COMP20008 Elements of Data Processing

Classification Methodologies
continued

Announcements

- Project marking
  - We expect to release marks + feedback for Phase 1 by Thursday 13th April
- Phase 2A (Concept formulation and preliminary investigation): Due 25th April
  - In workshops this week (10-13th April) – Half of the time will be devoted to discussion regarding Phase 2A of the project
- Phase 2B (peer feedback): Due 28th April
  - In lecture on Monday 24th April, we will discuss strategies for giving peer feedback

Plan today

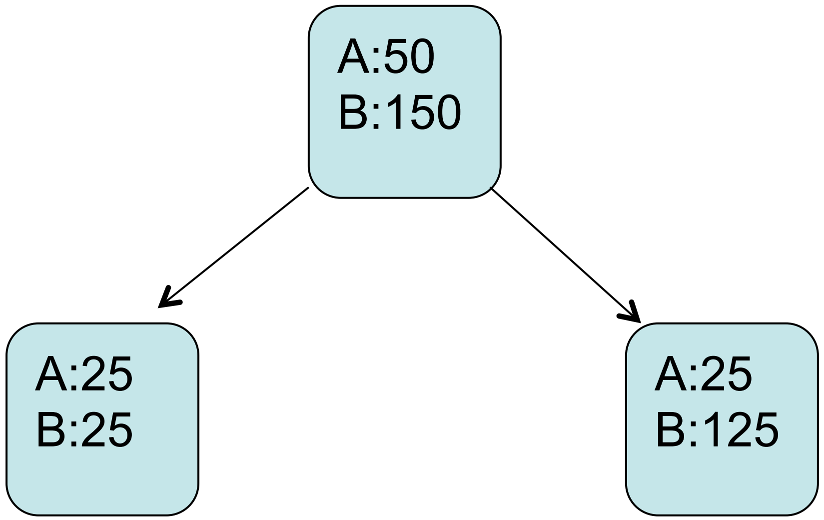
- Classification
  - Decision tree classification – wrap up from last week
  - k nearest neighbor classification
- Recap – where are we in the subject?
- Advice for Phase 2A of project

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

Question 2c) from 2016 exam

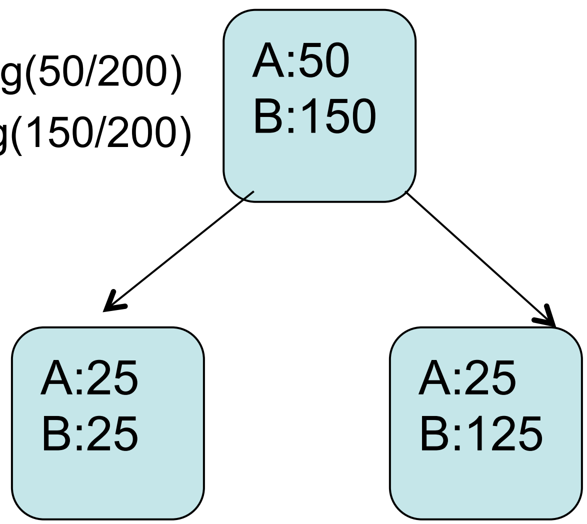
Given a dataset with two classes, A and B, suppose the root node of a decision tree has 50 instances of class A and 150 instances of class B. Consider a candidate split of this root node into two children, the first with (25 class A and 25 class B), the second with (25 class A and 125 class B). Write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



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)

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

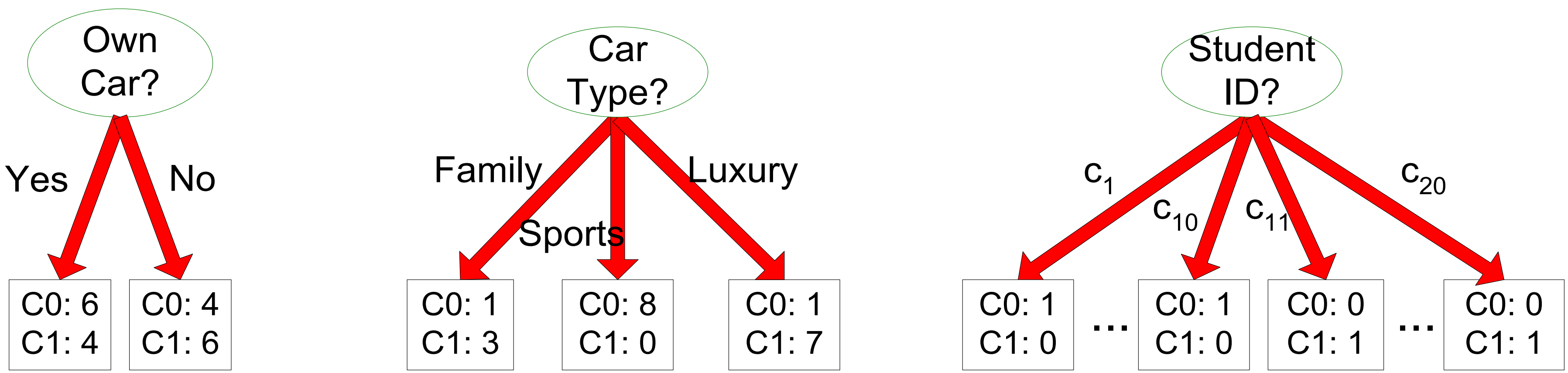
Split utility= Information Gain

=Entropy(root) – Entropy(root|split)

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

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

Student ID: Information gain=1

We should choose Student ID as the best split???!!!

Creating a decision tree

- Calc 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

Decision tree: advantages and disadvantages

- Advantages
  - Easy to interpret
  - Relatively efficient to construct
  - Fast for making a decision about a test instance

- Disadvantages
  - A simple greedy construction strategy, producing a set of ``If ..then'' rules
    - sometimes this is too simple for data with complex structure
  - May behave strangely for some types of features (E.g. student ID feature from earlier slide)

Decision tree classifier: training and testing

- Divide training data into:
  - Training set (e.g. 2/3)
  - Test set (e.g. 1/3)
- Learn decision tree using the training set
- Evaluate performance of decision tree on the test set

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

Metrics for Performance Evaluation

- Can be summarized in a Confusion Matrix (contingency table)
  - Actual class: {yes, no, yes, yes, …}
  - Predicted class: {no, yes, yes, no…}

	PREDICTED CLASS	
	Class=Yes	Class=No
ACTUAL CLASS Class=Yes	a	b
ACTUAL CLASS Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metrics for Performance Evaluation

	PREDICTED CLASS	
ACTUAL CLASS	Class=Yes	Class=No
Class=Yes	a (TP)	b (FN)
Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

– Actual class: {yes, no, yes, yes, no, yes, no, no}
– Predicted: {no, yes, yes, no, yes, no, no, yes}

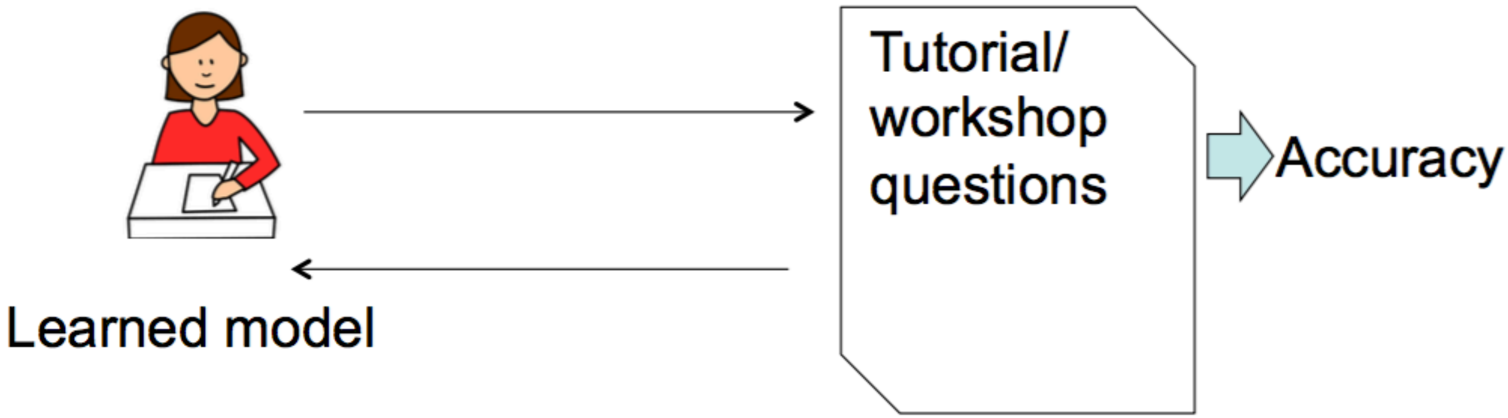
	PREDICTED CLASS	
	Class=Yes	Class=No
ACTUAL CLASS Class=Yes	a= 1 (TP)	b=3 (FN)
ACTUAL CLASS Class=No	c=3 (FP)	d=1 (TN)

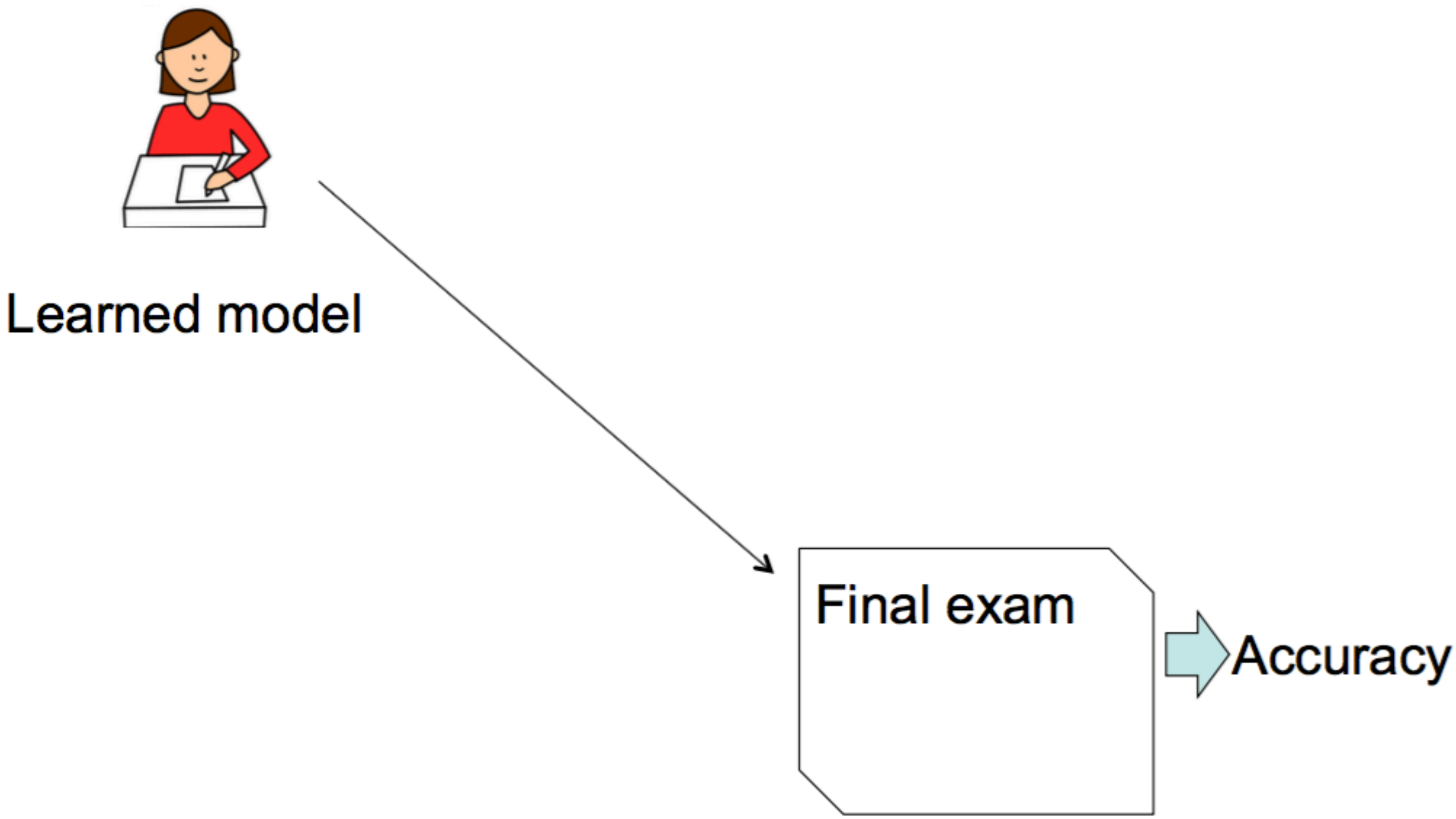
Limitations of accuracy

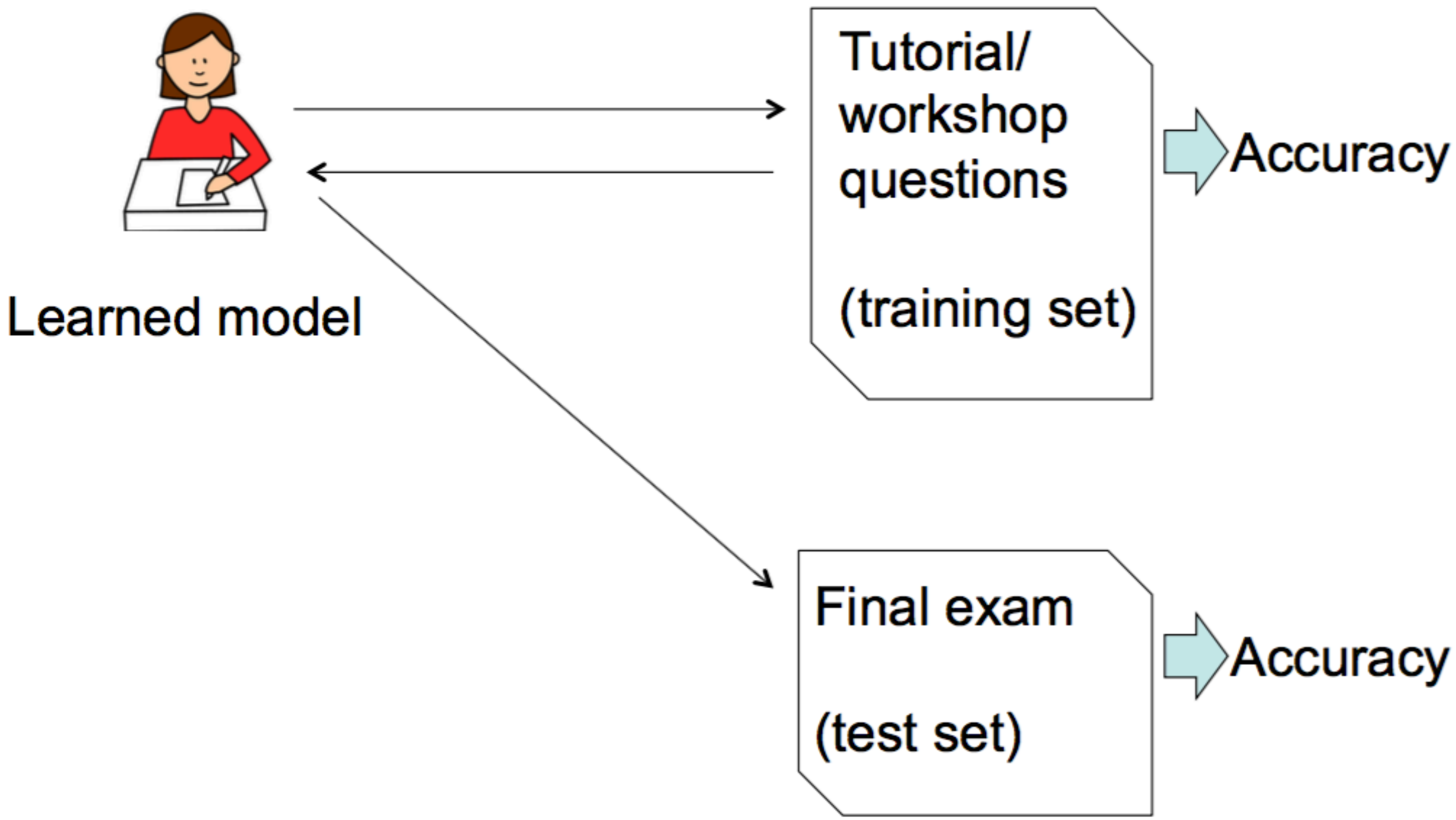
- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- • If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading here because model does not detect any class 1 example
  - Other metrics can be used instead of accuracy, that address this problem (but we won't cover these)

Why do we split the dataset into training and testing for evaluating accuracy?




Learned model
Final exam
Accuracy

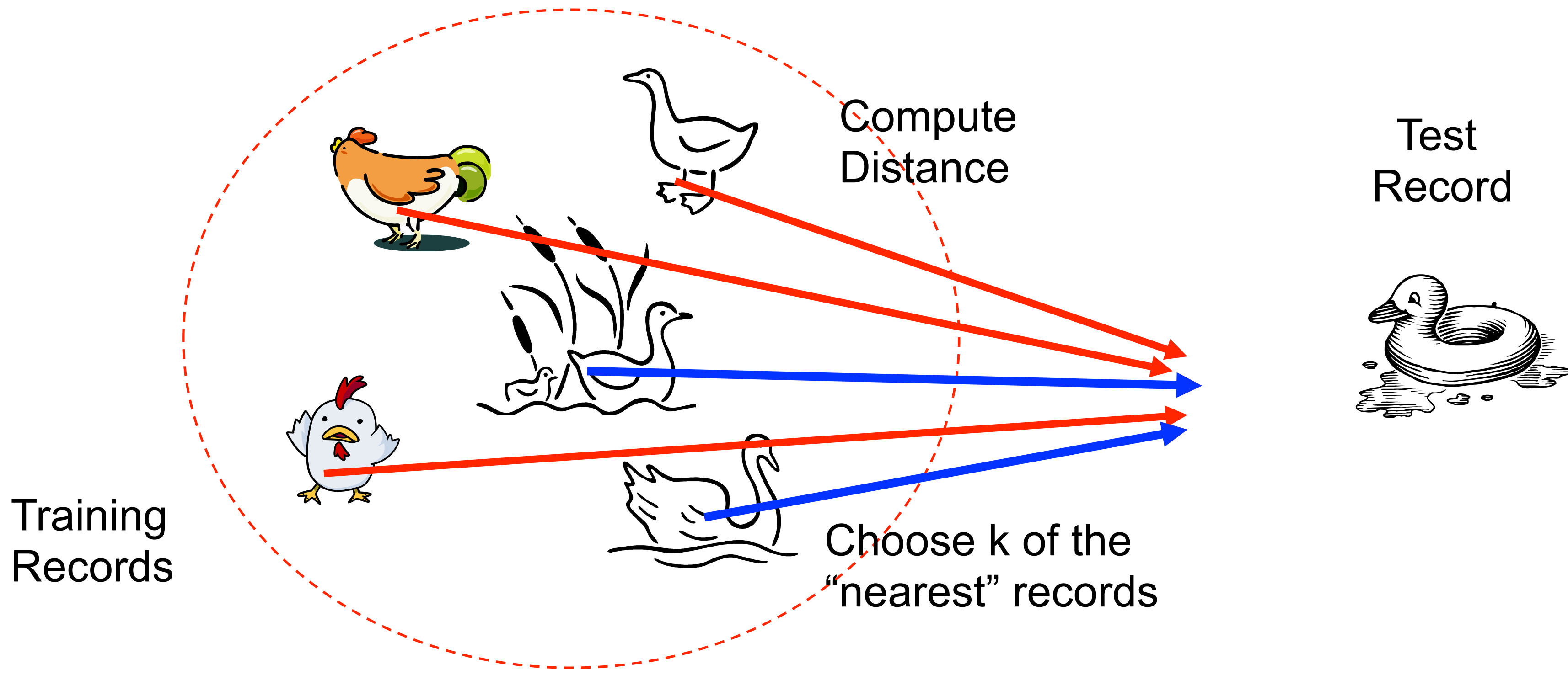

Learned model
Tutorial/
workshop
questions
(training set)
Accuracy
Final exam
(test set)
Accuracy

K nearest neighbor classifier

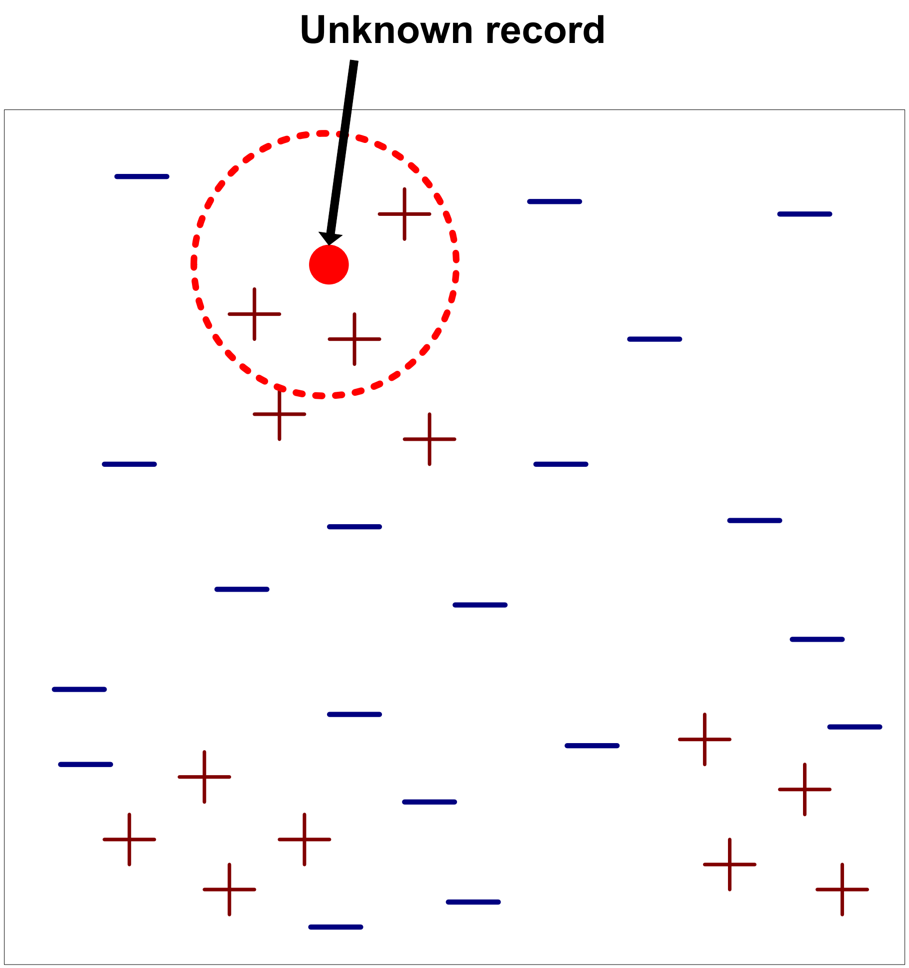
- Another widely used and intuitive algorithm for prediction

Nearest Neighbor Classifiers

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



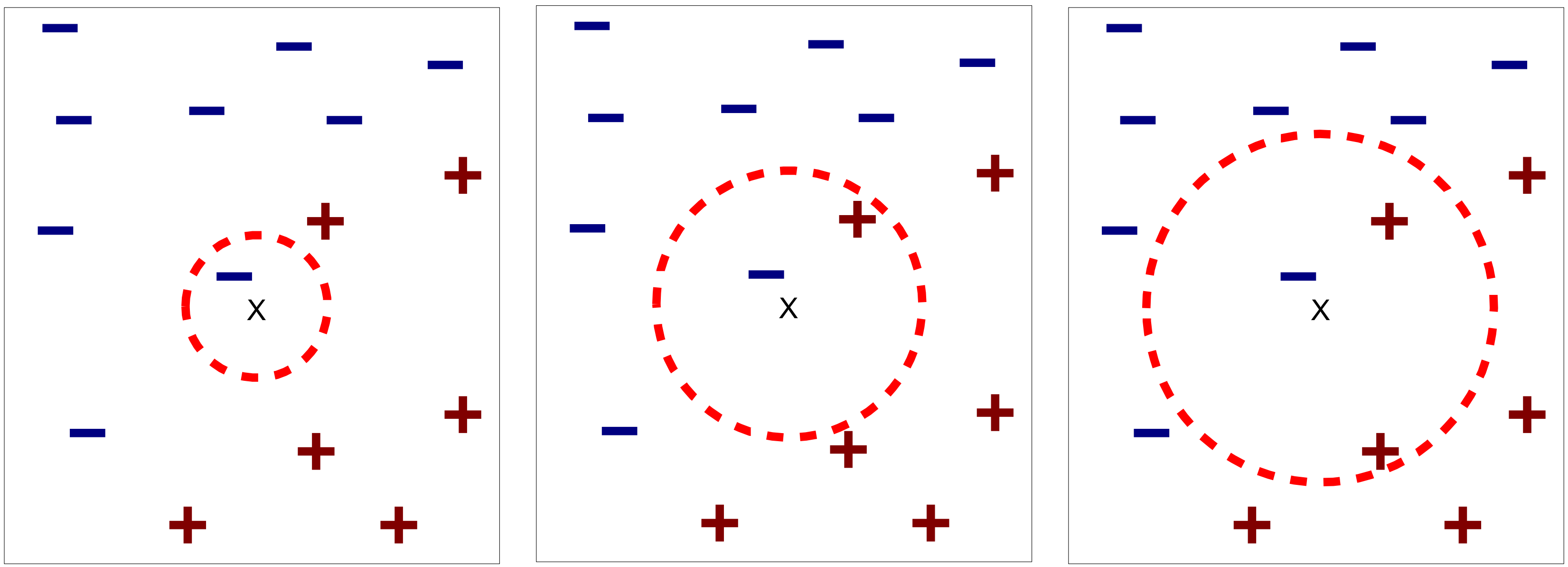
Nearest-Neighbor Classifiers



- 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)

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

Distance measure

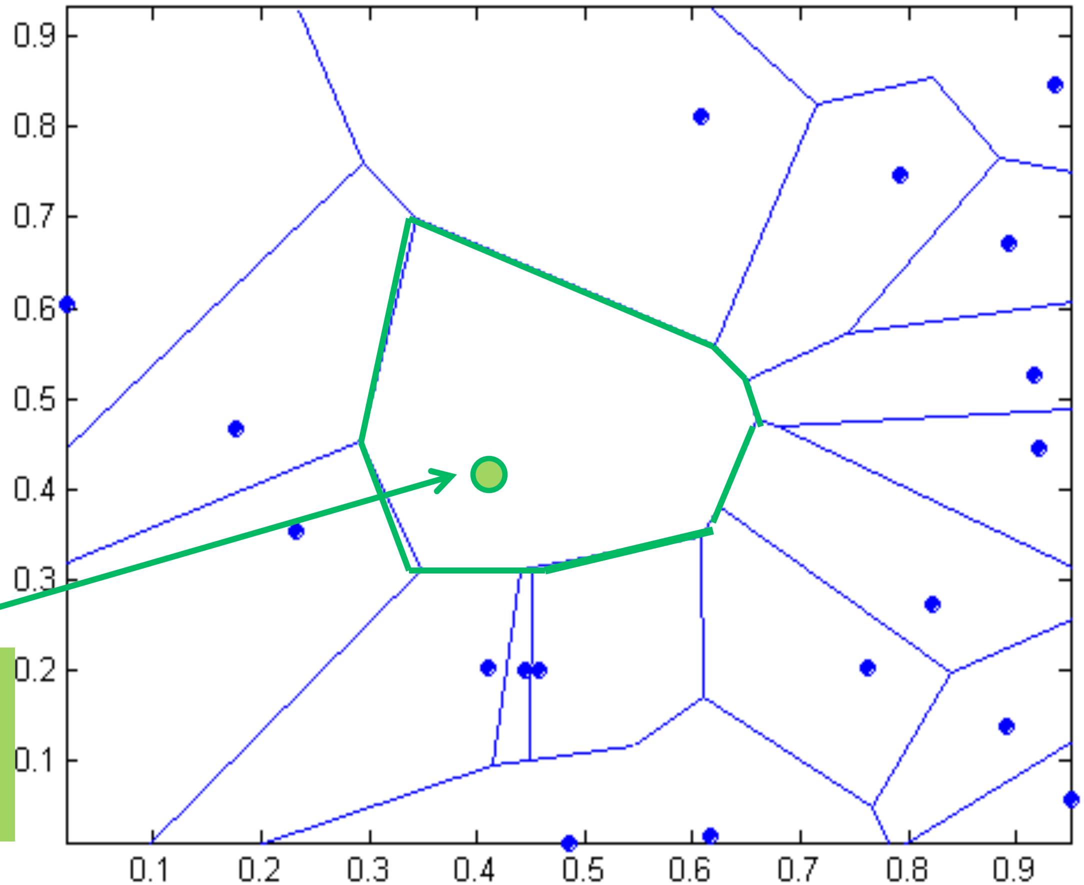
- Compute distance between two points:
  - Euclidean distance

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

  - Pearson coefficient (similarity measure)
- Determine the class from nearest neighbor list
  - take the majority vote of class labels among the k-nearest neighbors
  - Weigh the vote according to distance
    - weight factor, $w = 1/d^2$

1 nearest-neighbor

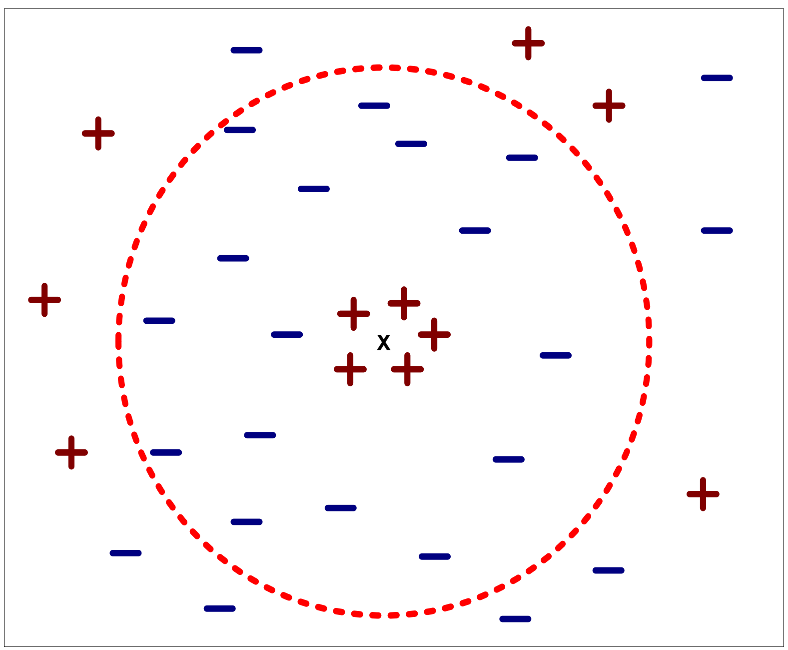
Voronoi Diagram defines the classification boundary

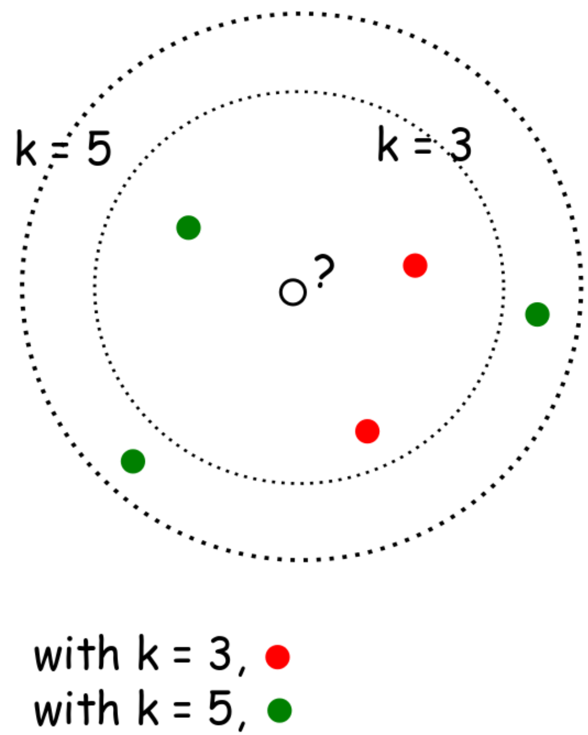


The area takes the class of the green point

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





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

K-NN classifier: advantages and disadvantages

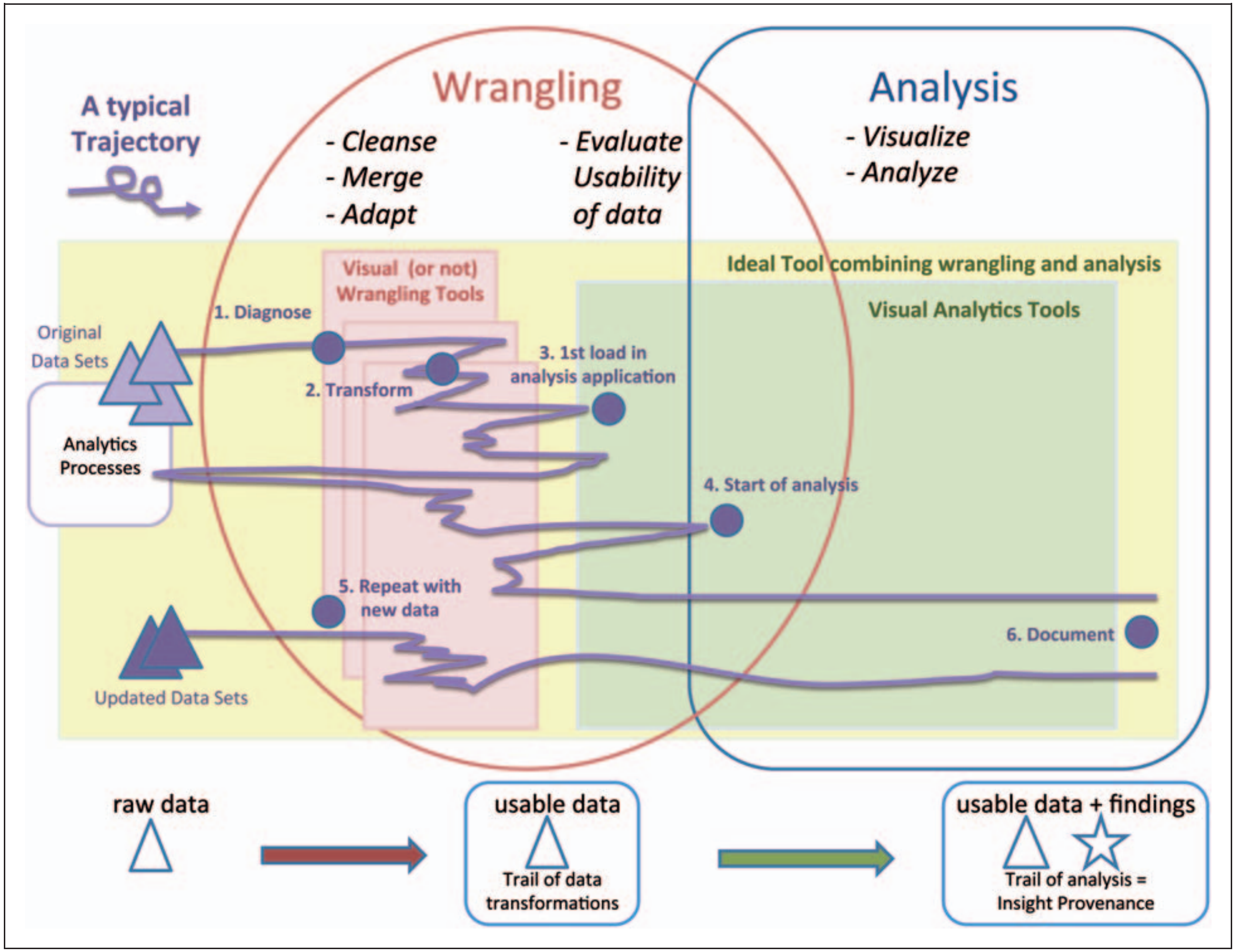
- Advantages
  - Intuitive way of making decisions
  - Can handle datasets with complex structure

- Disadvantages
  - Need to store training data in order to make the prediction
  - Classification may be slow for large datasets (Storage and neighbor computations)
  - Choices need to be made about parameters
    - What distance function to use?
    - What value of k to use?

Points to remember from this lecture

- Understand the use of accuracy as a metric for measuring the performance of a classification method.
- Understand how TP,TN,FP and FN are used in the accuracy calculation. The formula for accuracy will be provided on the exam
- understand the operation and rationale of the k nearest neighbor algorithm for classification
- understand the advantages and disadvantages of using k nearest neighbor or decision tree for classification

Where are we now?



THE UNIVERSITY OF MELBOURNE

Where have we got to?

- *Preprocessing (4 lectures): Weeks 1-3*
  - *Data types and processing, data cleaning including outliers, missing data*
- *Visualisation (3 lectures): Weeks 3-4*
  - *Plotting and visualisation methods, clustering, dimensionality reduction*
- *Analysis (4 lectures): Weeks 5-7*
  - *Correlations, basic prediction techniques*
- **Infrastructure and Distributed (5 lectures): Weeks 8-10**
  - **noSQL and cloud, data linkage and integration, blockchain**
- **Social, ethical and privacy issues (3 lectures): Weeks 11-12**
  - **K-anonymity, l-diversity, location privacy, ethics**

Phase 2A – Due 25th April

- Need to formulate a question, identify 2 open datasets to help answer the question and conduct some initial wrangling
- Approximately 13 hours work. A possible breakdown
  - (5 hours) Browse open datasets, select two and formulate question
  - (6 hours) Initial wrangling
  - (2 hours) Write report

References and Acknowledgement

This lecture was prepared 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