

COMP3308/3608 Artificial Intelligence

Week 7 Tutorial exercises Decision trees

Exercise 1. Entropy (Homework)

What is the information content (entropy) of a message telling the outcome of the flip of:

- an honest dice?
- a dice that has been rigged to come up six 50% of the time?

You may find this table useful:

x	y	$-(x/y) * \log_2(x/y)$	x	y	$-(x/y) * \log_2(x/y)$	x	y	$-(x/y) * \log_2(x/y)$	x	y	$-(x/y) * \log_2(x/y)$
1	2	0.50	4	5	0.26	6	7	0.19	5	9	0.47
1	3	0.53	1	6	0.43	1	8	0.38	7	9	0.28
2	3	0.39	5	6	0.22	3	8	0.53	8	9	0.15
1	4	0.5	1	7	0.40	5	8	0.42	1	10	0.33
3	4	0.31	2	7	0.52	7	8	0.17	3	10	0.52
1	5	0.46	3	7	0.52	1	9	0.35	7	10	0.36
2	5	0.53	4	7	0.46	2	9	0.48	9	10	0.14
3	5	0.44	5	7	0.35	4	9	0.52			

Exercise 2. Information gain

Consider the following set of training examples:

a1	a2	class
T	T	+
T	T	+
T	F	-
F	F	+
F	T	-
F	T	-

- What is the entropy of this collection of training examples with respect to the class?
- What is the information gain of a2?
- Which attribute will be selected as the first attribute to split on?
- Draw the decision tree after the selection of the attribute from the previous step. Indicate the child node(s) that: 1) need to be further expanded by adding additional tests and 2) can be converted to leafs, and show their class.

Exercise 3. Decision trees (Advanced only)

In the recursive construction of DTs, it sometimes occurs that a mixed set of positive and negative examples remains at a leaf node, even after all the attributes have been used. Consider binary classification and suppose that we have p positive examples and n negative examples at a leaf node and there are no more attributes left. Show that returning the probability $p/(p+n)$ minimizes the sum of squared errors of the set of these examples.

Hint: The error on a given training set can be written as a mathematical expression and viewed as a function of what is returned (i.e. the class chosen - in this case a single number 0 or 1 at the leaf as the task is binary classification) and the number of positive and negative examples, i.e. :

Let's x be the returned class ($x=0$ or 1). The error function E depends on x , p and n : $E=f(x,p,n)$.

The absolute error on the set of examples is: $E = p(1-x) + nx$. Let's check that it makes sense:

- if $x=0$ (negative class is returned), the n negative examples will be classified correctly but the p positive examples will be misclassified $\Rightarrow E=p$
- if $x=1$ (positive class is returned), the p positive examples will be classified correctly but the n negative examples will be misclassified $\Rightarrow E=n$

The sum of the squared errors is: $E^2=p(1-x)^2+nx^2$.

Show that E^2 is minimized for $x=p/(n+p)$.

Exercise 4. Using Weka

1. Load iris data (iris.arff). Choose evaluation on "training set".
2. Run the decision tree (j48) classifier under "trees. In Weka there are 2 versions: id3 which is only for nominal attributes and j48 for both numeric and nominal. "

Look at the decision tree that is produced, do you understand it?

Test mode: evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree

```

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petallength <= 4.9: Iris-versicolor (48.0/1.0)
| | petallength > 4.9
| | | petalwidth <= 1.5: Iris-virginica (3.0)
| | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| petalwidth > 1.7: Iris-virginica (46.0/1.0)

```

Number of Leaves : 5

Size of the tree : 9

This is the pruned DT. Weka uses pruning by default.

- a) Draw the DT.
- b) As you can see in some paths we test the same attribute more than once. Why? If the attributes are nominal, we never test the same attribute twice along one path.
- c) What are the number in brackets, e.g. (48.0/1.0)?

3. Change the evaluation to "Cross validation" (10 folds). Run ZeroR, OneR, Naïve Bayes, 3-nearest neighbor (IBL-3) and the decision tree (j48.) classifiers.

- a) Compare the classification accuracies. Which is the most accurate classifier?
- b) What output does each of the classifiers generate (e.g. a set of rules, etc.)? Suppose that you have to explain to your client why a particular decision is taken for a new example by these classifiers. The output of which classifier(s) are easier to understand and use for decision making?

Additional exercises (to be done at your own time)**Exercise 5. – Information gain**

Consider the following data from credit history of loan application.

credit history	debt	collateral	income	risk
bad	high	none	0-15k	high
unknown	high	none	15-35k	high
unknown	low	none	15-35k	moderate
unknown	low	none	0-15k	high
unknown	low	none	over 35k	low
unknown	low	adequate	over 35k	low
bad	low	none	0-15k	high
bad	low	adequate	over 35k	moderate
good	low	none	over 35k	low
good	high	adequate	over 35k	low
good	high	none	0-15k	high
good	high	none	15-35k	moderate
good	high	none	over 35k	low
bad	high	none	15-35k	high

You may find this table useful:

x	y	$-(x/y)^* \log_2(x/y)$	x	y	$-(x/y)^* \log_2(x/y)$	x	y	$-(x/y)^* \log_2(x/y)$	x	y	$-(x/y)^* \log_2(x/y)$	x	y	$-(x/y)^* \log_2(x/y)$
1	2	0.50	4	5	0.26	6	7	0.19	5	9	0.47	2	11	0.45
1	3	0.53	1	6	0.43	1	8	0.38	7	9	0.28	3	11	0.51
2	3	0.39	5	6	0.22	3	8	0.53	8	9	0.15	4	11	0.53
1	4	0.5	1	7	0.40	5	8	0.42	1	10	0.33	5	11	0.52
3	4	0.31	2	7	0.52	7	8	0.17	3	10	0.52	6	11	0.48
1	5	0.46	3	7	0.52	1	9	0.35	7	10	0.36	7	11	0.42
2	5	0.53	4	7	0.46	2	9	0.48	9	10	0.14	8	11	0.33
3	5	0.44	5	7	0.35	4	9	0.52	1	11	0.31	9	11	0.24

x	y	$-(x/y)^* \log_2(x/y)$	x	y	$-(x/y)^* \log_2(x/y)$	x	y	$-(x/y)^* \log_2(x/y)$
10	11	0.13	4	13	0.52	12	13	0.11
1	12	0.30	5	13	0.53	1	14	0.27
5	12	0.53	6	13	0.51	3	14	0.48
7	12	0.45	7	13	0.48	5	14	0.53
11	12	0.12	8	13	0.43	9	14	0.41
1	13	0.28	9	13	0.37	11	14	0.27
2	13	0.42	10	13	0.29	13	14	0.10
3	13	0.49	11	13	0.20			

- What is the amount of information (entropy) in the above set of examples with respect to the target classification?
- What is the information gain when using the attribute credit history to predict the risk?
- Which attribute will be selected as the first attribute to split on? Why?