Lecture 7 –Classification Methodologies: Decision Trees and KNN

COMP20008

School of Computing and Information Systems

© University of Melbourne

Announcements

Reminder – oral zoom assessments for Assignment 1. Held next week 15th September to 19th September. See schedule linked from Canvas announcements to check when your group will be examined.

Further info about oral assessments is at

https://canvas.lms.unimelb.edu.au/courses/211662/assignments/577382

Past exams are now available on Canvas

https://canvas.lms.unimelb.edu.au/courses/211662/modules/items/6784212

Code Peer Review due today 23:59 Melbourne time

© University of Melbourne

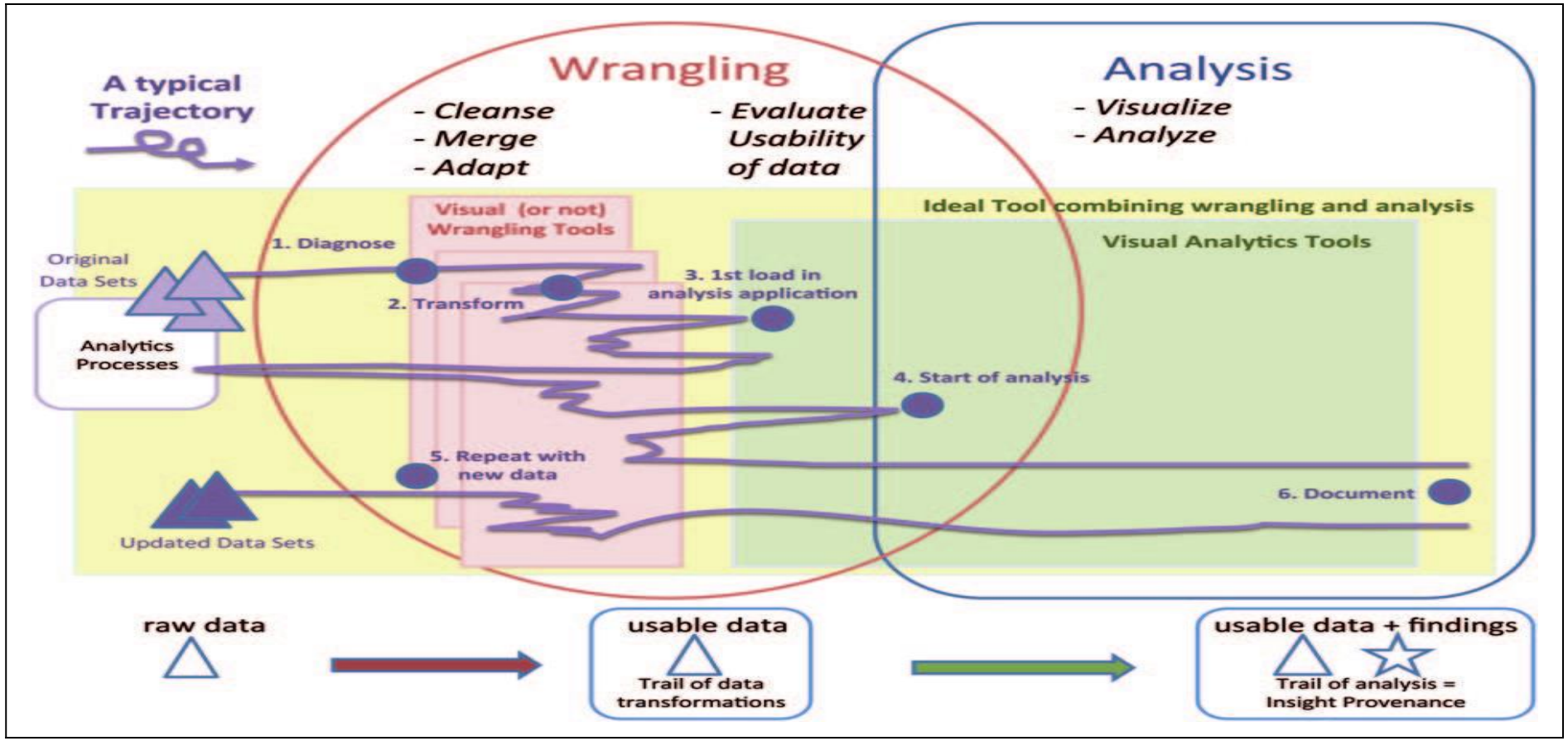
Lecture Agenda

Introduction to classification

- Decision trees
- K nearest neighbor

© University of Melbourne

Where are we now?



© University of Melbourne

Unsupervised and Supervised Learning

Two main types of machine learning techniques:

Unsupervised Learning – working with unlabeled data to discover patterns, groups, or relationships without prior knowledge of existing categories

- Clustering (Week 5)

Supervised Learning – works like a supervisor, teaching the machine using labelled data with correct classifications

- Classification (this week – Week 7)
- Linear Regression (next week – Week 8)

© University of Melbourne

Classification

We now start the topic of classification

- Making predictions about the future, based on historical data

Predictive modelling/classification

- The fun part of data science ?!

- A foundation for machine intelligence, AI, machines replacing humans and taking our jobs ….

© University of Melbourne

Classification

Predicting disease from microarray data (measured from a blood sample)

Training

	Gene 1	Gene 2	Gene 3	…	Gene n	Will develop cancer <1 year
Person 1	2.3	1.1	0.3	…	2.1	1
Person 2	3.2	0.2	1.2	…	1.1	1
Person 3	1.9	3.8	2.7	…	0.2	1
…	…	…	…	…	…	..
Person m	2.8	3.1	2.5	…	3.4	0

Testing

	Gene 1	Gene 2	Gene 3	…	Gene n	Develop Cancer<1year
Person m+1	2.1	0.9	0.6	…	1.9	?

Classification example

Animal classification

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

Test data

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
gila monster	cold-blooded	scales	no	no	no	yes	yes	?

https://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf

Classification example

Banking: classifying borrower

Tid	Home Owner (binary)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training set for predicting borrowers who will default on loan payments.

Test data

Tid	Home Owner	Marital status	Annual Income	Defaulted Borrower
11	No	Single	55K	?

Classification example

Detecting tax fraud/tax cheats

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Test data

Tid	Refund	Marital Status	Taxable Income	Tax Cheat
11	Yes	Married	125K	?

Classification: Definition

Given a collection of records (*training set*)

- Each record contains a set of *attributes*, one *class label*.

Find a predictive *model* for class label as a function of the values of other attributes.

$$y = f(x_1, x_2, \ldots, x_n)$$

- y: *discrete value*, target variable
- $x_1, \ldots x_n$: attributes, predictors
- f: is the predictive model (a tree, a rule, a mathematical formula, ..)

Goal: previously unseen records should be assigned a class as accurately as possible.

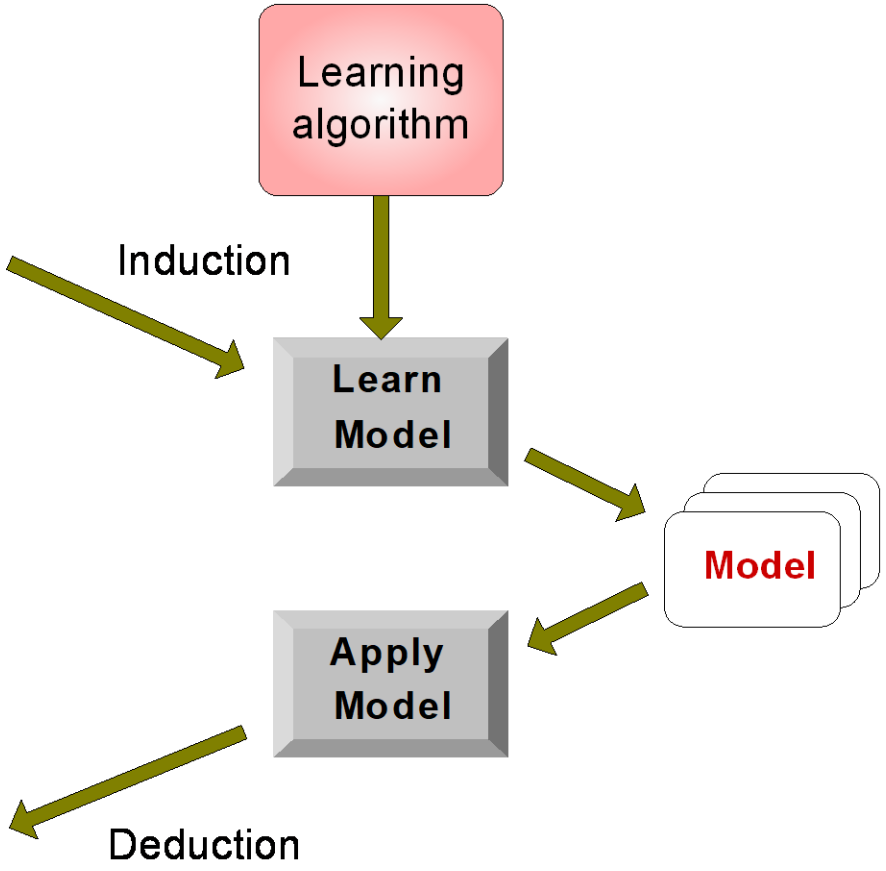
- A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

© University of Melbourne

Classification framework

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

© University of Melbourne

Regression

Given a collection of records (*training set*)

- Each record contains a set of *attributes*, one of *target variable*.

- Learn predictive model from data

$$y = f(x_1, x_2, \ldots, x_n)$$

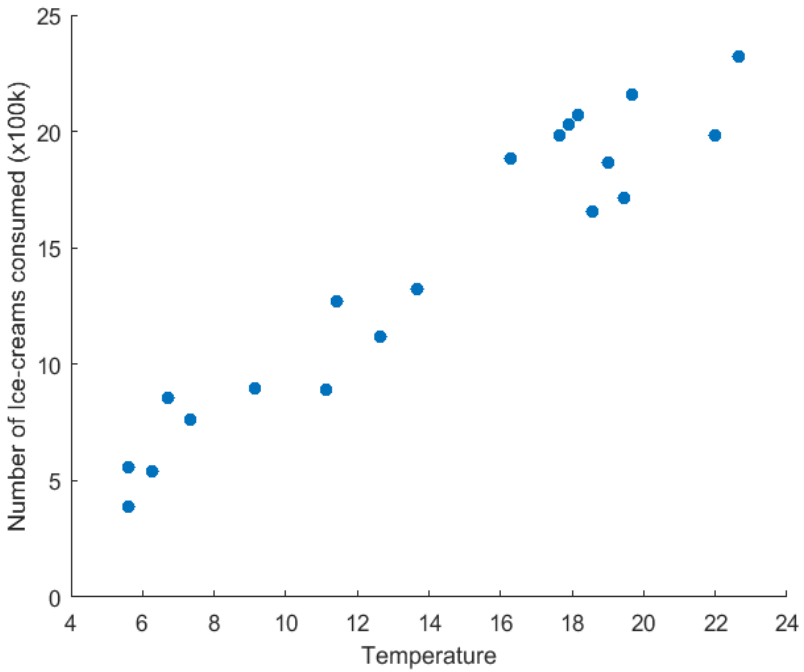
y: continuous real value, target variable

$x_1, \ldots x_n$: attributes, predictors

© University of Melbourne

Regression

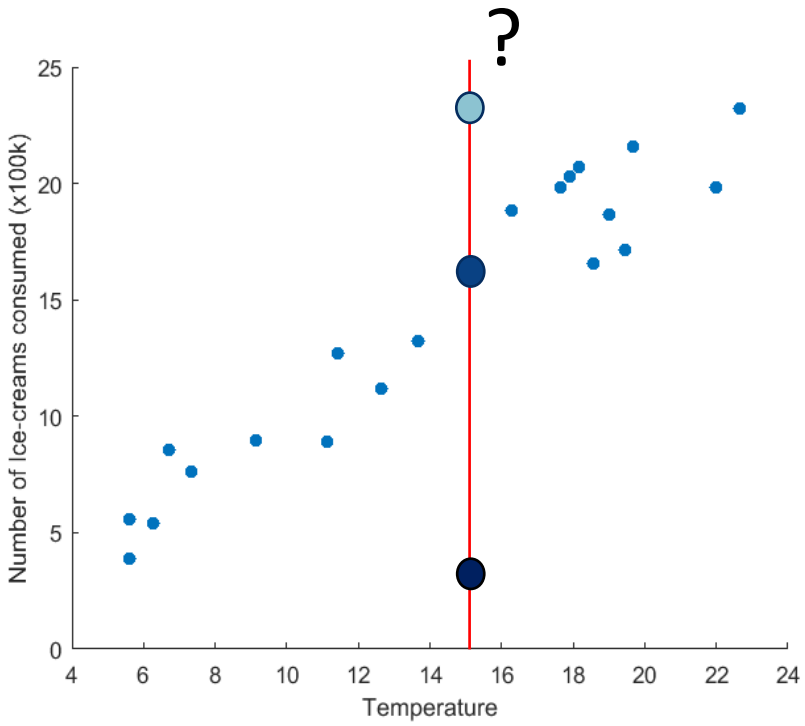
Predicting ice-creams consumption from temperature: $y = f(x)$



© University of Melbourne

Regression

Predicting ice-creams consumption from temperature: $y = f(x)$



Regression

Predicting activity level of a target gene

	Gene 1	Gene 2	Gene 3	…	Gene n	Gene n+1
Person 1	2.3	1.1	0.3	…	2.1	3.2
Person 2	3.2	0.2	1.2	…	1.1	1.1
Person 3	1.9	3.8	2.7	…	0.2	0.2
…	…	…	…	…	…	…
Person m	2.8	3.1	2.5	…	3.4	0.9

	Gene 1	Gene 2	Gene 3	…	Gene n	Gene n+1
Person m+1	2.1	0.9	0.6	…	1.9	?

Question 1

How would you describe the following task: Developing a predictive model for whether or not an email message is spam?

(A) Regression 0%

(B) Classification 0%

(C) Neither classification or regression 0%

Powered by Poll Everywhere

URL: https://go.unimelb.edu.au/8b7p

© University of Melbourne

Question 2

How would you describe the following task: Developing a predictive model for whether my house will sell or not in the auction tomorrow.

(A) Regression 0%

(B) Classification 0%

(C) Neither classification or regression 0%

URL: https://go.unimelb.edu.au/8b7p

© University of Melbourne

Question 3

How would you describe the following task: Partitioning a dataset into five groups.

(A) Regression

(B) Classification

(C) Neither classification or regression

URL: https://go.unimelb.edu.au/8b7p

© University of Melbourne

Question 4

How would you describe the following task: Strength of the relationship between a person's height and weight.

(A) Regression 0%

(B) Classification 0%

(C) Neither classification or regression 0%

URL: https://go.unimelb.edu.au/8b7p

© University of Melbourne

Question 5

How would you describe the following task: Constructing predictive model for the price that my house will sell for at the auction tomorrow.

(A) Regression

(B) Classification

(C) Neither classification or regression

URL: https://go.unimelb.edu.au/8b7p

© University of Melbourne

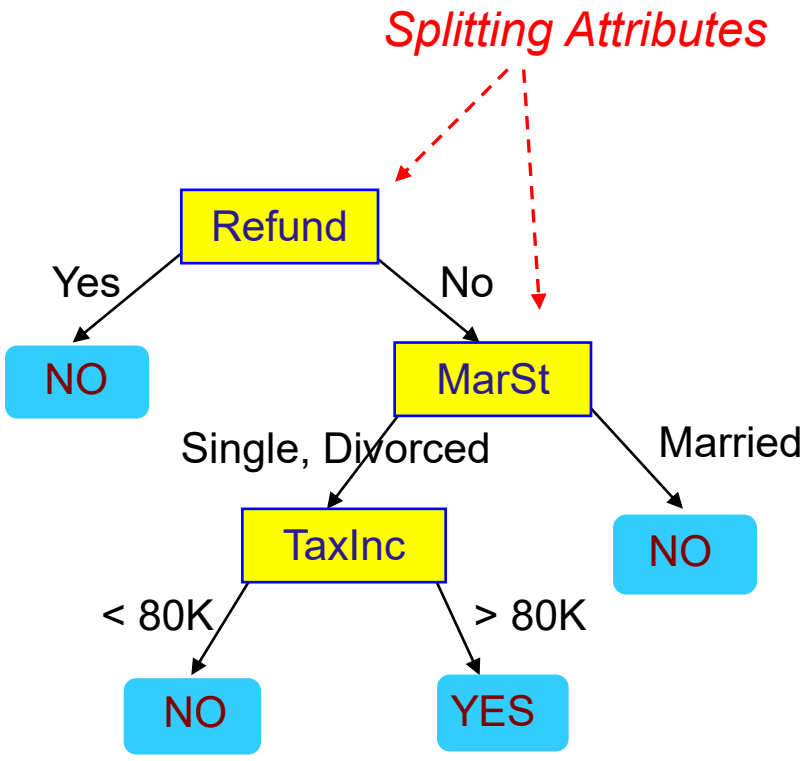
Supervised Learning Classification – Decision Tree

© University of Melbourne

Example of a Decision Tree: Prediction of tax cheating

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



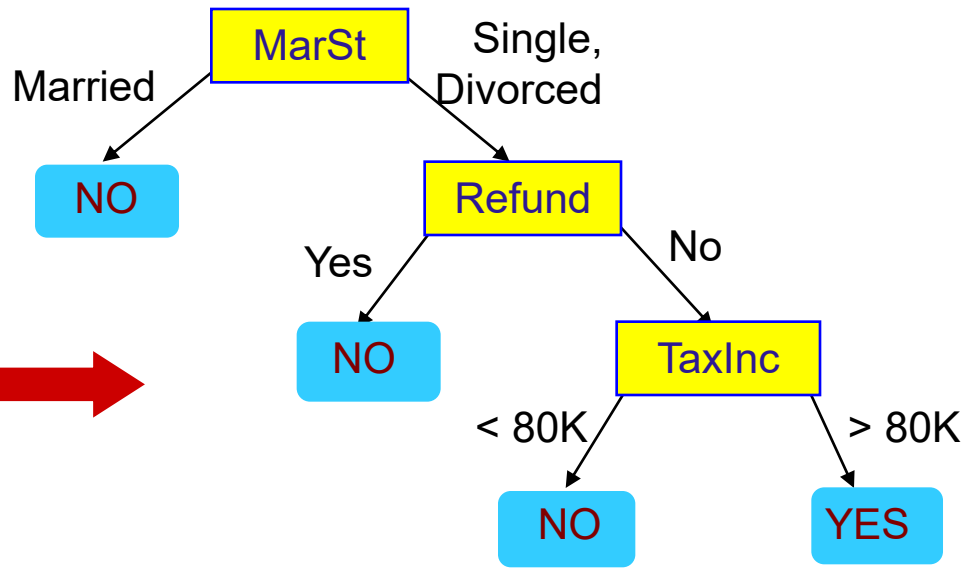
Model: Decision Tree

© University of Melbourne

Another Example of Decision Tree: prediction of tax cheating

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

There could be more than one tree that fits the same data!

© University of Melbourne

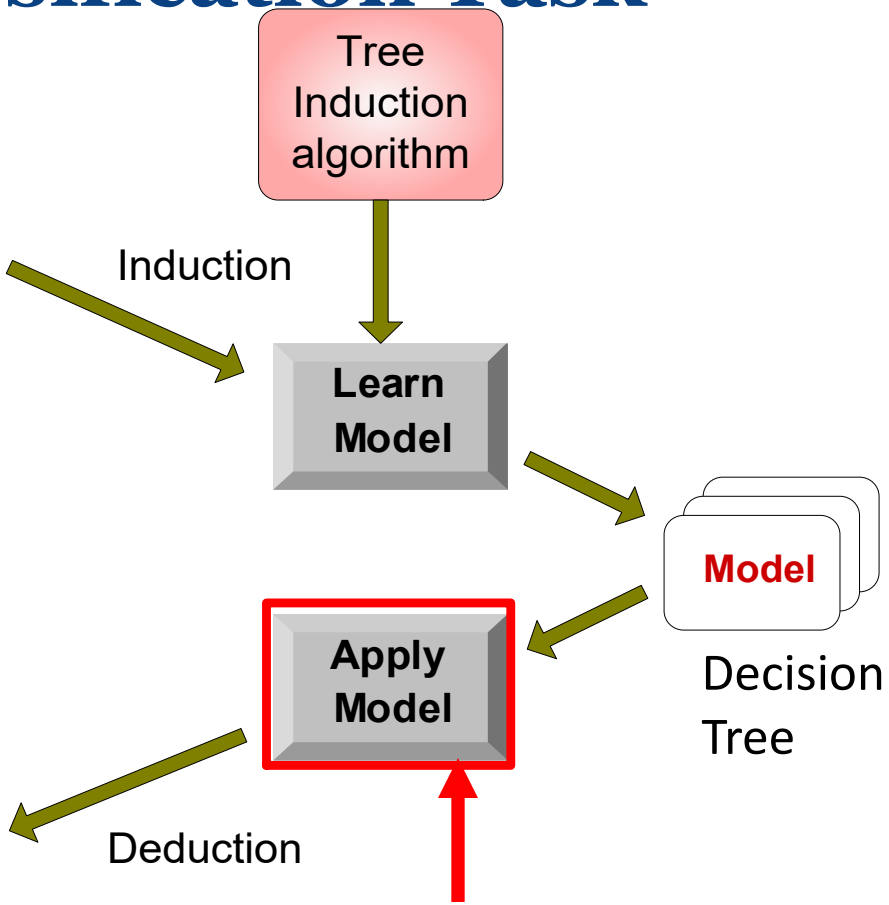
Decision Tree Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

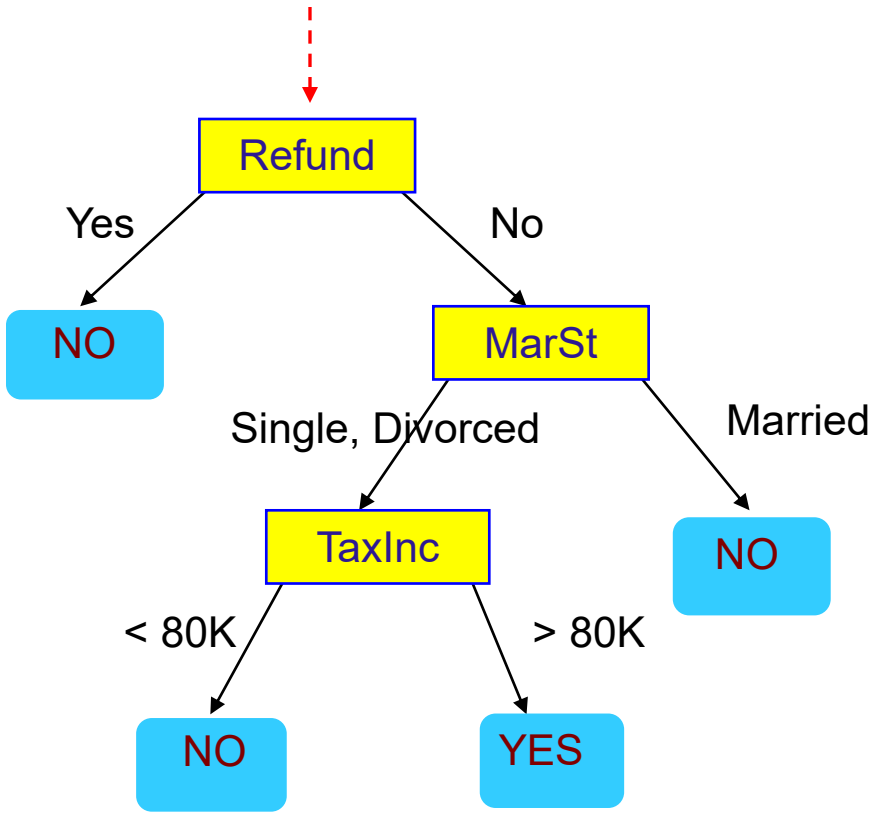
Test Set



© University of Melbourne

Apply Model to Test Data – Tax cheat prediction

Start from the root of tree.



Test Data

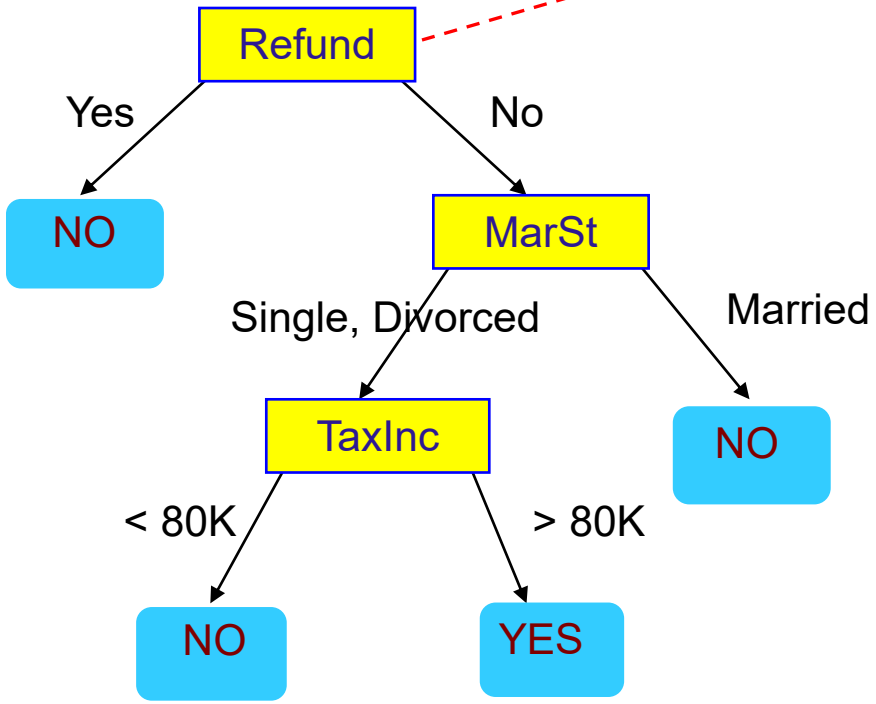
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

© University of Melbourne

Apply Model to Test Data – Tax cheat prediction

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

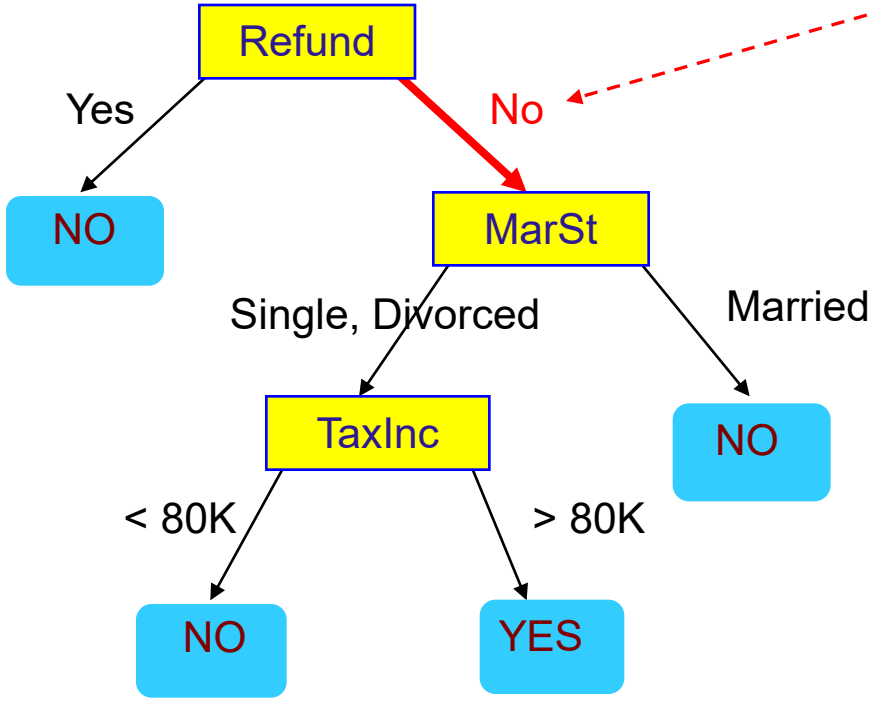


© University of Melbourne

Apply Model to Test Data – Tax cheat prediction

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

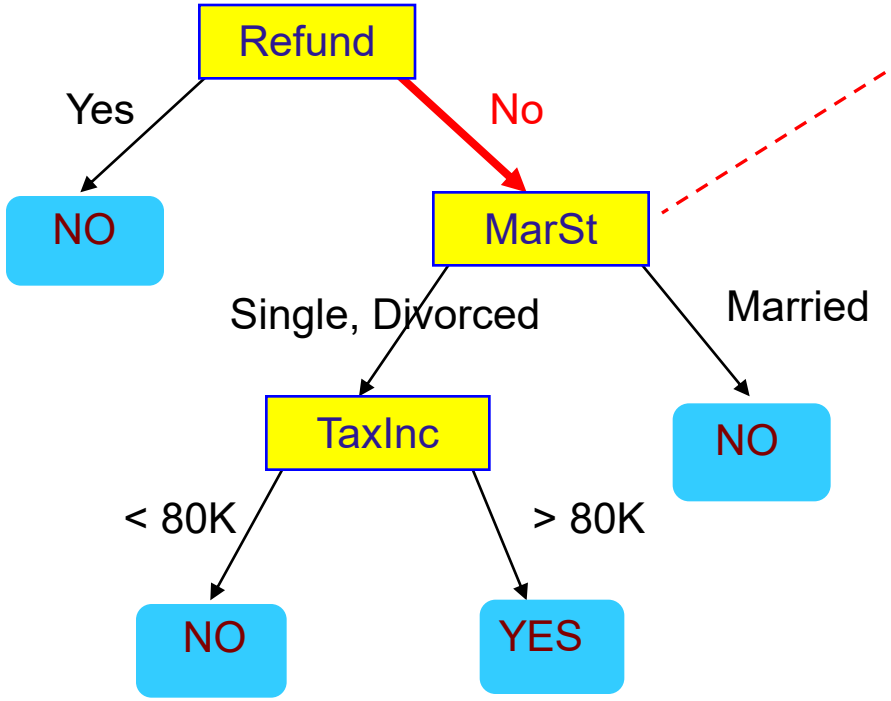


© University of Melbourne

Apply Model to Test Data – Tax cheat prediction

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

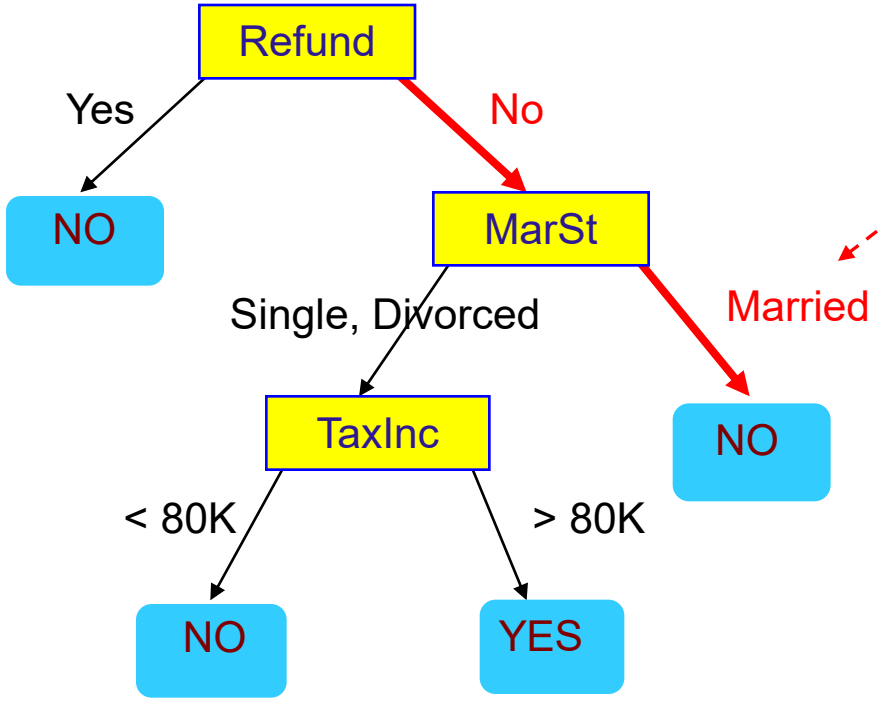


© University of Melbourne

Apply Model to Test Data – Tax cheat prediction

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

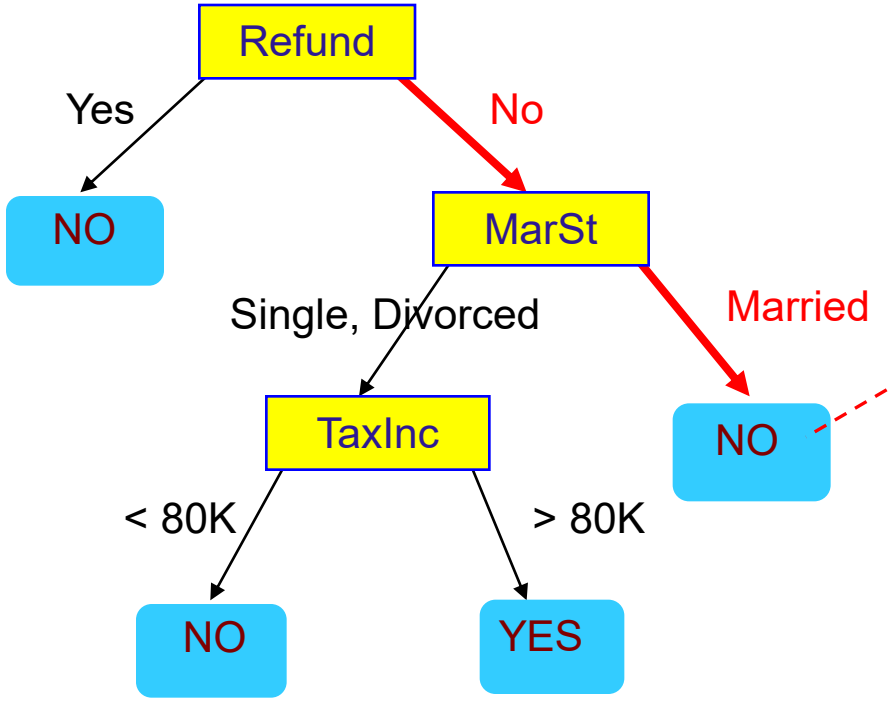


© University of Melbourne

Apply Model to Test Data – Tax cheat prediction

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

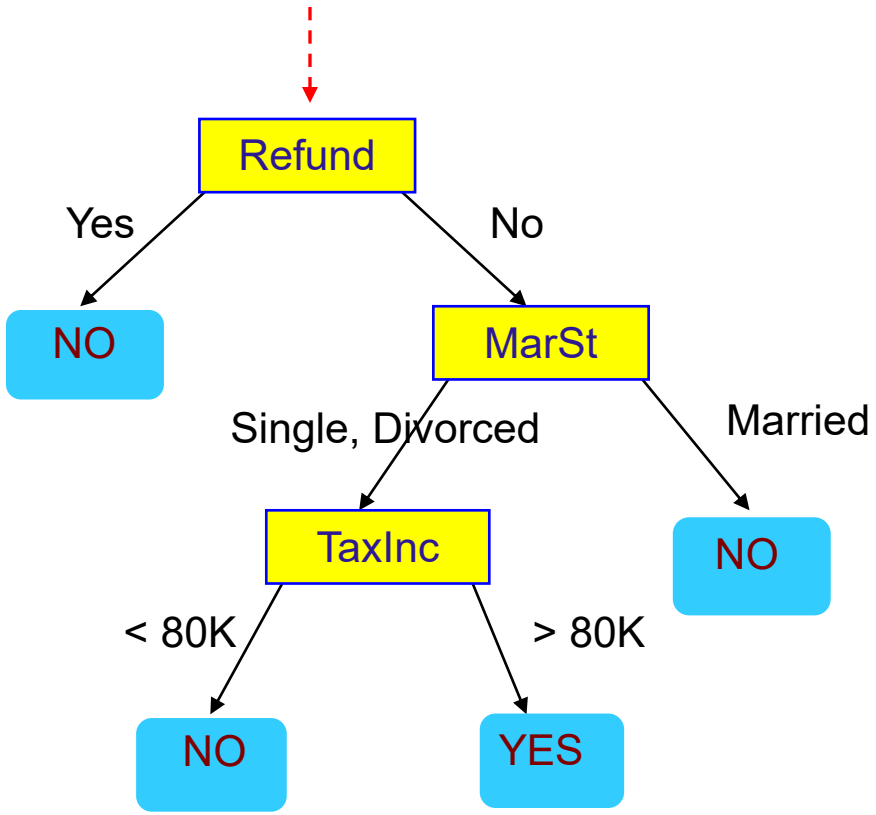


Assign Cheat to “No”

© University of Melbourne

Apply Model to Test Data – Tax cheat prediction

Start from the root of tree.



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Single	100K	?

© University of Melbourne

Decision Trees

- A flow-chart-like tree structure
- Internal node denotes a test on an attribute
- Branch represents an outcome of the test
- Leaf nodes represent class labels or class distribution

© University of Melbourne

Decision Tree Creation

Many Algorithms:

- We will look at a representative one (Hunt's algorithm)

How would you code an algorithm that automatically constructs a decision tree for the data below?

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

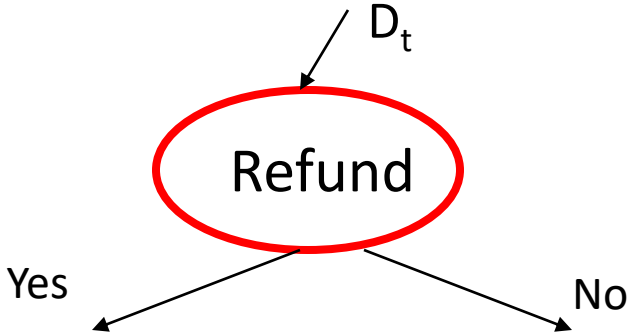
General Structure of Hunt's Algorithm

Let D_t be the set of training records that reach a node t

General Procedure:

- If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
- If D_t is an empty set, then t is a leaf node labeled by the default class, y_d
- If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

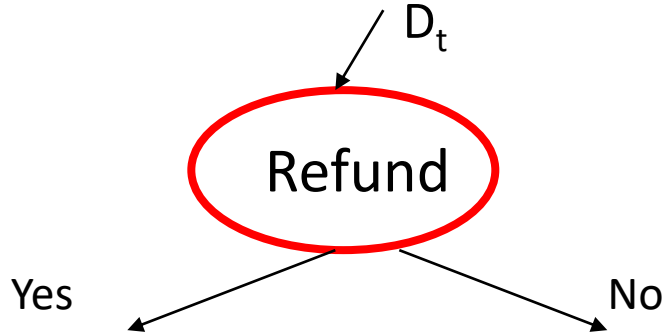


© University of Melbourne

Example

If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



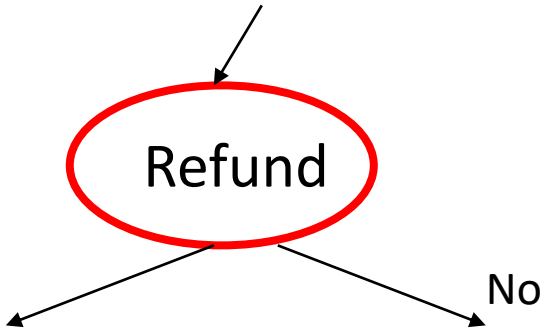
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
4	Yes	Married	120K	No
7	Yes	Divorced	220K	No

Tid	Refund	Marital Status	Taxable Income	Cheat
2	No	Married	100K	No
3	No	Single	70K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

© University of Melbourne

Stopping condition: leaf node

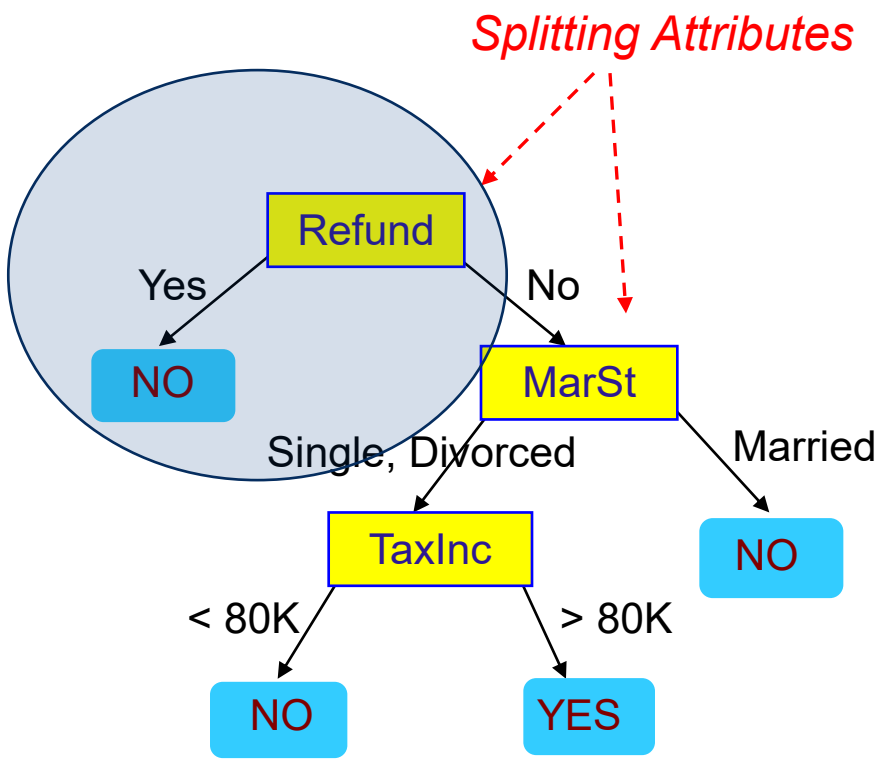
- If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
- If D_t is an empty set, then t is a leaf node labeled by the default class, y_d



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
4	Yes	Married	120K	No
7	Yes	Divorced	220K	No

© University of Melbourne

How do you know where to split on?



Model: Decision Tree

© University of Melbourne

Tree Induction (creation)

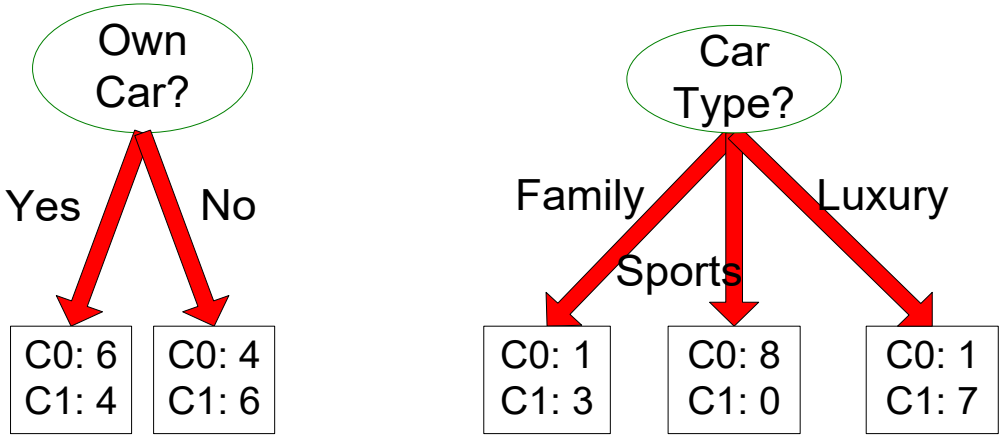
- Determine how to split the records
  - How to determine the best split?
  - How to specify the attribute test condition?

- Determine when to stop splitting
  - When node has only a single class of instances

© University of Melbourne

How to determine the Best Split

Before Splitting: **10 records of class 0,**
10 records of class 1



Which test condition is the best?

© University of Melbourne

How to determine the Best Split

Greedy approach:

- Nodes with homogeneous class distribution are preferred

Need a measure of node impurity:

C0: 5 C1: 5	C0: 9 C1: 1
Non-homogeneous, Low degree of purity	Homogeneous, High degree of purity

© University of Melbourne

Measures of Node Impurity

Entropy

- We have seen entropy last week, where it was used to measure the amount of uncertainty in an outcome

- *Entropy can also be viewed as an impurity measure*
  - The set {A,B,C,A,A,A,A,A} has low entropy: low uncertainty and **high purity**
  - The set {A,B,C,D,B,E,A,F} has high entropy: high uncertainty and **low purity**

© University of Melbourne

Node Impurity Criteria based on Entropy

Entropy (H) at a given node t:

$$H(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

- Measures homogeneity of a node.
  - Maximum (log n_c) when records are equally distributed among all classes (n_c is number of classes)
  - Minimum (0.0) when all records belong to one class

© University of Melbourne

Examples for computing Entropy

$$H(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Entropy = – 0 $\log_2$ 0 – 1 $\log_2$ 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (5/6) = 0.65

C1	**2**
C2	**4**

?

© University of Melbourne

Examples for computing Entropy

$$H(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Entropy = – 0 $\log_2$ 0 – 1 $\log_2$ 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (5/6) = 0.65

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

© University of Melbourne

What is entropy of this node?

$$H(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

C1	**13**
C2	**20**

P(C1) =

P(C2) =

Entropy =

© University of Melbourne

What is entropy of this node?

$$H(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

C1	**13**
C2	**20**

P(C1) = 13/33 P(C2) = $\frac{20}{33}$

Entropy = - (13/33)log(13/33) – (20/33)log(20/33)

© University of Melbourne

How good is a Split?

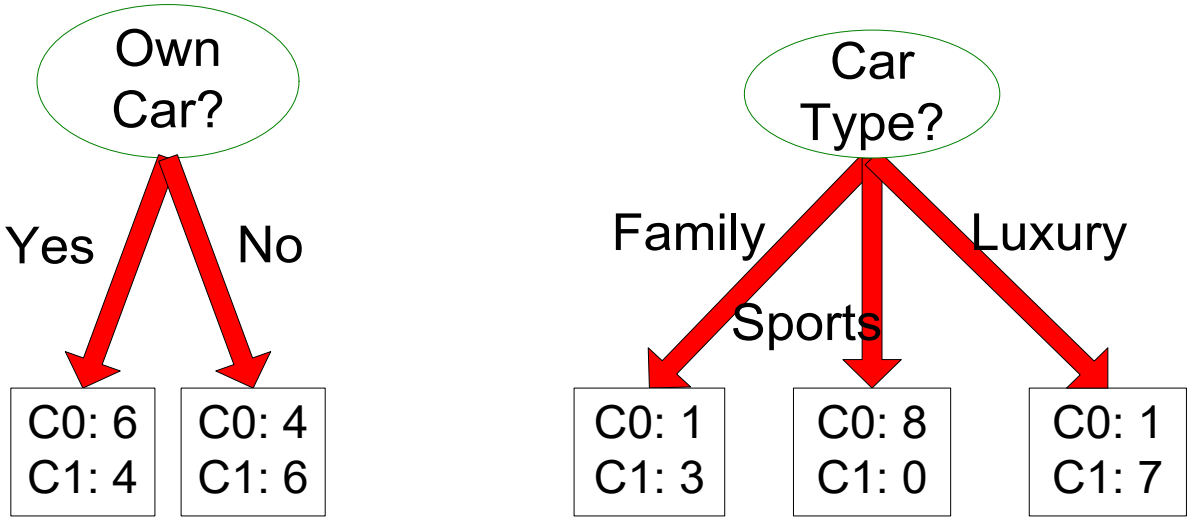
- Compare the impurity (entropy) of parent node (before splitting)
- With the impurity (entropy) of the children nodes (after splitting)

$$
\begin{aligned}
Gain = & \quad H(Parent) - H(Parent|Child) \\
= & \quad H(Parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} H(v_j)
\end{aligned}
$$

- $H(v_j)$: impurity measure of node v_j
- j: children node index
- $N(v_j)$: number of data points in child node v_j
- N: number of data points in parent node
- The larger the gain, the better

Note: Parent's entropy is fixed,
but we can control the children nodes

© University of Melbourne



© University of Melbourne

How good is a Split?

$$
\begin{aligned}
Gain = & \quad H(Parent) - H(Parent|Child) \\
= & \quad H(Parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} H(v_j)
\end{aligned}
$$

- Note: the information gain is equivalent to the mutual information between the class feature and the feature being split on
- Thus splitting using the information gain is to choose the feature with highest information shared with the class variable

© University of Melbourne

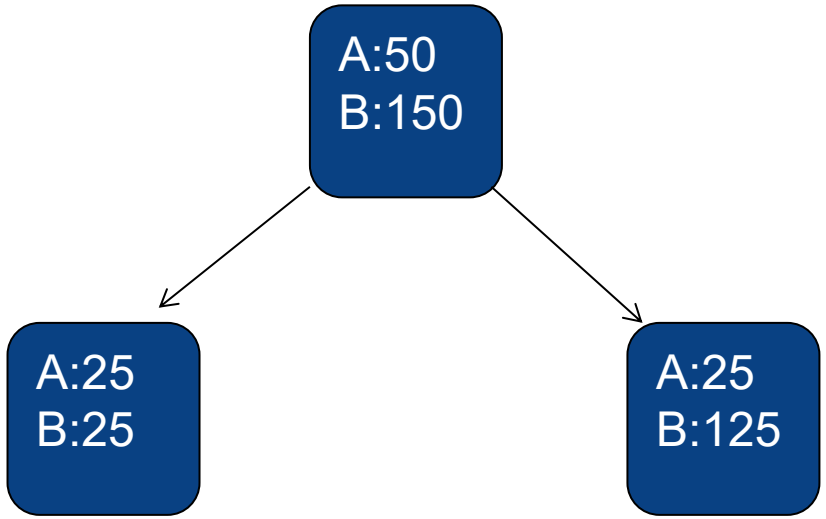
Question 2c) from 2016 exam

Given a dataset with two classes, A and B, suppose the root node of a decision tree has 50 instances of class A and 150 instances of class B. Consider a candidate split of this root node into two children, the first with (25 class A and 25 class B), the second with (25 class A and 125 class B). Write a formula to measure the utility of this split using the entropy criterion. Explain how this formula helps measure split utility.

© University of Melbourne

Question 2c) from 2016 exam

Given a dataset with two classes, A and B, suppose the root node of a decision tree has 50 instances of class A and 150 instances of class B. Consider a candidate split of this root node into two children, the first with (25 class A and 25 class B), the second with (25 class A and 125 class B). Write a formula to measure the utility of this split using the entropy criterion. Explain how this formula helps measure split utility

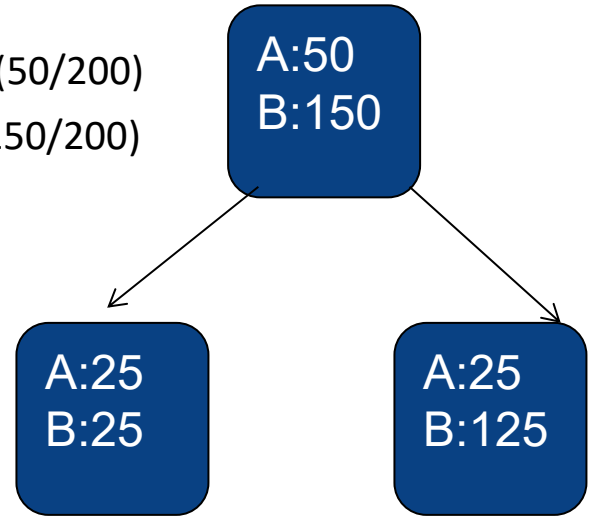


© University of Melbourne

Question 2c) from 2016 exam

Given a dataset with two classes, A and B, suppose the root node of a decision tree has 50 instances of class A and 150 instances of class B. Consider a candidate split of this root node into two children, the first with (25 class A and 25 class B), the second with (25 class A and 125 class B). Write a formula to measure the utility of this split using the entropy criterion. Explain how this formula helps measure split utility

Entropy(root)= -(50/200)log(50/200)
–(150/200)log(150/200)

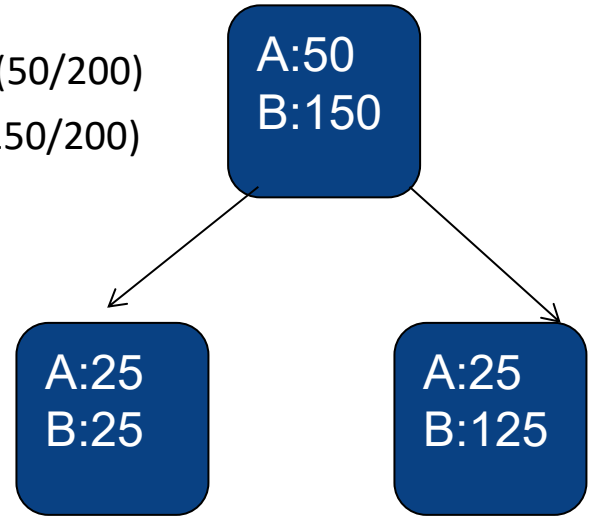


© University of Melbourne

Question 2c) from 2016 exam

Given a dataset with two classes, A and B, suppose the root node of a decision tree has 50 instances of class A and 150 instances of class B. Consider a candidate split of this root node into two children, the first with (25 class A and 25 class B), the second with (25 class A and 125 class B). Write a formula to measure the utility of this split using the entropy criterion. Explain how this formula helps measure split utility

Entropy(root)= -(50/200)log(50/200)
–(150/200)log(150/200)



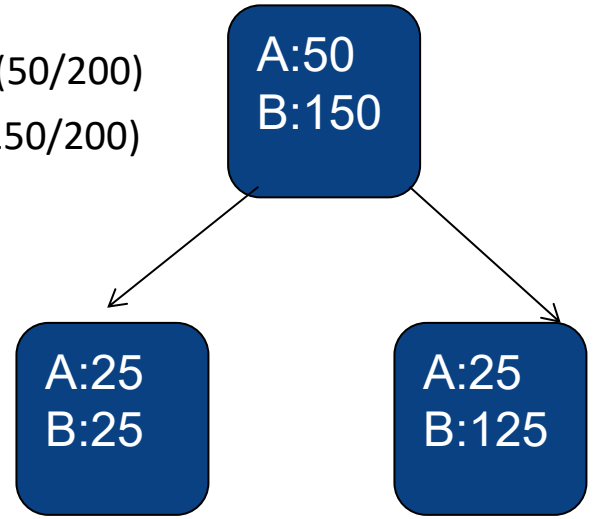
Entrop(left child)=
-(25/50)log(25/50)
-(25/50)log(25/50)

© University of Melbourne

Question 2c) from 2016 exam

Given a dataset with two classes, A and B, suppose the root node of a decision tree has 50 instances of class A and 150 instances of class B. Consider a candidate split of this root node into two children, the first with (25 class A and 25 class B), the second with (25 class A and 125 class B). Write a formula to measure the utility of this split using the entropy criterion. Explain how this formula helps measure split utility

Entropy(root)= -(50/200)log(50/200)
–(150/200)log(150/200)



Entrop(left child)=
-(25/50)log(25/50)
-(25/50)log(25/50)

Entropy(right child)=
-(25/150)log(25/150)
–(125/150)log(125/150)

© University of Melbourne

Question 2c) from 2016 exam

Given a dataset with two classes, A and B, suppose the root node of a decision tree has 50 instances of class A and 150 instances of class B. Consider a candidate split of this root node into two children, the first with (25 class A and 25 class B), the second with (25 class A and 125 class B). Write a formula to measure the utility of this split using the entropy criterion. Explain how this formula helps measure split utility

Split utility= Information Gain

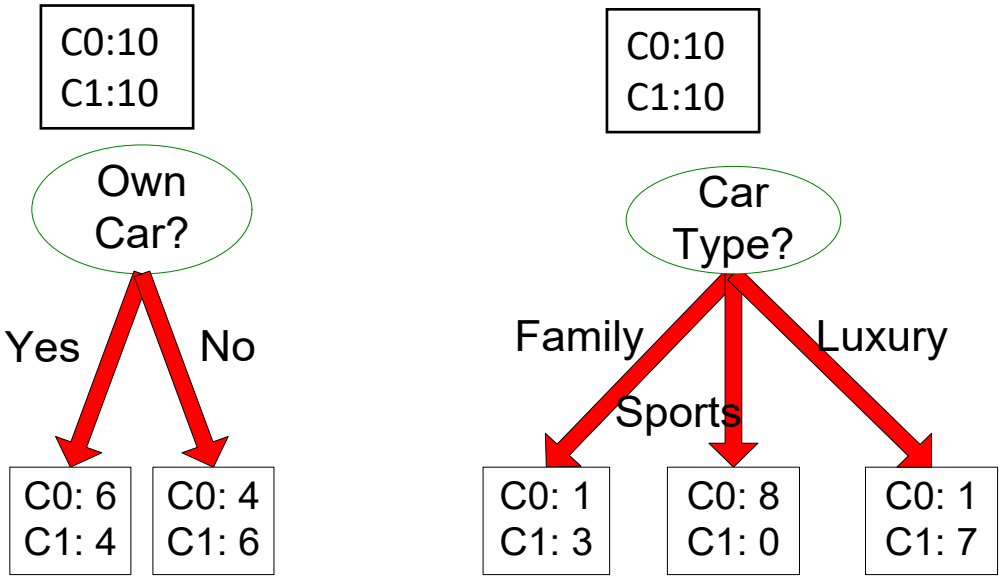
=Entropy(root) – Entropy(root|split)

=Entropy(root) – [(50/200)*Entropy(left child)
+(150/200)*Entropy(right child)]

© University of Melbourne

How to determine the Best Split?

Before Splitting: **10 records of class 0,**
10 records of class 1



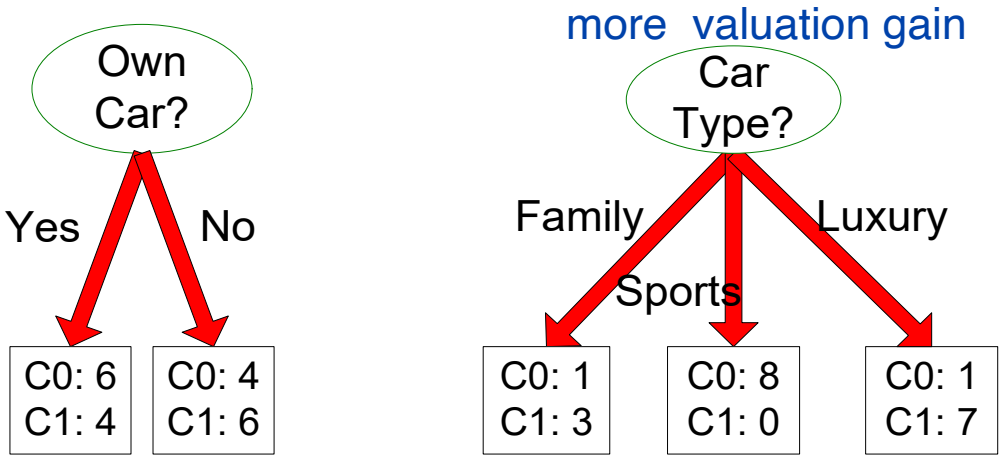
Which test condition is the best?

- Compute the gain of all splits
- Choose the one with largest gain

© University of Melbourne

How to determine the Best Split?

Before Splitting: 10 records of class 0,
10 records of class 1



Own Car: Information gain=0.029

Car type: Information gain=0.62

We should choose Car type as the best split

© University of Melbourne

Creating a decision tree for tax fraud/cheating

- Calculate information gain [Left Child],[Right Child] for each of the following
  - Refund [Yes], [No]
  - Marital status [Single],[Married],[Divorced]
  - Taxable income
    - [60,60], (60,220]
    - [60,70], (70,220]
    - [60,75],(75,220]
    - [60,85],(85,220]
    - [60,90],(90,220]
    - [60,95],(95,220]
    - [60,100],(100,220]
    - [60,120],(120,220]
    - [60,125],(125,220]
- Choose feature+split with the highest information gain and use this as the root node and its split
- Do recursively, terminating when a node consists of only Cheat=No or Cheat=Yes.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

© University of Melbourne

Tree Induction (creation)

- Determine how to split the records
  - How to determine the best split?
  - How to specify the attribute test condition?

- Determine when to stop splitting
  - When node has only a single class of instances

© University of Melbourne

How to Specify Test Condition?

Depends on attribute types

- Nominal
- Ordinal
- Continuous

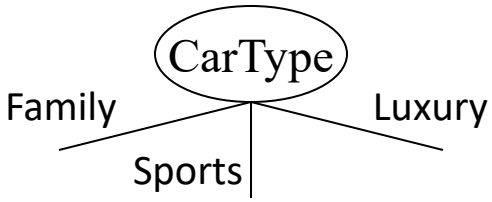
Depends on number of ways to split

- 2-way split
- Multi-way split

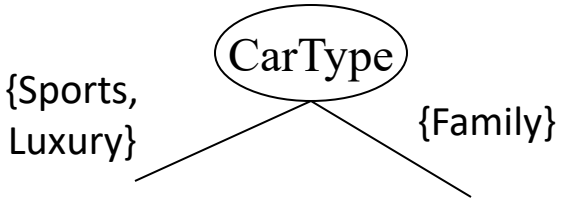
© University of Melbourne

Splitting Based on Nominal Attributes

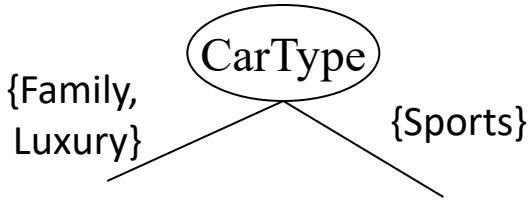
Multi-way split: Use as many partitions as distinct values.



Binary split: Divides values into two subsets.
Need to find optimal partitioning.



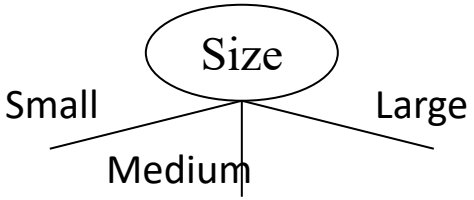
OR



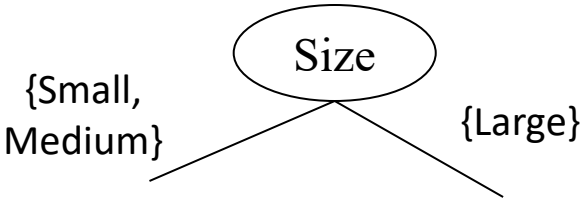
© University of Melbourne

Splitting Based on Ordinal Attributes

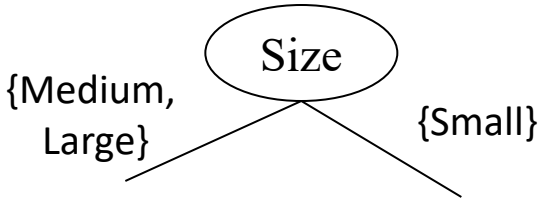
Multi-way split: Use as many partitions as distinct values.



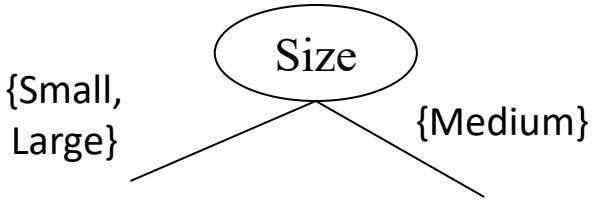
Binary split: Divides values into two subsets.
Need to find optimal partitioning.



OR



What about this split?



© University of Melbourne

Splitting Based on Continuous Attributes

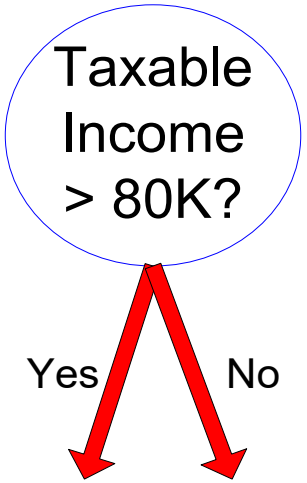
Different ways of handling

- Discretisation to form an ordinal categorical attribute
  - Static – discretize once at the beginning
  - Dynamic – ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering.

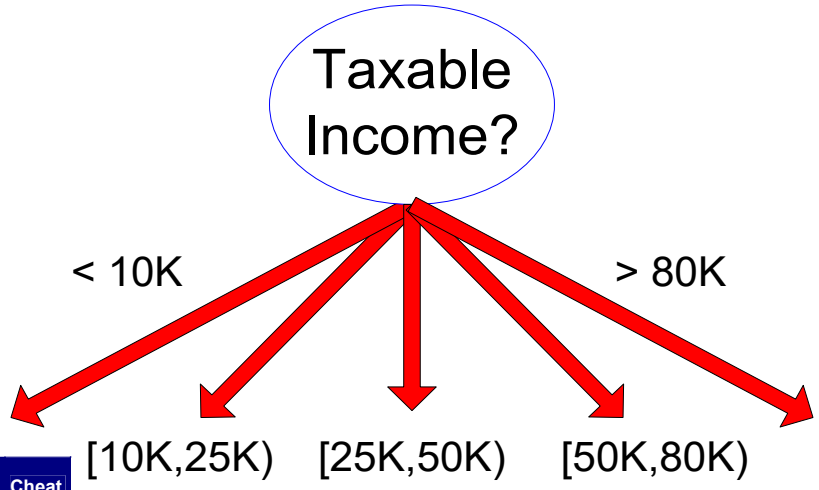
- Binary Decision: $(A < v)$ or $(A \geq v)$
  - consider all possible splits and find the best cut
  - can be more computationally intensive

© University of Melbourne

Splitting Based on Continuous Attributes



(i) Binary split



(ii) Multi-way split

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

© University of Melbourne

Demo of Decision Tree

Hours_Studied	Attendance	Result
High	High	Pass
High	Low	Pass
Low	High	Pass
Low	Low	Fail

Step 1: Entropy before splitting
H(Result) = - ¾ log ¾ + ¼ log ¼ = 0.811
Therefore, this is the entropy of the parent node

Step 2: Try splitting on Hours_Studied
H(Hours_Studied| Result (Parent Node))

H(Hours_Studied = High) = - (2/2 log 2/2 + 0 log 0) = 0
H(Hours_Studied = Low) = - (1/2 log 1/2 + 1/2 log 1/2) = 1

Step 2.1 Weighted entropy after the split:

Sizes: $|S| = 4$, $|S_{\text{High}}| = 2$, $|S_{\text{Low}}| = 2$.

$$H_{\text{children}} = \frac{|S_{\text{High}}|}{|S|} H(S_{\text{High}}) + \frac{|S_{\text{Low}}|}{|S|} H(S_{\text{Low}}) = \frac{2}{4} \cdot 0 + \frac{2}{4} \cdot 1 = \frac{1}{2}$$

Step 2.3 Information Gain:

H(Result) – H(children) = 0.811 – 0.5 = 0.3113

© University of Melbourne

Demo of Decision Tree

Hours_Studied	Attendance	Result
High	High	Pass
High	Low	Pass
Low	High	Pass
Low	Low	Fail

Step 3: Try splitting on Attendance instead
H(Attendance| Result (Parent Node))

H(Attendance = High) = - (2/2 log 2/2 + 0 log 0) = 0
H(Attendance = Low) = - (1/2 log 1/2 + 1/2 log 1/2) = 1

Step 3.1 Weighted entropy after the split:

Sizes: $|S| = 4$, $|S_{\text{High}}| = 2$, $|S_{\text{Low}}| = 2$.

$$H_{\text{children}} = \frac{|S_{\text{High}}|}{|S|} H(S_{\text{High}}) + \frac{|S_{\text{Low}}|}{|S|} H(S_{\text{Low}}) = \frac{2}{4} \cdot 0 + \frac{2}{4} \cdot 1 = \frac{1}{2}$$

Step 3.3 Information Gain:

H(Result) – H(children) = 0.811 – 0.5 = 0.3113

In this toy example, same exact values here

© University of Melbourne

Demo of Decision Tree

Hours_Studied	Attendance	Result
High	High	Pass
High	Low	Pass
Low	High	Pass
Low	Low	Fail

Step 4 conclude and find best split based on information gain:

By Hours_Studied information gain: H(Result) – H(children) = 0.811 – 0.5 = 0.3113

By Attendance information gain: H(Result) – H(children) = 0.811 – 0.5 = 0.3113

We have a tie! ..Now let's see what this looks like in code

© University of Melbourne

Supervised Learning – K nearest neighbor classifier

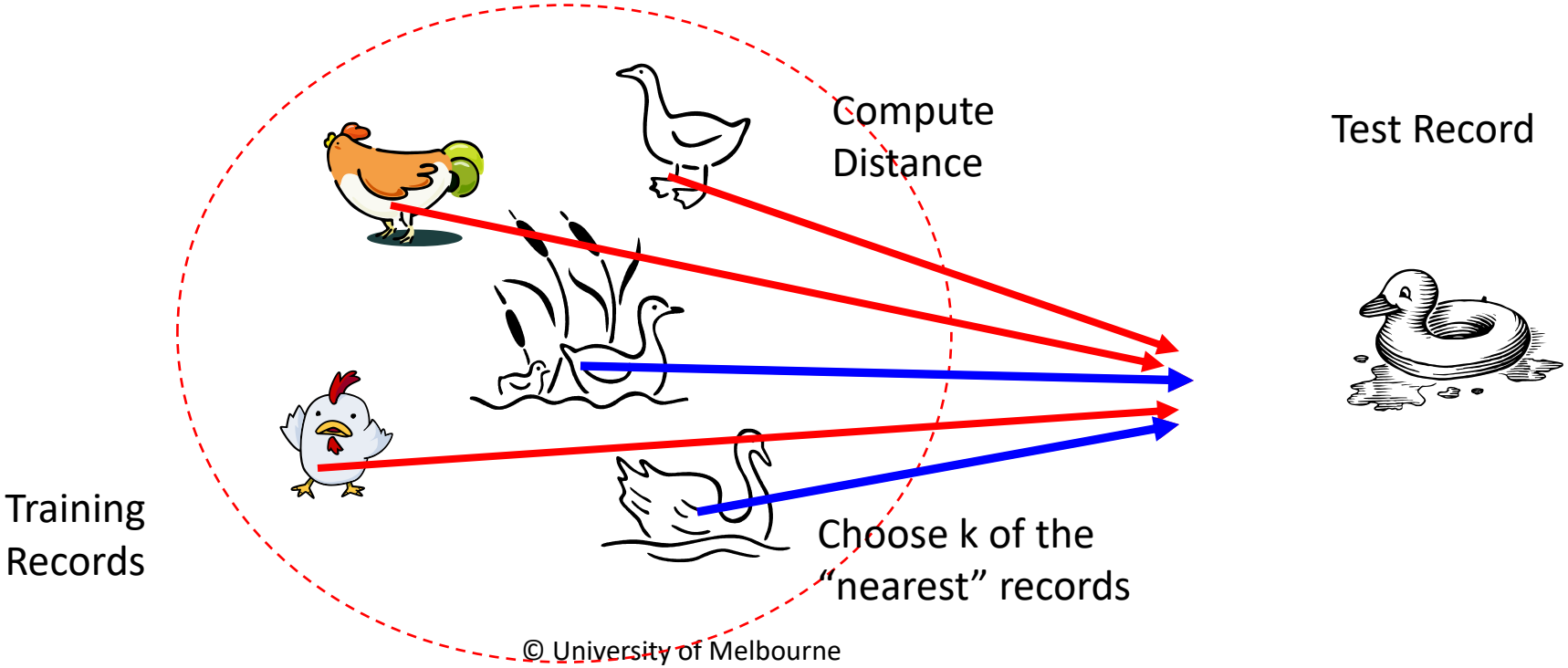
Another widely used and intuitive algorithm for prediction

© University of Melbourne

Nearest Neighbor Classifiers

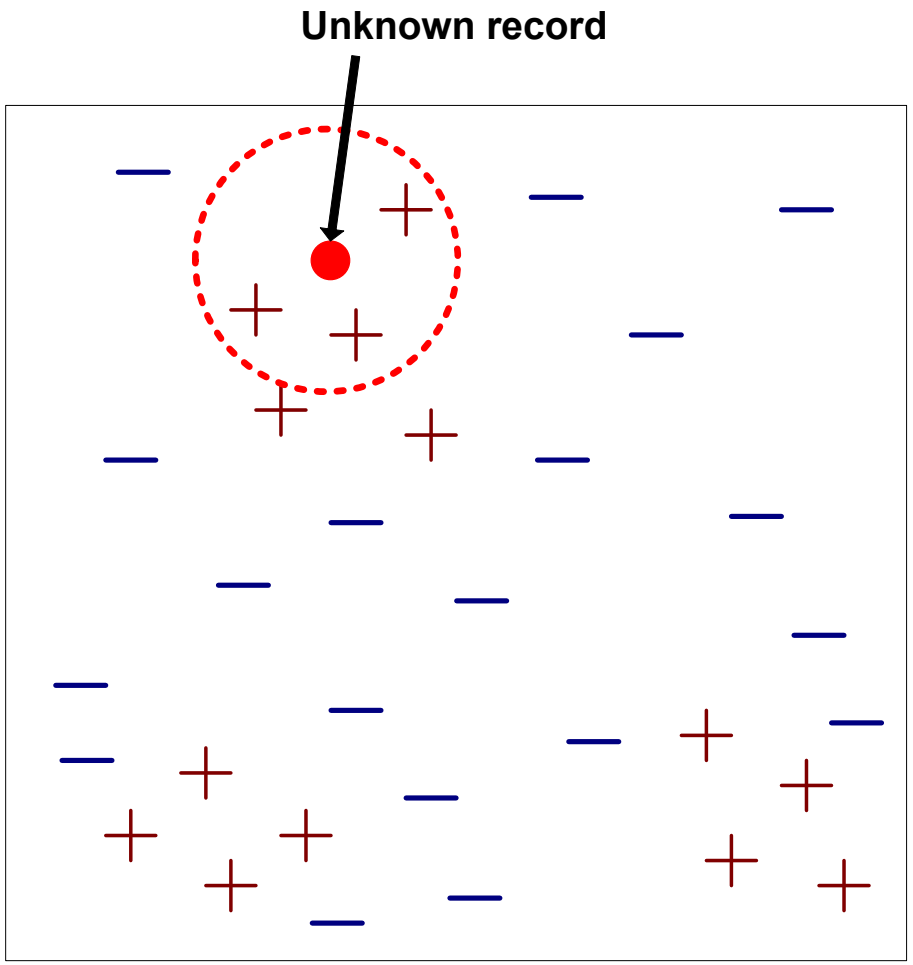
Basic idea:

- *"If it walks like a duck, quacks like a duck, then it's probably a duck"*



© University of Melbourne

Nearest Neighbor Classifiers

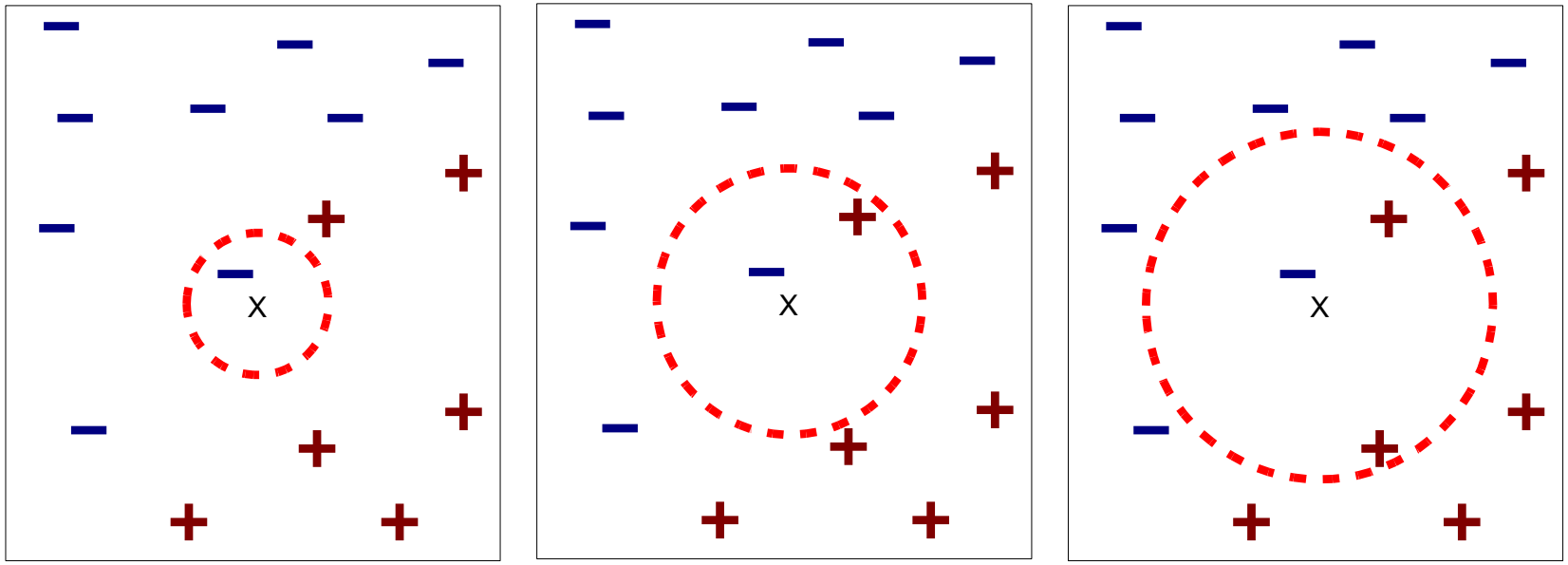


- Requires three things
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve

- To classify an unknown record:
  1. Compute distance to other training records
  2. Identify *k* nearest neighbors
  3. Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

© University of Melbourne

Definition of Nearest Neighbor



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

© University of Melbourne

Distance measure

– Compute distance between two points $p=(p_1,p_2,\ldots)$, $q=(q_1,q_2,\ldots)$

- Euclidean distance

$$d(p,q)=\sqrt{\sum_i (p_i-q_i)^2}$$

- Can also use Pearson coefficient (similarity measure)

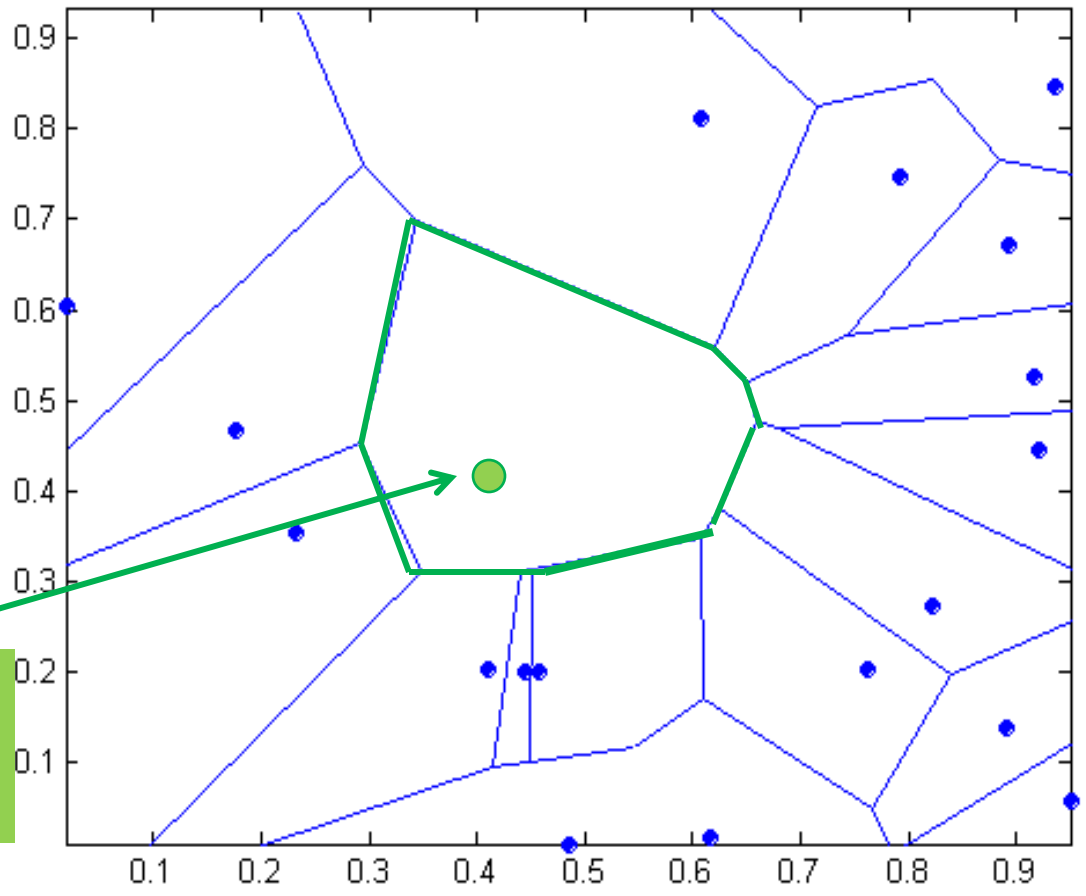
Determine the class from nearest neighbor list:
Once you have the Euclidean distance/similarity measure

- take the majority vote of class labels among the k-nearest neighbors
- Or weight the vote according to distance
  – weight factor, $w = \frac{1}{d^2}$

© University of Melbourne

1 nearest-neighbor

Voronoi Diagram defines the classification boundary

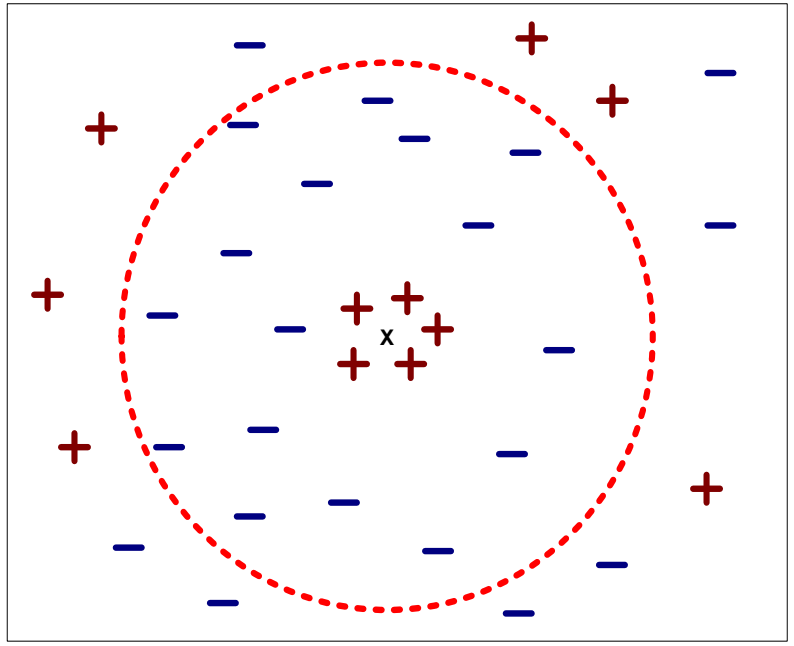


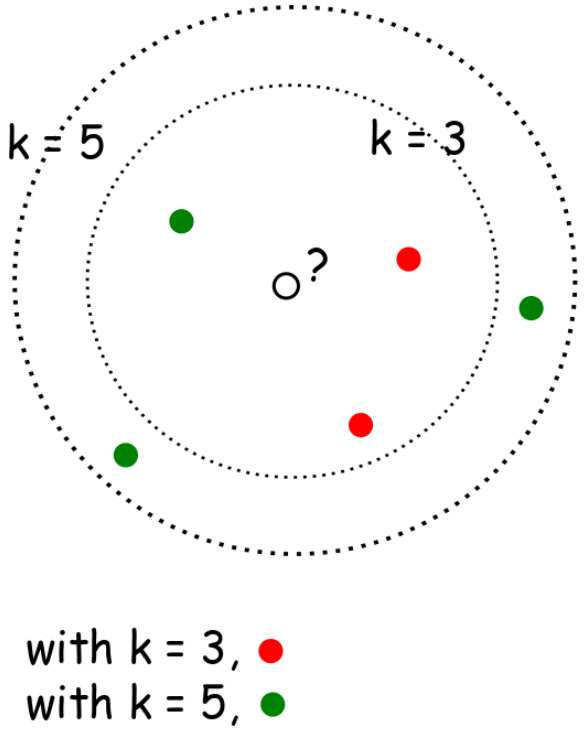
The area takes the class of the green point

K- Nearest Neighbor classifier

Choosing the value of k:

- If k is too small, sensitive to noise points
- If k is too large, neighborhood may include points from other classes





© University of Melbourne

http://2.bp.blogspot.com/-EK4tA2525EM/U-c6Q4jJuwI/AAAAAAAADjw/hdqRXuunpnQ/s1600/knn.png

2024 Sem 1 exam question

6. Consider a classification scenario using a hybrid classification scheme that combines decision trees and k-nearest neighbor (k-NN) classification. At training time, a decision tree is constructed with a maximum depth of three (3), meaning the maximum distance between the root and any leaf is three (3). Consequently, leaf nodes in the tree may contain a mixture of classes. At test time, a classification decision is made by: (i) assigning the test instance to the appropriate leaf node, and (ii) using a k-NN strategy to compare the test instance against all training samples assigned to that leaf node.

   Answer the following questions for this scenario, listing any assumptions made.

   (a) (2 marks) Explain two benefits of this hybrid decision tree+k-NN strategy compared to a pure decision tree classifier.

   (b) (2 marks) Explain two drawbacks of this hybrid decision tree+k-NN strategy compared to a pure decision tree classifier.

© University of Melbourne

Break

© University of Melbourne

Points to remember from this section

Understand what is meant by the terms classification and regression and why it is useful to build models for these tasks

Understand how a decision tree may be used to make predictions about the class of a test instance

Understand the key steps in building a decision tree

- How to split the instances, how to specify the attribute test condition, how to determine the best split and how to decide when to stop splitting

Understand the use of entropy as a node impurity measure for decision tree node splitting. Understand the benefits of entropy for this task and why it is effective for assessing the goodness of a split

Understand the concept of k nearest neighbor classification and in particular the role of the parameter k in determining decision behaviour

© University of Melbourne

References and Acknowledgement

This lecture was adapted from lecture by Prof. James Bailey and using some material adapted from:

https://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf

CS059 - Data Mining -- Slides

http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap4_basic_classification.ppt

© University of Melbourne

Data Science Research

Discussions relevant to Assignment 2

© University of Melbourne

Plan

- Key steps in Data Analysis
- Formulating a Research Question
- Interpreting your results

© University of Melbourne

Key steps in data analysis

	Set Expectations	Collect Information	Revise Expectations
Question	Question is of interest to the audience	Literature Search / Experts	Sharpen question
EDA -Exploratory Data Analysis	Data are appropriate for question	Make exploratory plots of data	Refine question of collect more data
Formal Modelling	Primary model answers	Fit secondary models, sensitivity analysis	Revise formal model to include more predictors
Interpretation	Interpretation of analyses provides a specific and meaningful answer to question	Interpret totality of analyses with focus on effect sizes and uncertainty	Revise EDA and/or models to provide specific and interpretable answer
Communication	Process and results of analysis are understood, complete and meaningful to audience	Seek feedback	Revise analyses or approach to presentation

© University of Melbourne

Good research questions

- Good research questions are:
  - **Interesting**: Someone actually cares about the research outcome
  - **Not already answered**: Not so relevant for the assignment, but in industry you should look through the scientific literature to check
  - **Plausible**: There is some believable framework that could lead to your predicted outcome
  - **Answerable**: Data exists or can be feasibly and ethically collected to answer the question
  - **Specific**: Measurable or investigable; you can investigate it
  - **Clear and simple**: Readers can understand what you are hoping to find out

© University of Melbourne

Types of research question

- Research questions may be:
  - **Descriptive**: Seeks to summarise a characteristic of a set of data, e.g. average number of people living in each suburb. May not be very interesting.
  - **Exploratory**: Identify patterns, trends or relationships between variables. E.g. which dietary factors lead to high blood pressure.
  - **Inferential**: Restate proposed hypothesis as a question and perform analysis with a different dataset
  - **Predictive**: Identify factors that that can predict an outcome. E.g. who is likely to develop cancer within the next 12-months
  - **Causal**: Will changing one factor result in a change in another factor. Difficult to do without control of the study design.
  - **Mechanistic**: *How* does a change in one factor lead to a change in another

© University of Melbourne

Good research questions?

- What was the average temperature in Melbourne in July 2024?
- Did Melbourne's weather affect energy usage in July 2024?
- Is there a relationship between energy usage and rates of diabetes?
- Does the relationship between gaming and mental health observed in undergraduate students generalise to the entire adult population?
- Why am I failing all my subjects?

© University of Melbourne

Interpreting your data

- Before starting your analysis, think about whether the data can lead to an interpretable result.
- For example consider you have the following data:
  - People taking a vitamin tablet each day
  - People taking Panadol each day
- Is there any way you could make a conclusion about the effects vitamin tablets have on headaches from this data?
- Consider confounding factors
- Beware of making causal statements without the data to support them
- Consider possible biases/limitations in your data

© University of Melbourne

Exploratory data analysis steps

1. Formulate your question
2. Read in your data
3. Check the packaging
4. Look at the top and the bottom of your data
5. Check your “n”s
6. Validate with at least one external data source
7. Make a plot
8. Try the easy solution first
9. Follow up

© University of Melbourne

Key tasks to consider

- **Pre-processing:** What pre-processing may be required to correct data quality and/or representation issues? How might different datasets be linked together and what problems may arise? Is feature selection required and how might it be best performed?
- **Exploratory data analysis**: How can you quickly gain & communicate an understanding of the datasets? What stats and visualisations might help?
- **Modelling**: What techniques (e.g. correlation, clustering, classification, regression) are applicable to your research question & the data? Within a technique, which algorithm(s) might you use?

© University of Melbourne

Communication

Some key concepts to consider:

1. **Audience**: Know your audience and when you have control over who the audience is, select the right audience for the kind of feedback you are looking for.
2. **Content**: Be focused and concise, but provide sufficient information for the audience to understand the information you are presenting and question(s) you are asking.
3. **Style**: Avoid jargon. Unless you are communicating about a focused highly technical issue to a highly technical audience, it is best to use language and figures and tables that can be understood by a more general audience.
4. **Attitude**: When communicating with your teammates & colleagues, adopt a collaborative attitude and engage in a dialogue. Your goal is not to "defend" your question or work, but rather to get their input so that you can do your best work.

© University of Melbourne

References and recommended reading

- Much of this material has been drawn from the following textbook:

Peng, R. D., & Matsui, E. (2016). *The Art of Data Science: A guide for anyone who works with Data.*

- Freely available at https://bookdown.org/rdpeng/artofdatascience/, I encourage you to browse through it when starting your second assignment

© University of Melbourne