Workshop Week 8
Exercises 3-6

3- Consider the following data set for a binary class problem and consider building a decision tree using this data.

Feature A	Feature B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

- Write a formula for the information gain when splitting on feature A.

- Contingency Table after splitting on feature A

	A = T	A = F
+	4	0
-	3	3

- The overall entropy before splitting :

$$E_{Orig} = -\,0.4 \log 0.4 \ -0.6 \log 0.6 = 0.9710$$

- The information gain after splitting on A is:

$$E_{A=T} = -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7} = 0.9852$$

$$E_{A=F} = -\frac{3}{3}\log\frac{3}{3} - \frac{0}{3}\log\frac{0}{3} = 0$$

$$\Delta = \ E_{Orig} - \frac{7}{10}E_{A=T} - \frac{3}{10}E_{A=F} = 0.2813$$

3- Consider the following data set for a binary class problem and consider building a decision tree using this data.

Feature A	Feature B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

- Write a formula for the information gain when splitting on feature B.

- Contingency Table after splitting on feature B

	B = T	B = F
+	?	?
-	?	?

- The overall entropy before splitting :

$$E_{Orig} = -0.4 \log 0.4 \ -0.6 \log 0.6 = 0.9710$$

- The information gain after splitting on B is:

$$E_{B=T} = -?\log ? - ?\log ? = ?$$

$$E_{B=F} = -?\log ? - ?\log ? = ?$$

$$\Delta = \ E_{Orig} - ?\,E_{B=T} - ?\,E_{B=F} \ = ?$$

3- Consider the following data set for a binary class problem and consider building a decision tree using this data.

Feature A	Feature B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

- Write a formula for the information gain when splitting on feature B.

- Contingency Table after splitting on feature B

	B = T	B = F
+	3	1
-	1	5

- The overall entropy before splitting :

$$E_{Orig} = -\,0.4 \log 0.4 \;-0.6 \log 0.6 = 0.9710$$

- The information gain after splitting on B is:

$$E_{B=T} = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113$$

$$E_{B=F} = -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6} = 0.6500$$

$$\Delta = E_{Orig} - \frac{4}{10}E_{B=T} - \frac{6}{10}E_{B=F} = 0.2565$$

- The information gain after splitting on A is:

$$\Delta = E_{Orig} - \frac{7}{10}E_{A=T} - \frac{3}{10}E_{A=F} = 0.2813$$

- The information gain after splitting on B is:

$$\Delta = E_{Orig} - \frac{4}{10}E_{B=T} - \frac{6}{10}E_{B=F} = 0.2565$$

- Therefore attribute ? will be chosen to split the node

- Therefore attribute A will be chosen to split the node

4- Consider the following simple dataset: Classify the point x=5.0 according to its 1-, 3-, 5- and 9-nearest neighbors.

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-

- 1-NN: x=5.0 should be classified as ? +
- 3-NN: x=5.0 should be classified as ? –
- 5-NN: x=5.0 should be classified as ? +
- 9-NN: x=5.0 should be classified as ? –

5- Describe two ways one could use a decision tree to make a classification where the test instance may have missing feature values.

Answer:

- Could either i) Impute the missing values and then use the decision tree (the merit of doing this will depend on the accuracy of imputation), or ii) When at prediction time we encounter a node in the decision tree which tests a variable A, and for that variable we have in our instance a missing value than all the possibilities are explored. Thus, for each possible sub-node a prediction is made. We keep the distribution for each sub node and we add them. Finally the class chosen for prediction is the class with the biggest density value.

- See also http://uksim.info/aims2015/CD/data/8675a122.pdf

6- Suppose Alice takes a dataset D with 100 instances, 4 features, plus a class label feature.

- She computes the correlation of each of the 4 features with the class label using mutual information and discards the two features with lowest correlation.
- She now has a processed version D’ of the dataset (2 features, class label feature and 100 instances).
- She splits D’ into two - 80% training (80 instances) and 20% testing (20 instances).
- She learns a decision tree model on the training set and evaluates the model accuracy on the testing set.
- She reports the accuracy as being 90%. Why might this estimate of 90% accuracy be over-optimistic? Give reasons.

6- Suppose Alice takes a dataset D with 100 instances, 4 features, plus a class label feature.

Answer:

The 90% estimate would be biased, since the testing data (class label info) was looked at when doing feature selection. This provided information to the feature selection process that should not have been seen. (like seeing the final exam before it is held). Consequently the model that was trained using the results from the feature selection was developed on information that should not have been seen. The reported accuracy will thus likely be over optimistic.