Result export

Created by	Ad Feelders
Creation date	10/7/25, 2:24 PM
Chapters	Result details, Answers
Empty chapters	Include

Index

Result details

Candidate:	Yi-An Liu
Code:	7578997
User groups:	#AlleKandidaten
Course:	Data Mining (INFOMDM)
Blueprint name:	Data Mining Homework Set 1
Blueprint type:	Graded Practice
Status:	Closed
Start date:	9/18/25, 10:46 PM
End date:	9/18/25, 10:48 PM
Number of questions:	5

Answers

Question 1

We want to determine the best split in a node containing the following data on numeric attribute x and class label y. The class label can
take on three different values, coded as A, B and C.

x	2	2	3	4	4	5	6	7	8	9
y	A	A	A	B	B	B	A	C	C	C

We use the gini-index as impurity measure.

What is the best split on x?

Score: 2 out of 2 pts.

○	Between x=2 and x=3	0 pts.
○	Between x=3 and x=4	0 pts.
○	Between x=4 and x=5	0 pts.
○	Between x=5 and x=6	0 pts.
◉	Between x=6 and x=7	✓ 2 pts.
○	Between x=7 and x=8	0 pts.
○	Between x=8 and x=9	0 pts.

Question 2

We want to determine the optimal split in a node that contains the following data:

x	a	a	a	b	b	c	c	d	d
y	0	0	1	0	1	1	1	0	0

Here x is a categorical attribute with possible values {a,b,c,d} and y is a binary class label with values 0 and 1. We use the gini-index
as impurity measure.

The best split on x is:

- ○ x ∈ {a,b} — 0 pts.
- ○ x ∈ {a,c} — 0 pts.
- ○ x ∈ {a,b,c} — 0 pts.
- ◉ x ∈ {a,b,d} — ✓ 2 pts.
- ○ x ∈ {a,d} — 0 pts.
- ○ x ∈ {b,c,d} — 0 pts.
- ○ x ∈ {a,c,d} — 0 pts.

Question 3

Answered on: 2025-09-18 22:47:01	Duration: 17 sec.	Weight factor: 1	Score: 2 of 2 pts.

The following binary classification tree has been grown on a training set with n=100 examples:

Node	t1	t2	t3 - leaf	t4 - leaf	t5	t6	t7 - leaf	t8 - leaf	t9 - leaf
Class A	50	30	30	0	20	10	10	0	10
Class B	50	10	0	10	40	40	0	40	0

The nodes have been numbered according to depth first preorder traversal. We listed the number of cases a node contains of each class below it.

Perform cost-complexity pruning on this tree, and answer the following questions.

The value of α_2 = 0.1 1 pt.   0.1 1 pt.

The value of α_3 = 0.2 1 pt.   0.2 1 pt.

Question 4

Answered on: 2025-09-18 22:47:16	Duration: 20 sec.	Weight factor: 1	Score: 2 of 2 pts.

Let n denote the number of observations and k the number of different classes occurring in node t. Let x be a numeric variable with all values distinct. We use the gini-index as impurity measure. Let S denote the number of splits we have to evaluate in order to determine the best split on x in t.

Give an expression for S in terms of n and/or k for the most unfavorable distribution of the class labels: S =

| n-1 | 1 pt. | | n-1 1 pt. | | n - 1 1 pt. |

Give an expression for S in terms of n and/or k for the most favorable distribution of the class labels: S =

| k-1 | 1 pt. | | k-1 1 pt. | | k - 1 1 pt. |

Question 5

| Answered on: 2025-09-18 22:47:32 | Duration: 25 sec. | Weight factor: 1 | Score: 2 of 2 pts. |

In learning classification trees, determination of the appropriate size of the tree is an important problem. One can control the size of the tree by using a so-called stopping rule to stop growing the tree early. One possibility to implement this idea is to use parameters nmin and minleaf. If a node contains less than nmin cases, then it becomes a leaf node. A split is not allowed if it creates a child node with less then minleaf cases. Assume the tree growing algorithm only makes binary splits.

Consider the following two parameter settings:
1. nmin=12 and minleaf=10
2. nmin=18 and minleaf=10
Would you expect the tree in case (1) to have a lower, higher, or the same error rate on the training sample (resubstitution error) as the tree in case (2)?

Score: 1 out of 1 pt.

○	lower		0 pts.
○	higher		0 pts.
◉	the same	✓	1 pt.

Answer the same question for the following parameter settings:
1. nmin=20 and minleaf=5
2. nmin=10 and minleaf=5

Score: 1 out of 1 pt.

○	lower		0 pts.
◉	higher	✓	1 pt.
○	the same		0 pts.