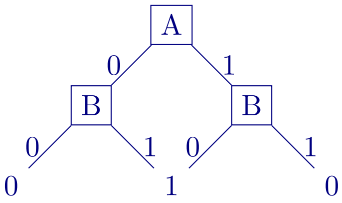
Answer: (a)



(b)



(c)



(d) left to the discretion of the reader.

Growing Decision Trees

Question 2. In order to evaluate the quality of the tree which is grown by ID3, one could compare its performance to a baseline performance. The baseline performance is often the performance of a very simple machine learning algorithm. Consider the following approach:

You have a dataset consisting of 25 examples of each of two classes. You plan to use leave-one-out cross validation. As a baseline, you use a simple majority classifier (a majority classifier is given a set of training data and then always outputs the class that is in the majority in the training set, regardless the input). Such a majority classifier is expected to score about 50%, but with this example of leave-one-out cross-validation, it does not. What will be its performance and why?

Answer: The entire set contains 25 positive examples and 25 negative examples. With leave one out cross validation, you randomly select 49 examples for training and one for testing. Suppose the test instance should be classified positive. In that case the training set contains 24 positive and 25 negative examples. The majority voter hence classifies the instance as negative, which is wrong. A similar reasoning holds for a negative test instance. Hence, the performance of the classifier is always zero in this case.

Question 3. Consider the following data on the hair color, body weight, body height and the usage of lotion of eight different people. The table shows whether the people got sunburned after an afternoon in the sun.

hair	**height**	**weight**	**lotion**	**sunburn?**
blonde	avg	light	False	True
blonde	tall	avg	True	False
brown	short	avg	True	False
blonde	short	avg	False	True
red	avg	heavy	False	True
brown	tall	heavy	False	False
brown	avg	heavy	False	False
blonde	short	light	True	False

a) Using the ID3 algorithm, perform average entropy calculations on the following complete dataset for each of the four attributes. Select the attribute which minimizes the entropy; draw the first level of the decision tree.

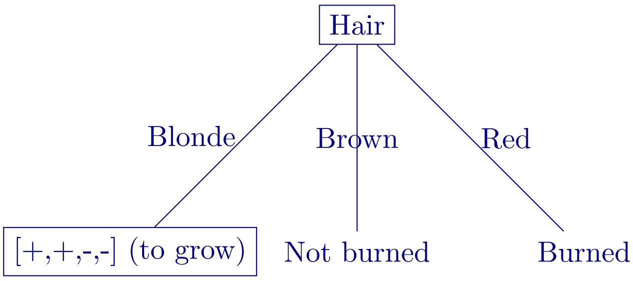
b) Grow the tree until you reach the proper identification of all the samples.

Answer: the entropy H of a set S is given by: $H(s) = \sum_{x \in X} -p(x) log_2(p(x))$
(a) Entropy of the complete dataset is therefore: $H(S) = -\frac{3}{8} log_2(\frac{3}{8}) - \frac{5}{8} log_2(\frac{5}{8}) = 0.954$
then for each feature:

- $H(Hair, blonde)[++, --]$(2 positive and 2 negative samples)$= -\frac{1}{2} log_2(\frac{1}{2}) - \frac{1}{2} log_2(\frac{1}{2})$

- $H(Hair, brown)[- - -] = 0$
- $H(Hair, red)[+] = 0$
- $IG(Hair) = 0.954 - \frac{4}{8} \times 1 - \frac{3}{8} \times 0 - \frac{1}{8} \times 0 = 0.454$
- ...
- $IG(Height) = 0.954 - \frac{3}{8} \times 0.918 - \frac{2}{8} \times 0 - \frac{3}{8} \times 0.918 = 0.266$
- $IG(Weight) = 0.954 - \frac{2}{8} \times 1 - \frac{3}{8} \times 0.918 - \frac{3}{8} \times 0.918 = 0.016$
- $IG(Lotion) = 0.954 - \frac{5}{8} \times 0.971 - \frac{3}{8} \times 0 = 0.347$

The attribute with the biggest gain is ”Hair”, so we build the first layer of the tree with it.



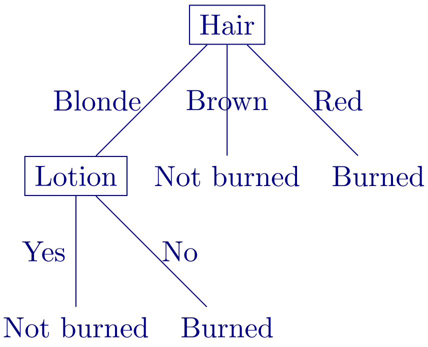
(b) We continue growing the tree, by the computing the entropies (with updated S excluding already classified samples)

$H(S)[++, --] = 1$

For each features left:

- $H(Height, average)[+] = 0$
- $H(Height, tall)[-] = 0$
- $H(Height, short)[+, -] = 1$
- $IG(Height) = 1 - \frac{2}{4} \times 1 = 0.5$
- ...
- $IG(Weight) = 1 - \frac{1}{2} \times 1 - \frac{1}{2} \times 1 = 0$
- $IG(Lotion) = 1 - 0 - 0 = 1$

The attribute with the biggest gain is ”Lotion”, so we build the next layer of the tree with it.



Question 4. We want to train a decision tree using python, however, the library that we use (sklearn[1]) uses the CART algorithm, instead of ID3, and does not allow categorical data.

a) How can you modify the data to mitigate this issue ?

b) Use the sample code (see page 1) to grow a tree and visualize it, using python and sklearn.

Answer: (a) We should use, one hot encoding, ordinal encoding, or label encoding for continuous variables.
(b) see github for implementation.

Question 5. Consider the herewith provided alternative dataset:

hair	**height**	**weight**	**lotion**	**sunburn**
?	avg	light	False	True
blonde	tall	avg	True	False
brown	short	avg	True	False
blonde	short	avg	False	True
red	avg	heavy	False	True
brown	tall	heavy	False	False
brown	avg	?	False	False
blonde	short	light	?	False

Is this dataset suitable for training your model and will you be able to build an optimal decision tree? Explain briefly how you can deal with this situation.

Answer: It is suitable if processed correctly. To handle missing data in a dataset you can: drop the data (if you have a lot), fill the gaps with the most frequent value for the feature, fill the data with the most common value for the same output class, fill the gaps with the mean/median of a given feature (for continuous vars), etc.

Evaluating Hypothesis

Question 6. When testing a hypothesis h, we used a sample set containing 30 samples. We learnt that 3 samples were misclassified. Calculate the 95% and the 50% confidence interval. You observe that the 95% confidence interval is almost 4 times bigger than the 50% confidence interval.
What is the meaning of this interval? Using 95% confidence intervals, is it possible that the true error rate is: 1% instead of 10%? 0% instead of 10% ?

[1] `https://scikit-learn.org/stable/modules/tree.html`

Answer: The following equation is used to estimate the true error given the sample error:

$$error_D(h) = error_s(h) \pm z_n \sqrt{\frac{error_s(h)(1 - error_s(h))}{n}}$$

For 95% confidence interval, $z_n = 1.96$, for 50%, $z_n = 0.67$. Here: $error_s(h) = 3/30 = 0.1$, then:
95% CI:

$$error_D(h) = 0.1 \pm 1.96\sqrt{\frac{0.1(1 - 0.1)}{30}} = [-0.00735, 0.20735]$$

50% CI:

$$error_D(h) = 0.1 \pm 0.67\sqrt{\frac{0.1(1 - 0.1)}{30}} = [0.0634, 0.1366]$$

So, although we experienced a sample error of 0.1, the true error might seem to be with 95% chance in between $[-0.00735, 0.20735]$. A 0.01 is indeed in the interval, it is possible that the true error is actually 1% instead of the 10% we experienced.

However, the true error cannot be 0%, even if 0 is in the interval: As we have experienced that 3 samples were classified wrong, the error is above 0.

Obviously, the true error, cannot be in the range $[-0.00735, 0.20735]$ as an error rate cannot be negative. Moreover, it cannot be 0 as we have evidence that there are at least 3 misclassifications. Hence $(0, 0.20735]$ is a better estimation of the true error.

Question 7. When testing a hypothesis h, we used a sample set containing n samples. We learnt that 10% of the samples were misclassified. We want to be 95% sure (this means "with 95% confidence") that the true error rate is between 5% and 15%. How many samples do we need in order to be able to assure this?

Answer:

$$error_D(h) = 0.1 \pm 1.96\sqrt{\frac{0.1(1 - 0.1)}{n}} = [0.05, 0.15] = 0.1 \pm 0.05$$

$$n = 138.2$$

As the number of samples is a discrete number, we need 139 samples to be 95% confident that the true error is in the given interval.

Bayes theorem and Naïve Bayes classifier

Question 8. For the course X, we experienced that on average 10% of students pass. We also noticed over the last couple of years that from all the students who passed, 90% did attend the exercise sessions. From all the students who did not pass, 95% did not attend the exercise sessions; Are your chances for passing course X increased by attending the exercise sessions?

Answer:

$$P(pass) = 0.1 \rightarrow P(fail) = 0.9$$
$$P(present|pass) = 0.9 \rightarrow P(absent|pass) = 0.1$$
$$P(absent|fail) = 0.95 \rightarrow P(present|fail) = 0.05$$

We want to calculate $P(present)$, we use the theorem of total probability:

$P(present) = P(present|pass)P(pass)+P(present|fail)P(fail) = 0.9{*}0.1+0.05{*}0.9 = 0.135$

We also compute $P(pass|present)$:

$$P(pass|present) = \frac{P(present|pass)P(pass)}{P(present)} = 0.9 * 0.1/0.135 = 0.66$$

So yes, coming to the exercise pays off since $0.66 > 0.1$(overall course success). Also:

$$P(pass|absent) = \frac{P(absent|pass)P(pass)}{P(absent)} = 0.1 * 0.1/0.865 = 0.011$$

Question 9. An HIV test gives a positive result with probability 98% when the patient is indeed affected by HIV, while it gives a negative result with 99% probability when the patient is not affected by HIV. If a patient is drawn at random from a population in which 0.1% of individuals are affected by HIV and he is found positive, what is the probability that he is indeed affected by HIV?

Answer:

$$P(+|HIV) = 0.98 \rightarrow P(-|HIV) = 0.02$$
$$P(-|!HIV) = 0.99 \rightarrow P(+|!HIV) = 0.01$$

We sample from a polulation where $P(HIV) = 0.001$, the patient is found positive. We compute

$P(+) = P(+|HIV){*}P(HIV)+P(+|!HIV){*}P(!HIV) = 0.98{*}0.001+0.01{*}0.999 = 0.01097$

Now we compute $P(HIV|+)$:

$$P(HIV|+) = \frac{P(+|HIV)P(HIV)}{P(+)} = \frac{0.98 * 0.001}{0.01097} = 0.0893$$

Id	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Question 10. At the parking lot of company X, a lot of cars get stolen. See below for an overview of the last 10 cars which were parked. I now park my brand new RED DOMESTIC SUV, what is the maximum a posteriori hypothesis (MAP): will the car be stolen or not according to a naïve bayes classifier?

Answer: An initial idea is to compare the probability that the car will be stolen, given the fact that the car is a red, domestic, SUV with the probability that the car will not be stolen, given the fact that the car is a red, domestic, SUV (RDS):

$$P(stolen|RDS) = \frac{P(RDS|stolen)P(stolen)}{P(RDS)}$$

$$P(!stolen|RDS) = \frac{P(RDS|!stolen)P(!stolen)}{P(RDS)}$$

Naïve Bayes allows us to look at: $H(stolen) = P(RDS|stolen)P(stolen)$, and $H(!stolen) = P(RDS|!stolen)P(!stolen)$ such that we drop $P(RDS)$ as it is a constant independent of our possible classes.
The Naïve Bayes output class will be: $C_{NB} = argmax(H(stolen), H(!stolen))$

$$H(stolen) = P(RDS|stolen)P(stolen) = P(R|stolen)P(D|stolen)P(S|stolen)P(stolen)$$

$$= 3/5 * 2/5 * 1/5 * 1/2 = 6/250 = 0.024$$

$$H(!stolen) = P(RDS|!stolen)P(!stolen) = P(R|!stolen)P(D|!stolen)P(S|!stolen)P(!stolen) = 0.072$$

$H(!stolen) > H(stolen)$, thus it is more likely that the car will not be stolen. Note that by normalizing the obtained results to sum to one, we can calculate the conditional probabilities for each of our classes, given the observed attribute values (red, domestic, SUV). For *stolen* this is: $\frac{0.024}{0.024+0.072} = 0.25$ and $!stolen = 0.75$