COMP307 Week 2 (Tutorial)

1. Announcements
   - Assignment 1
     - When to start?
     - What to do now?
   - Helpdesk sessions
2. Nature of this course
   - Uncertainty:
     - Uncertain (AI)
     - Certain (other courses)
   - Solutions/answers:
     - Unique?
     - Best?
     - Good?
   - Algorithms/methods:
     - Step-by-step => the correct version (others)
     - Main ideas => many different versions (AI)
   - Details vs guidelines
3. Search (Lectures 2 & 3)
   - Search strategies
     - Uninformed
     - Informed
     - Beyond classic search
   - Classic vs. local search
   - (Genetic) Beam search
4. Machine Learning (Lecture 4)
   - Type of learning systems
     - Supervised learning
     - Unsupervised learning
   - Machine learning tasks
     - Classification
     - Regression
     - Clustering
   - Line fitting
     - Over-fitting

COMP 307/AIML 420 — *Introduction to* AI

Assignment 1: Basic Machine Learning Algorithms

15% of Final Mark — Due: 11:59pm Sunday 28 March 2021

1 Objectives

The goal of this assignment is to help you understand the basic concepts and algorithms of machine learning, write computer programs to implement these algorithms, use these algorithms to perform classification tasks, and analyse the results to draw some conclusions. In particular, you should be familiar with the following topics:

- Machine learning concepts,
- Machine learning common tasks, paradigms and methods/algorithms,
- Nearest neighbour method for classification,
- Decision tree learning method for classification,
- Perceptron/linear threshold unit for classification, and
- k-means method for clustering and k-fold cross validation for experiments.

These topics are (to be) covered in lectures 4–7. The textbook and online materials can also be checked.

2 Question Description

Part 1: k-Nearest Neighbour Method (30 Marks for COMP307, and 37 Marks for AIML420)

In this part you will implement the k-Nearest Neighbour method, and evaluate it on the *wine* data set described below. Additional questions on k-means and k-fold cross validation need to be answered/discussed.

Problem Description

The *wine* data set is taken from the UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets/wine`). The data set contains 178 instances in 3 classes, having 59, 71 and 48 instances, respectively. Each instance has 13 attributes: *Alcohol, Malic_acid, Ash, Alcalinity_of_ash, Magnesium, Total_phenols, Flavanoids, Nonflavanoid_phenols, Proanthocyanins, Color_intensity, Hue, OD280/OD315_of_diluted_wines*, and *Proline*. We have split the dataset into two subsets: one for training and the other for testing.

Requirements

Your program should classify each instance in the test set `wine-test` according to the training set `wine-training`. Note that the *final* column in these files list the class label for each instance.

Your program should take two file names as command line arguments, and classify each instance in the test set (the second file name) according to the training set (the first file name).

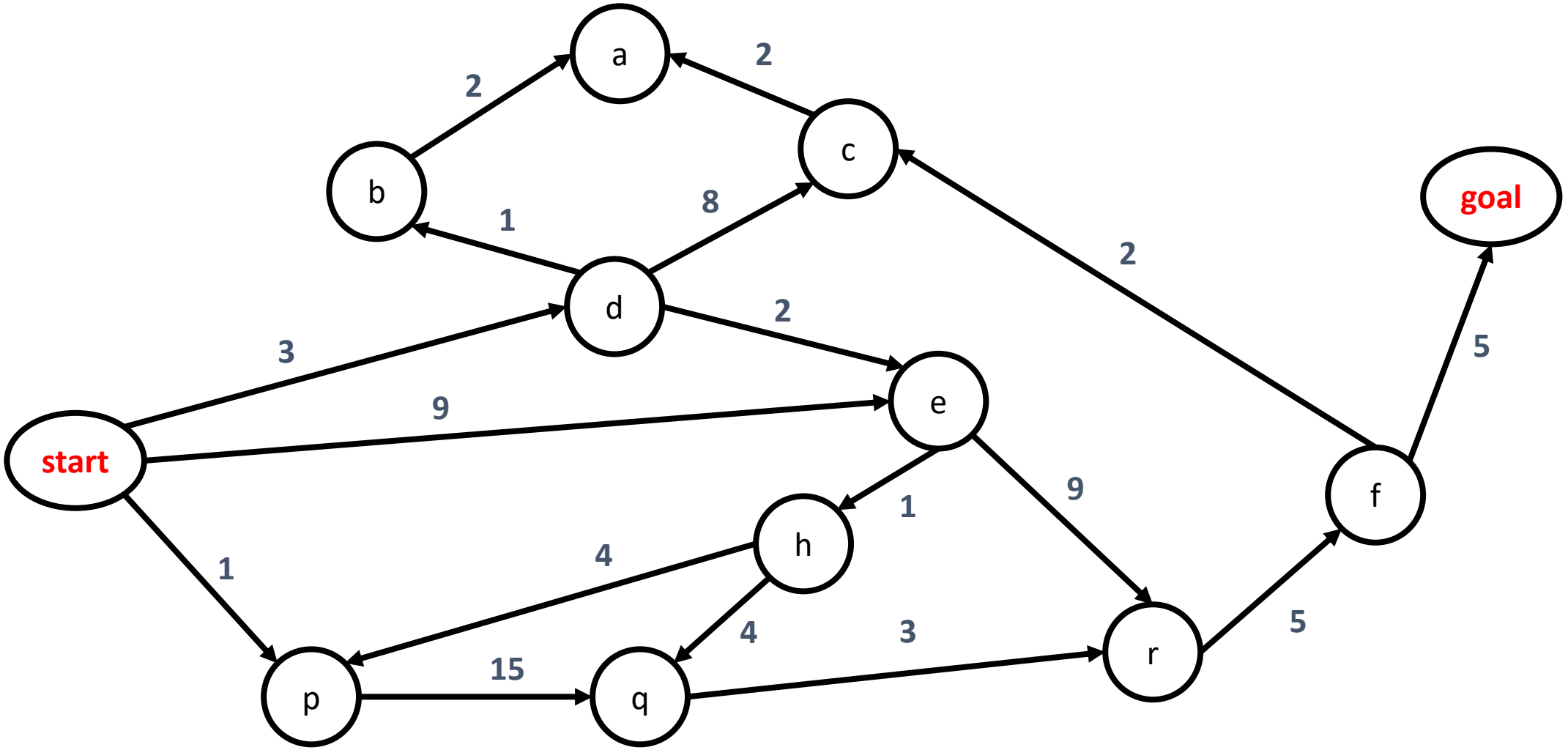
You may write the program code in `Java, C/C++, Python`, or any other programming language, **as long as it can be easily run on the ECS systems**.

You should submit the following files electronically:

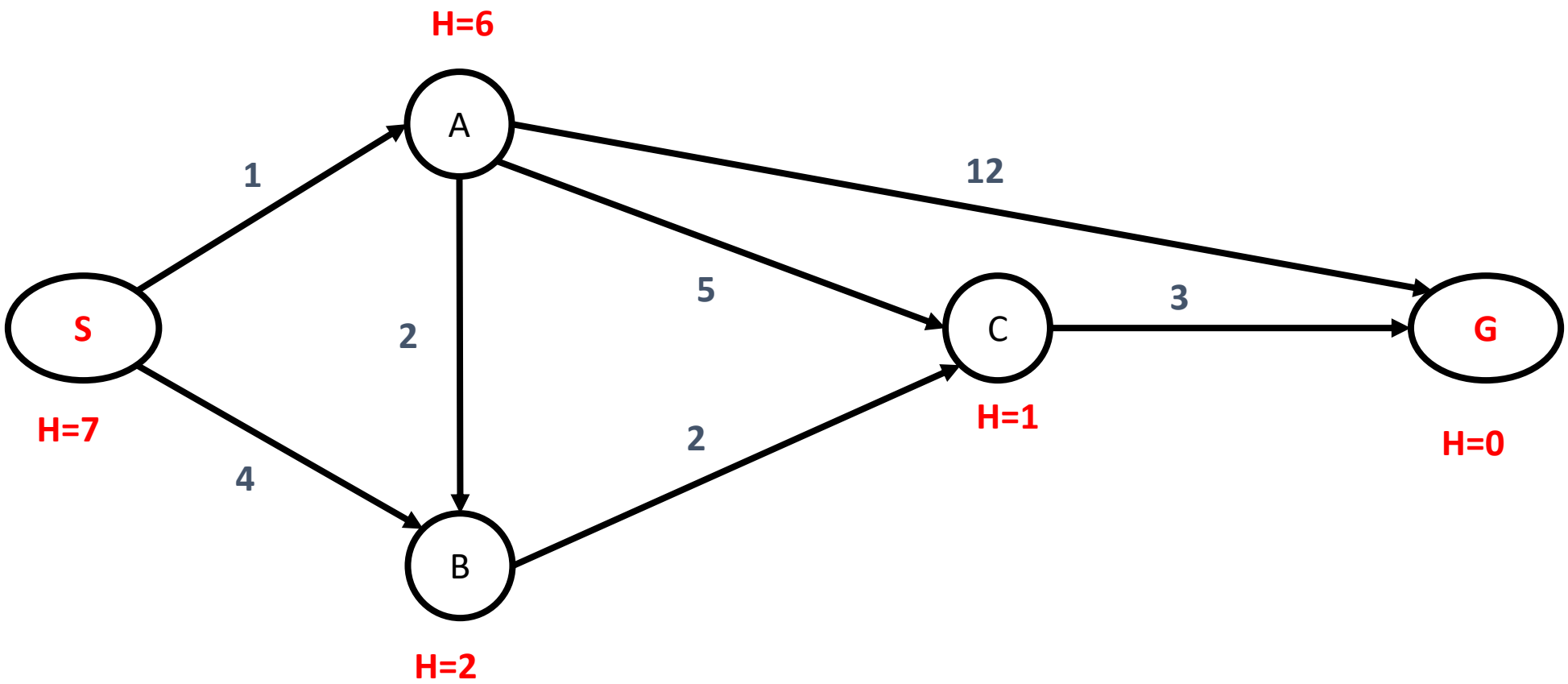
- (15 marks) `Program code` for your k-Nearest Neighbour Classifier (the source code as well as the executable program that runs on the ECS School machines[1]).
- `readme.txt` which describes how to run your program.

[1] If you use Python or another interpreted language, the source and executable programs are the same!

Uniform-Cost Search



A* Search (1)



Classic vs. Local Search

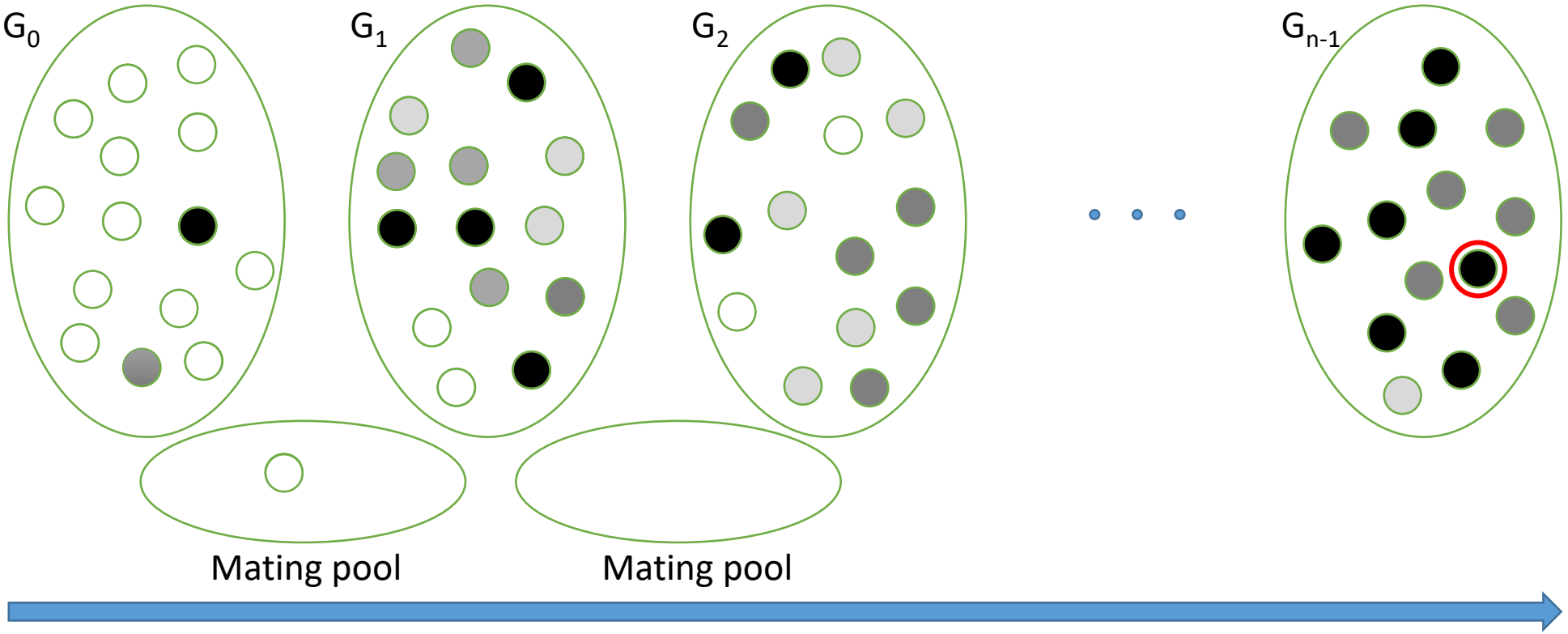
Classic Search

- Assumptions:
  - Observable
  - Deterministic
  - known environment where the solution is a sequence of actions
- The path is a solution

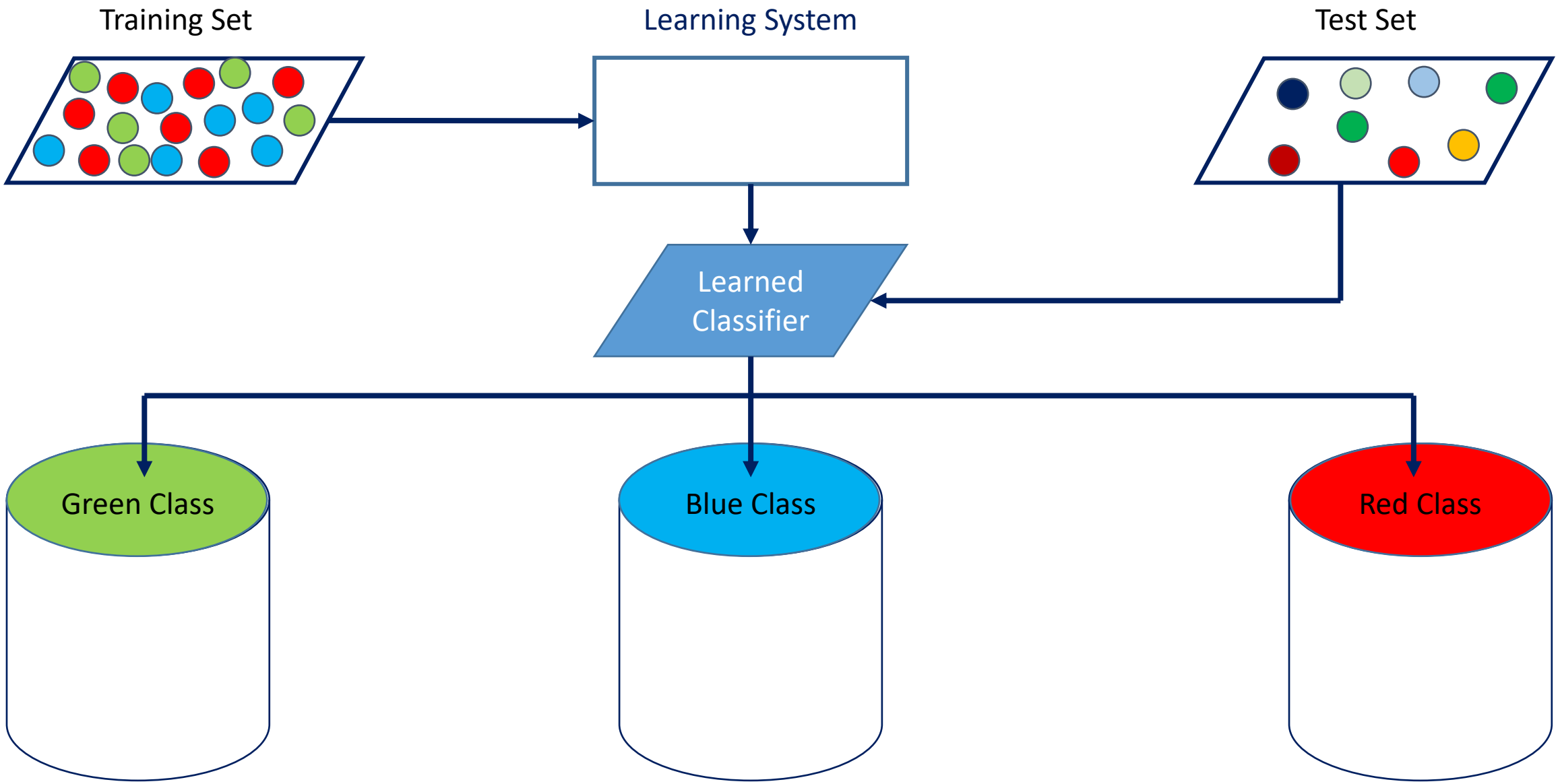
Local Search

- Useful for solving optimisation problems
- Aim to find the best state according to an objective function
- Only keep one state and its evaluation
- Choose the best successor

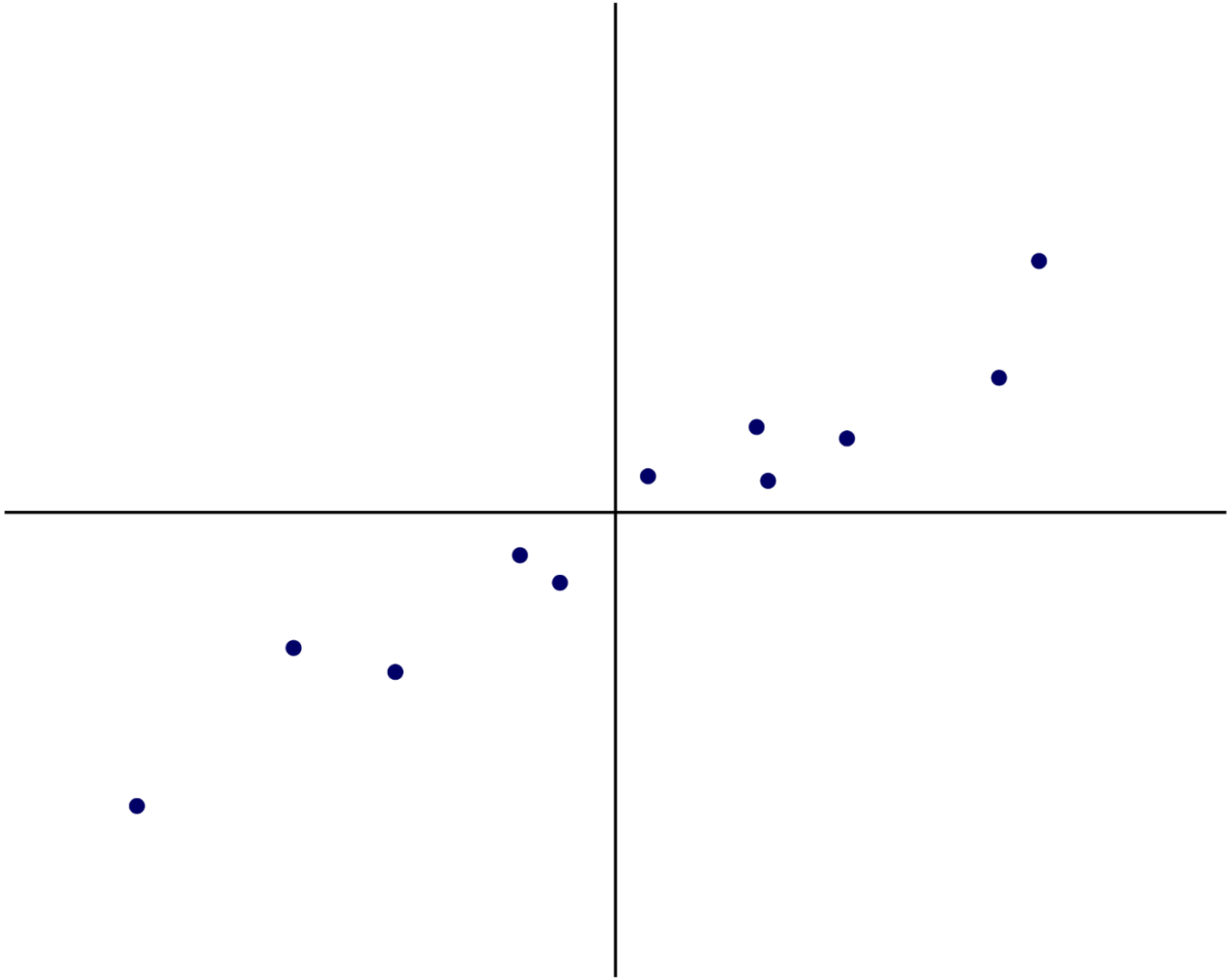
(Genetic) Beam Search



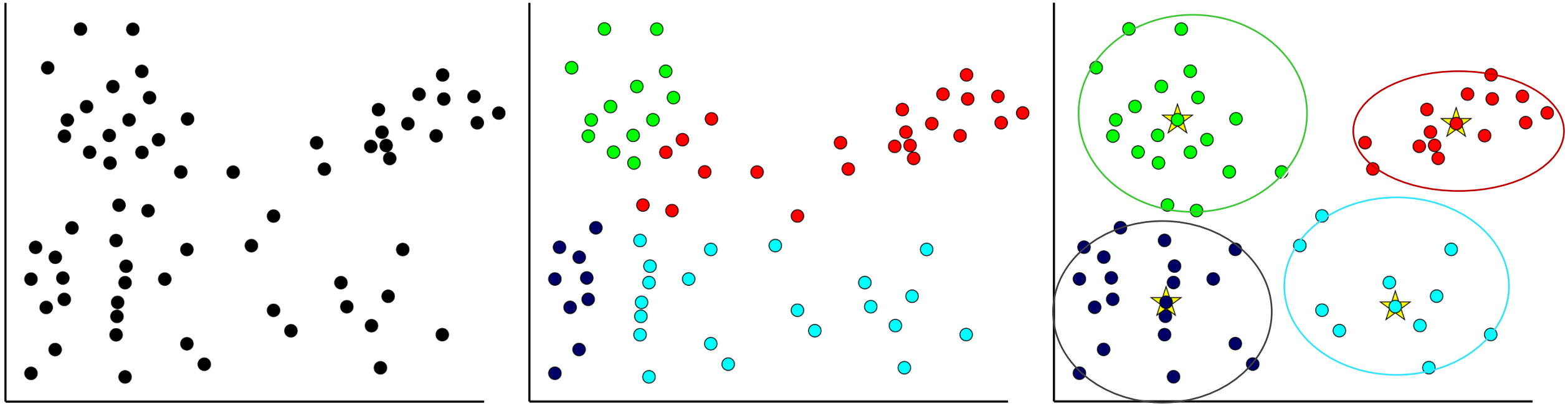
Supervised Learning (Classification)



Supervised Learning (Regression)



Unsupervised Learning (Clustering)



Over-fitting