Intelligent Systems
Exercises Block 2 Chapter 3
Classification Trees

Escuela Técnica Superior de Ingeniería Informática
Dep. de Sistemas Informáticos i Computación
Universitat Politècnica de València

28 de octubre de 2016

Questions

1 $\boxed{C}$ (Exam 18th January 2013) About the components involved in a Decision Classification Tree (DCT), which of the following statements is FALSE?

A) The queries of the splits are of the form $\mathbf{y} \in B?$, $B \subseteq E$
B) The quality of a split is measured by the impurity reduction produced by the split, and the there are different impurity measures such as the entropy impurity or misclassification impurity
C) In a DCT, it is desirable that the resubstitution error estimate is 0
D) A good class label for a terminal node $t \in \hat{T}$ es: $c^\star = \arg\max_c \hat{P}(c \mid t)$

2 $\boxed{A}$ (Exam 18th January 2013) We want to build a classification tree T from a given set of learning samples for a four-class ($C = 4$) classification problem. During the process of building the tree, we obtain three nodes with the following estimated probabilities that each node belongs to a class:

t	$\hat{P}(1 \mid t)$	$\hat{P}(2 \mid t)$	$\hat{P}(3 \mid t)$	$\hat{P}(4 \mid t)$
t_1	2^{-2}	2^{-2}	2^{-2}	2^{-2}
t_2	2^{-1}	2^{-1}	0	0
t_3	2^{0}	0	0	0

Which node is the most impure according to the entropy concept?

A) t_1. $\mathcal{I}(t_1) = -(4 \cdot 2^{-2} \log_2 2^{-2}) = 2$
B) t_2. $\mathcal{I}(t_2) = -(2 \cdot 2^{-1} \log_2 2^{-1}) = 1$
C) t_3. $\mathcal{I}(t_3) = -(1 \cdot 2^{0} \log_2 2^{0}) = 0$
D) There is not a single node that is more impure than the others.

3 $\boxed{A}$ (Exam 30th January 2013) Let t be a node in a classification tree with $N(t)$ elements of $\mathbb{R}^D$, such that $N_c(t)$ elements belong to class c. Show which of the following statements is FALSE:

A) A split with minimal entropy $H(t)$ is an optimal split
B) t can be considered a terminal node if the impurity reduction obtained with the best split is not large enough; that is, if $max_{j,r}\ \Delta\mathcal{I}(j, r, t) < \epsilon$ where $j \in \{1, 2, \ldots, D\}$ is a dimension and $r \in \mathbb{R}$ is a threshold of a one-dimension split.
C) A good way to assess the impurity of t is through the entropy, $H(t)$, as the number of bits associated to the decision among the classes represented in t.
D) If t is considered a terminal or leaf node, it is recommendable to assign it a class label $c^\star$ such that $N_{c^\star}(t)/N(t)$ is maximum.

4 $\boxed{C}$ (Exam 15th January 2014) Consider a classification problem in C classes $c = 1, \ldots, C$ for which we have learnt a classification tree T. Let t be a node whose impurity is given by the entropy, $H(t)$, associated to the conditional (posterior) probability of class c in node t, $P(1 \mid t), \ldots, P(C \mid t)$. t will be considered as a totally pure node:

A) When classes have the same probability; that is, $P(1 \mid t) = \cdots = P(C \mid t) = \frac{1}{C}$.
B) It exists a class c^* whose probability is higher than the rest of classes; that is, $P(c^* \mid t) > P(c \mid t)\ \forall c \neq c^*$.
C) It exists a class c^* with probability 1; that is, $P(c^* \mid t) = 1$.
D) None of the above.

5 $\boxed{\text{A}}$ (Exam 15th January 2014) We have a classification problem in two classes $c = 1, 2$ for objects represented by means of two-dimensional feature vectors, i.e. $\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2$. We have four training samples: $\mathbf{y_1} = (1, 0.2)^t$, belongs to class 1; and $\mathbf{y_2} = (2, 0.2)^t$, $\mathbf{y_3} = (3, 0.8)^t$ and $\mathbf{y_4} = (1, 0.8)^t$, belong to class 2. We want to build a Decision Classification Tree (DCT) for this problem by using the impurity reduction (in terms of entropy) to measure the quality of a split. In the case of the root node, and assuming we only consider the feature y_1, which of the following statements is **true**? (Note: $log(1/3) = -1.585, log(2/3) = -0.585$).

A) The best split is $y_1 \leq 1$.
B) The best split is $y_1 \leq 2$.
C) The best split is $y_1 \leq 3$.
D) None of the above.

6 $\boxed{\text{C}}$ (Exam 15th January 2014) Consider a classification problem in C classes $c = 1, \ldots, C$ for which we have learnt a classification tree T. Let t be a terminal node of T with posterior probabilities $\hat{P}(1 \mid t), \ldots, \hat{P}(C \mid t)$. A simple criterion to assign a class label to t is:

A) The class with minimum posterior probability.
B) The class with posterior probability close to the mean (i.e. $\frac{1}{C}$).
C) The class with maximum posterior probability.
D) None of the above.

7 $\boxed{\text{B}}$ (Exam 28th January 2014) Consider a classification problem in two classes, $c = 1, 2$, for objects represented by means of real-valued two-dimensional feature vectors; that is, $\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2$. Let T be a classification tree for this problem and t a non-terminal node of T. Let B_1 and B_2 represent the minimum and maximum values of the two features for classes 1 and 2, respectively; particularly, B_1 and B_2 represent $[\min y_1, \min y_2] \times [\max y_1, \max y_2]$ for each class, respectively. Let's assume that $B_1 = [1.5, 0.6] \times [2.3, 3.5]$ for class 1 and $B_2 = [2.5, 1.3] \times [3.8, 3.2]$ for class 2. In terms of impurity reduction (measured with the entropy), which is the best split for t?

A) $y_1 \leq 3.8$
B) $y_1 \leq 2.3$
C) $y_2 \leq 1.3$
D) $y_2 \leq 3.5$

8 $\boxed{\text{C}}$ (January 13, 2015) Let's consider a decision of classification among 4 classes, A, B, C, D, whose probabilities are $P_A = P_B = P_C = P_D$. According to the entropy concept, the impurity of this decision is ...

A) $+\infty$
B) the minimum possible impurity
C) 2 bits
D) lower than the impurity of a decision where $P_A = P_B \neq P_C = P_D$

9 $\boxed{\text{C}}$ (January 13, 2015) During the execution of the algorithm DCT (algorithm for learning a Decision Classification Tree), how many recursive calls are made?

A) two recursive calls in all cases
B) no recursive calls because DCT is an iterative algorithm
C) no recursive calls if the node is declared as a terminal node or two calls otherwise
D) one recursive call if the node is declared as a terminal node or two calls otherwise

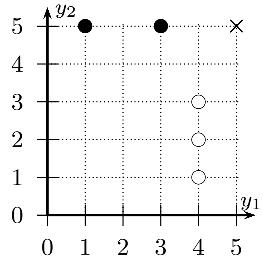
10 $\boxed{\text{D}}$ (January 2016) Let's assume we apply the Decision Classification Tree (DCT) algorithm for a two-class problem, A and B. The DCT algorithm reaches a node t which includes two data: one sample that belongs to class A and the other belongs to class B. The entropy impurity of t, $\mathcal{I}(t)$, is:

A) $\mathcal{I}(t) < 0.0$
B) $0.0 \leq \mathcal{I}(t) < 0.5$
C) $0.5 \leq \mathcal{I}(t) < 1.0$
D) $1.0 \leq \mathcal{I}(t)$

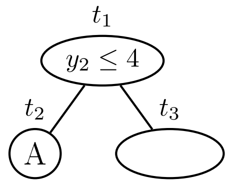
$$\mathcal{I}(t) = -\hat{P}(A \mid t) \log_2 \hat{P}(A \mid t) - \hat{P}(B \mid t) \log_2 \hat{P}(B \mid t) = -\tfrac{1}{2} \log_2 \tfrac{1}{2} - \tfrac{1}{2} \log_2 \tfrac{1}{2} = 1$$

Problems

1. (January 26, 2015) In order to learn a classification tree, we have six two-dimensional samples that belong to three classes, A, B and C. The feature vectors of these samples are shown in the figure on the right ($A = \circ$, $B = \bullet$ and $C = \times$).
After applying some recursive calls of the DCT algorithm (with $\epsilon = 0.5$ bits), we obtain the three-node subtree shown in the figure below. This subtree results from an optimal division of the samples into two subsets through the *"split"* $(2, 4.0)$ (that is, $y_2 \leq 4$, where y_2 is the vertical axis). The table below shows the values of some of the parameters obtained during the process of building this subtree.



Node	Split	$P(A \mid t_i)$	$P(B \mid t_i)$	$P(C \mid t_i)$	$P_{t_i}(L)$	$P_{t_i}(R)$	$\mathcal{I}(t_i)$	$\Delta\mathcal{I}(t_1)$
t_1	(2,4)	1/2	1/3	1/6	1/2	1/2	1.459	1.000
t_2	–	1	0	0	–	–	0	–
t_3		0	2/3	1/3				
t_4								
t_5								



a) Explain how the values $P(A \mid t_1)$, $P(B \mid t_1)$, $P(C \mid t_1)$, $P_{t_1}(R)$ and $\mathcal{I}(t_1)$ of the table are obtained.

b) Compute the impurity of the node t_3.

c) Find the optimal *"split"* for the node t_3 and fill out the blank cells of the table.

a) The root node (t_1) represents the six 6 available samples. 3 samples belong to class A, 2 to class B and 1 to class C. Therefore: $P(A \mid t_1) = 3/6 = 1/2$, $P(B \mid t_1) = 2/6 = 1/3$, $P(C \mid t_1) = 1/6$

The *split* $(2, 4.0)$ $(t_2 \leq 4)$ divides the tree into two subtrees: one rooting in t_2, which represents 3 samples such that $y_2 \leq 4$, and the other one rooting in t_3, which represents the other 3 data such that $y_2 > 4$. Then: $P_{t_1}(R) = 3/6 = 1/2$

Finally: $\mathcal{I}(t_1) \;=\; -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{3}\log_2\frac{1}{3} - \frac{1}{6}\log_2\frac{1}{6} \;\approx\; 1.459$ bits

b) $\mathcal{I}(t_3) \;=\; 0 - \frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \;\approx\; 0.918$ bits

c) The node t_3 represents the vectors $((1,5)^t, B), ((3,5)^t, B), ((5,5)^t, C)$. There are only two possible partitions, one corresponding to the *split* $y_1 \leq 2$ and the other one to the *split* $y_1 \leq 4$. The impurity reductions are:

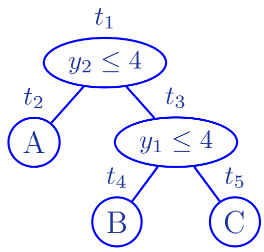
$\Delta\mathcal{I}(1,2,t_3) = \mathcal{I}(t_3) - \frac{1}{3}\mathcal{I}(t_4) - \frac{2}{3}\mathcal{I}(t_5) \;\approx\; 0.918 - 0 - \frac{2}{3}\cdot 1 \;=\; 0.251$ bits

$\Delta\mathcal{I}(1,4,t_3) = \mathcal{I}(t_3) - \frac{2}{3}\mathcal{I}(t_4) - \frac{1}{3}\mathcal{I}(t_5) \;\approx\; 0.918 - 0 - 0 \;=\; 0.918$ bits

The highest impurity reduction is for the split $(1,4)$; i.e, $y_1 \leq 4$.

The resulting tree and its parameters are shown below in the figure and table, respectively.

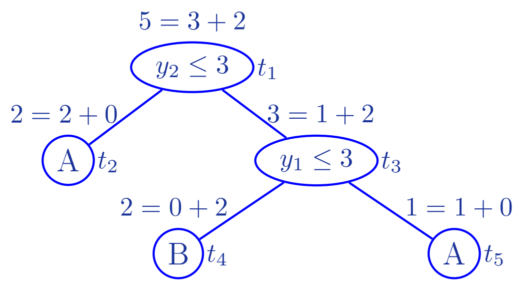
Node	Split	$P(A \mid t_i)$	$P(B \mid t_i)$	$P(C \mid t_i)$	$P_{t_i}(L)$	$P_{t_i}(R)$	$\mathcal{I}(t_i)$	$\Delta\mathcal{I}(t_1)$
t_1	(2,4)	1/2	1/3	1/6	1/2	1/2	1.459	1.000
t_2	–	1	0	0	–	–	0	–
t_3	(1,4)	0	2/3	1/3	2/3	1/3	0.918	0.918
t_4	–	0	1	0	–	–	0	–
t_5	–	0	0	1	–	–	0	–



2. **(1 point)** (January 2016) We have the 5 two-dimensional samples shown in the table to learn a classification tree. For each sample, we show its feature vector and the class it belongs to. The first *split* is $(2,3)$; that is, $y_2 \leq 3$; and the second and last split is $(1,3)$; that is, $y_1 \leq 3$.

y_1	2	2	2	4	6
y_2	2	4	6	6	2
c	A	B	B	A	A

a) Represent graphically the classification tree and classify the sample $(4,4)^t$



The sample $(4,4)^t$ goes through the tree until it reaches t_5. Therefore, the classification hypothesis is class A.

b) For each non-terminal node, t, calculate:

- Probability of the classes, $P(c \mid t),\ c \in \{A, B\}$
  $P(A \mid t_1) = 3/5, \quad P(B \mid t_1) = 2/5; \qquad P(A \mid t_3) = 1/3, \quad P(B \mid t_3) = 2/3$

- Probability of choosing the left node and the right node, $P_t(L),\ P_t(R)$
  $P_{t_1}(L) = 2/5, \quad P_{t_1}(R) = 3/5 \qquad P_{t_3}(L) = 2/3, \quad P_{t_3}(R) = 1/3$

c) Calculate the number of bits of the impurity, $\mathcal{I}(t_1)$, of the root node, t_1

$$\begin{aligned} \mathcal{I}(t_1) &= -P(A \mid t_1)\, log_2 P(A \mid t_1) - P(B \mid t_1)\, log_2 P(B \mid t_1) \\ &\approx -0.6(-0.737) \ -0.4(-1.322) \ = \ 0.971 \text{ bits.} \end{aligned}$$

d) For each terminal node, t, calculate:

- Probability of the terminal node, $P(t)$
  $P(t_2) = 2/5, \quad P(t_4) = 2/5, \quad P(t_5) = 1/5$

- Impurity in bits, $\mathcal{I}(t)$
  $\mathcal{I}(t_2) = \mathcal{I}(t_4) = \mathcal{I}(t_5) = 0$ bits.

e) Estimated resubstitution error (misclassification error) of the tree.
Since the three terminal nodes are pure nodes, the estimated resubstitution error is 0.