UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

IN3050/4050 — Introduction to Artificial Intelligence and Machine Learning
Trial Exam Spring 2020
Exam content: Around 2 working days
Deadline: June 2 at 2:30 PM
Permitted materials: All

This is a trial exam – the instructions for the regular exam will be as below. Note that you do not need to follow those instructions for this trial exam - in particular, you should not try to submit your answers through Inspera. Suggested solutions for this exam will be provided around 1 week after its release. Due to the short preparation time, we are focusing our resources on quality-checking the main exam. Therefore, please let us know if you find any potential mistakes or problems with the questions below. Send them to in3050-support@ifi.uio.no.

Instructions for the exam:

- **General information about home exams in Inspera can be found here.**
- **For the main exam, the answers should be delivered as a single PDF file through Inspera. Info on how to deliver files in Inspera can be found here.**
- **You are free to make plots/figures in any program you want, and add them to your delivered PDF. That includes using drawing tools on the computer, and drawing figures by hand and taking a picture of them. Some tips are available here.**
- **Your delivery should be anonymous. Do not write your name.**
- **If you have questions related to this exam, please submit them here.**
- **Please check the "Messages" in the Course Website daily. We will post any clarifications around the content of this exam there.**
- **You are not required to write any code for any of these tasks – all calculations may be done by hand, and we do want you to show each step of your calculation.**
- **The tasks of the exam are given points summing to 100, giving you an indication of how much each task is weighted and roughly how much time you should spend on it.**

1) Simulated Annealing (3p)

a) What is the role of the temperature parameter in simulated annealing? (1p)
b) What would happen if we start with a very low temperature (keeping low through search)? Which search algorithm would this be similar to? (1p)
c) What would happen if we start with a very high temperature and never decrease? (1p)

2) Evolutionary Algorithm: Minimizing f(x) (10p)

In lecture 3, slide 32, we saw an example of an EA done by hand for a full generation (also found in Eiben & Smith, page 35). The simple problem was to optimize the value of the function $f(x)=x^2$ for integers in the range 0-31. Now, you are going to do the same, but for a minimization problem. That is, minimize the function $f(x) = x^2$ in the range 0-31, using the same genotype, phenotype, mutation and crossover as the example from the lecture.

a) Suggest a way to change the fitness function to make this a minimization problem. The original example used the following selection function:

$$p_i = f(i) / \sum_{j \in P} f(j)$$

Will you need to change this, or are the changes to the fitness function sufficient? Briefly explain why/why not? (2p)

Perform a full round of the EA on this minimization problem by filling in the tables below (don't fill in cells marked with a "-"). Note that the individuals in the population are different from those in the example from the lecture. Calculating the x-value for an individual is done with straightforward binary-to-decimal decoding. That is, each digit in the genotype is multiplied by 2^i, where i indicates the position in the genotype from right to left, starting with 0: [g_4, g_3, g_2, g_1, g_0]. We have filled in two phenotype values to help you get started.

b) Parent selection:

When calculating the actual count, instead of randomly sampling, you can just *round the expected count to the nearest integer* (if that gives you too few or too many individuals, try to select fairly according to the expected count). (2p)

String no.	*Initial population*	*x value*	*Fitness f(x) you defined above*	*Prob$_i$*	*Expected Count*	*Actual count*
1	0 1 1 0 1	13				
2	0 0 1 1 1	7				
3	1 1 1 0 0					
4	1 0 0 1 1					
Sum	-	-				
Average	-	-				
Best	-	-				

c) Crossover:

Insert individuals into the mating pool in the same order they are listed above, according to the actual count. Use the indicated crossover points to cross over individual pairs (1,2) and (3,4). (1p)

String no.	*Mating pool*	*Crossover point*	*Offspring after crossover*	*x Value*	*Fitness f(x)*
1		3			
2		3			
3		1			
4		1			
Sum	-	-	-	-	
Average	-	-	-	-	
Best	-	-	-	-	

d) Mutation:
Now, mutate the offspring. Normally, we would do that randomly, but instead we will tell you how to mutate this time. Perform the normal binary mutation, on the following genes:

-Gene 1 in individual 1
-Gene 4 in individual 2
-Gene 2 in individual 4

(1p)

String no.	*Offspring after crossover*	*Offspring after mutation*	*x Value*	*Fitness f(x)*
1				
2				
3				
4				
Sum	-	-	-	
Average	-	-	-	
Best	-	-	-	

e) What is the best resulting x value and associated fitness value? Has the EA resulted in an increase to the average and best fitness? Why/why not? (2p)

f) Assume we want to hybridize our EA with a local search. The local search searches through all solutions that are 1 bit-flip away from our current solutions (in other words, the search tests 5 “neighbours” of the current individual), and selects the best one found. We insert the local search right after mutation, as the final part of the evolutionary loop before the next generation begins. Assume we have the following individual after mutation: [1 0 0 1 0]. With your fitness function above, what is the fitness of this individual?
What is the genotype and fitness after the local search, assuming Lamarckian Evolution?
What is the genotype and fitness after the local search assuming Baldwinian Evolution? (2p)

3) Evolutionary Algorithms – Variation Operators (7p)

Assume the following two genotypes in a permutation-style EA problem:
g1 - [1 3 5 4 2 6 7]
g2 - [7 4 3 5 2 1 6]

Show the steps in your calculations below, and not just the final answer.

a) Why is it important to use specialized variation operators when genotypes are permutations? (1p)

b) What is the result of performing:

   a. Inversion Mutation on each genotype, from the second to the fourth gene (0.5p)

   b. Swap mutation on each genotype, with the second and fifth gene as swap points (0.5p)

c) Partially Mapped Crossover between g1 and g2, assuming the initial segment is taken from the first parent, from gene position 4 to 6 (the segment [4 2 6]). You should only create the first child. (2p)

d) Edge Crossover between g1 and g2. Whenever you have to make a random choice, choose instead the lowest number. Start with 1 as the initial element. Show your calculated edge table. (3p)

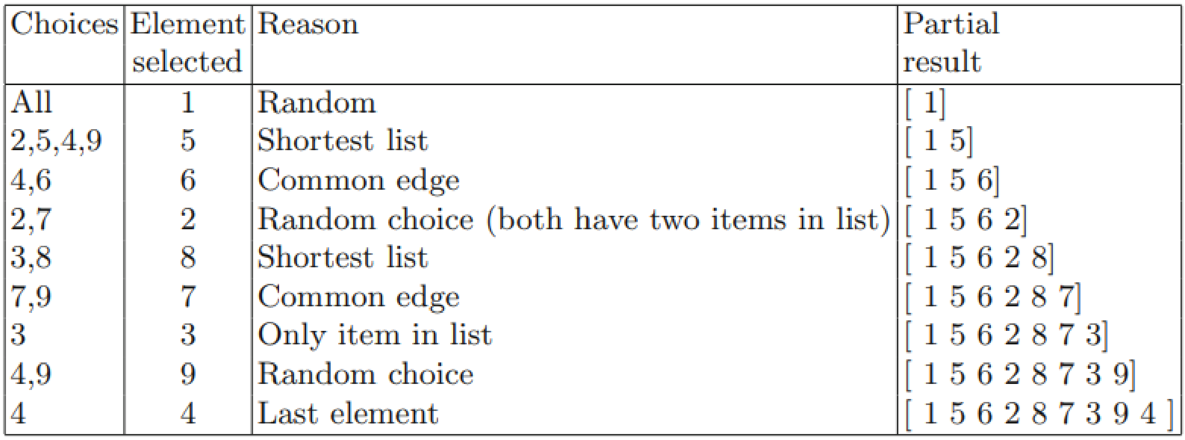
Edge table:

Element	Edges
1	
2	
3	
4	
5	
6	
7	

Show how you construct the new individual by setting up a table similar to the one in the book. We have filled in the first element for you:

Choices	Element Selected	Reason	Partial Result
All	1	Random choice	[1]

The example table from the book is given below for convenience – but note that your edges and calculations will be different.

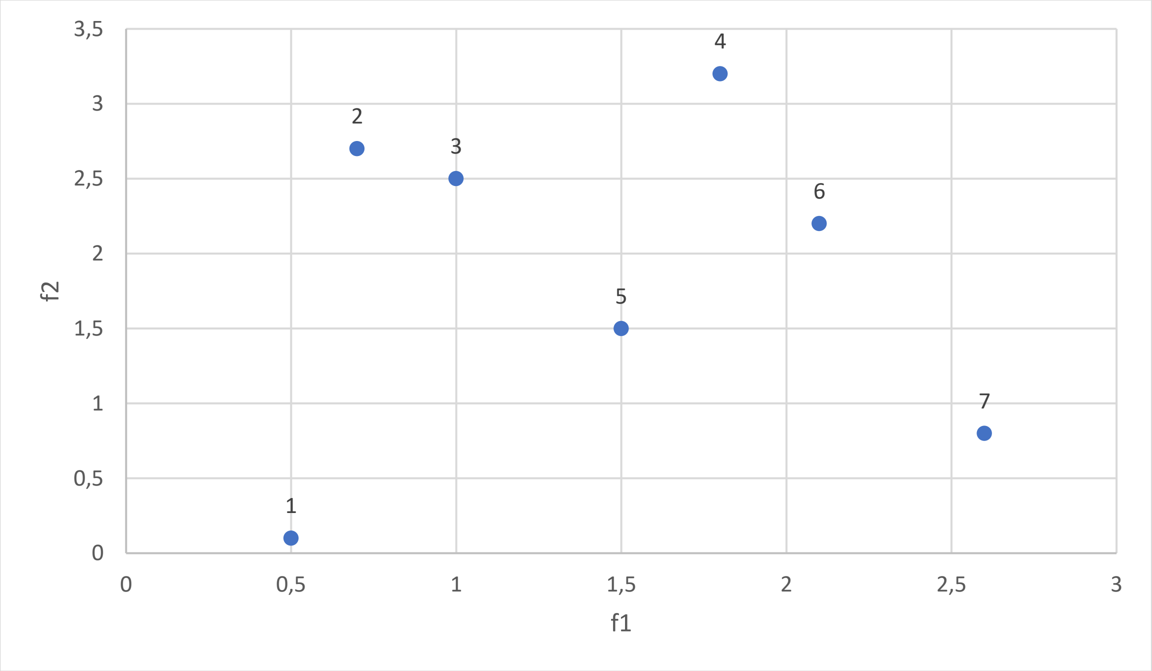


Choices	Element selected	Reason	Partial result
All	1	Random	[1]
2,5,4,9	5	Shortest list	[1 5]
4,6	6	Common edge	[1 5 6]
2,7	2	Random choice (both have two items in list)	[1 5 6 2]
3,8	8	Shortest list	[1 5 6 2 8]
7,9	7	Common edge	[1 5 6 2 8 7]
3	3	Only item in list	[1 5 6 2 8 7 3]
4,9	9	Random choice	[1 5 6 2 8 7 3 9]
4	4	Last element	[1 5 6 2 8 7 3 9 4]

Table 4.3. Edge crossover: example of permutation construction

4) Pareto Optimality (3p)

For an optimization problem we wish to optimize solutions according to two different objectives, f1 and f2. The fitness values according to the two objectives for 7 different solutions are plotted in the figure below.



a) What requirements do the solutions in a Pareto optimal set need to fulfill? (1p)

Find the Pareto optimal set of solutions when

b) Maximizing f1 and f2 (1p)
c) Maximizing f1 but minimizing f2 (1p)

5) Classification (10 p)

Given the following data

Item	X1	X2	Class
A	-0.3	0.3	no
B	0.3	0.6	yes
C	0.6	0.6	yes
D	0.6	0.3	no

a) The goal is to train a perceptron classifier on these data. Set the bias to -1 and the learning rate to 0.1. Also assume that each of the weights is initially set to 0.1. Run the perceptron algorithm sequentially through the data in the order given from A to D. Show how the points get classified and update the weights. (3p)

b) Repeat two more rounds. Has the algorithm converged? (2p)

c) Suppose you had instead used a batch strategy for training. Run one round A-D with the same initial weights and show how the weights are updated. (2p)

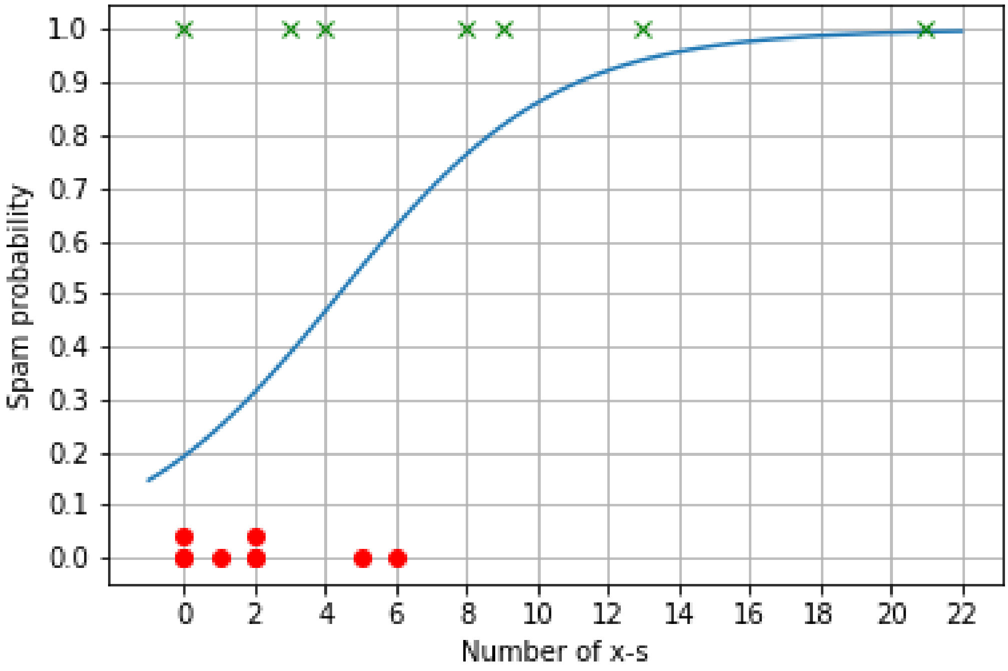
d) Assume that we use a linear regression classifier instead. Run one round batch update with the linear regression classifier and show how the weights are updated. (3p)

6) Logistic Regression (6 p)

Kim is building a spam filter. She has the hypothesis that counting the occurrences of the letter ‘x’ in the e-mails will be a good indicator of spam or no-spam. She collects 7 spam messages and 7 no-spam messages and counts the number of x-s in each. Here is what she finds.

- Number of ‘x’-s in each spam: [0, 3, 4, 8, 9, 13, 21]
- Number of ‘x’-s in each no-spam: [0, 0, 1, 2, 2, 5, 6]

She trains a logistic regression classifier on the data and plots the classifier against the data.



Assume the logistic regression model and answer the following questions:

a) Consider an e-mail with no ‘x’-s. According to the model, what is roughly the probability of this message being a spam message and what the probability of it not being spam? (1p)

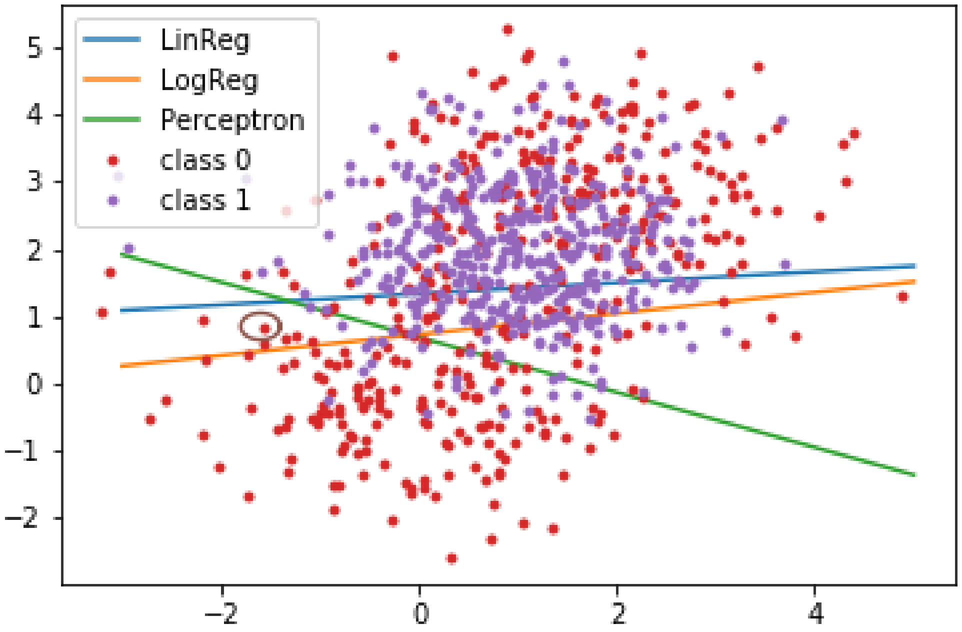
b) How many x-s must an e-mail contain to guarantee it is a spam mail? (1p)

c) How is a logistic regression model normally turned into a binary classifier? If you turn the model into a classifier in this way, what is the accuracy of the classifier on the training data? (2p)

d) It is most important that no no-spams are classified as spams. How can this goal be described in terms of precision and recall? How can the logistic regression classifier be modified to try to achieve this goal? (2p)

7) Majority Voting Classifier (4 p)

We have trained three different classifiers on the same training data (from mandatory assignment 2); a linear regression classifier, a logistic regression classifier, and a perceptron classifier. We have plotted the decision boundaries for all three classifiers on the training data in the figure. They all classify the points above their boundaries as class 1 (purple) and the points below the boundary as class 0 (red). By referring to the figure and the circled point, explain how a majority voting classifier works. In particular, describe the decision boundary for the majority vote classifier.



8) Backpropagation (10p)

(From INF3490/4490 — Biologically Inspired Computing, exam 2014)

SiO, the student welfare organization, would like to have a system for sorting utensils after washing. You are going to help them design a camera-based classifier system for sorting knives, forks, spoons and teaspoons into separate bins. You have a machine vision library available that lets you identify where there is a utensil in the camera images, and it extracts a large number of features for each identified object that we can use as inputs.

a) What class of learning algorithm would be best to use in this case: supervised, unsupervised or reinforcement learning? Justify your answer. (1p)

b) We would like to make a system for distinguishing the utensils using a multi-layer perceptron network. How many output neurons should the network have, and what would each of them represent? (1p)

c) Sketch the steps in the forward and backward phase of the multi-layer perceptron algorithm (backpropagation). Use words and not equations. (3p)

d) An error term is used for updating the weights of the output layer:
Explain the different parts (including what they represent) of the equations (you don´t need to use indices).

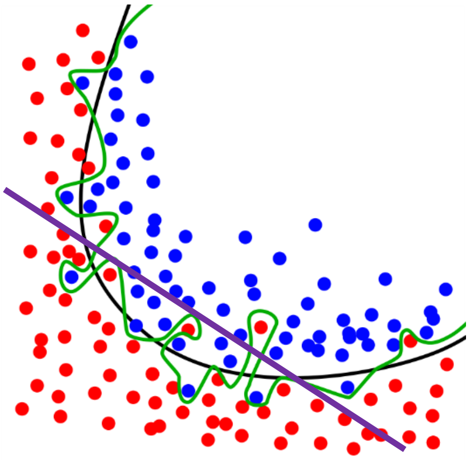
$$\delta_o(\kappa) = (y_\kappa - t_\kappa)\, y_\kappa (1 - y_\kappa)$$

How is the term above used for updating the weights in the hidden layer? (2p)

e) What are the different approaches to how often weights are updated during training? (1p)

f) How would you find out when to stop the training? (2p)

9) Overfitting and Bias (10 points)



a) What do we mean by ‘overfitting’ in machine learning, and why can it become a problem? You may refer to the figure in the discussion or make examples. (2p)

b) The term ‘bias’ is used in various ways in machine learning. We are here considering the use as in “the bias-variance trade-off” and “inductive bias”. What do we mean by ‘bias’ in these contexts? Again, you may refer to the figure in the discussion or make examples. (2p)

c) You are facing a regression problem. You first try a linear regression model. What are the possible problems you may encounter with ‘overfitting’ and ‘bias’? (1p)

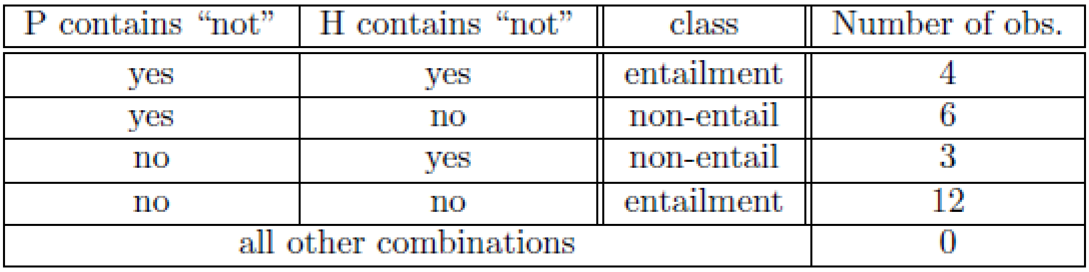
d) You want to address the bias problem. You still want to use linear regression as learning algorithm. Which other settings can you change to get a model which faces less problems with the bias, and how will you change them? Which new problems may these improvements cause? (2p)

e) You are training a logistic regression classifier. Which problems can arise with respect to overfitting and bias? (1p)

f) We will concentrate on overfitting. Which methods are there for avoiding overfitting when using gradient descent for logistic regression? Explain the main parts of each method. (2p)

10) Decision Tree Classifier (10 points)

(The following example is of course simplified.) Kim is training an entailment classifier on 25 training items. Each item consists of a premise, P, and a hypothesis, H. The test items belong to one of two classes: Entailment or Non-entailment. Kim has decided to use two features only, whether the premise contains the word *not* and whether the hypothesis contains *not*. The 25 observations are summarized in the following table.



P contains "not"	H contains "not"	class	Number of obs.
yes	yes	entailment	4
yes	no	non-entail	6
no	yes	non-entail	3
no	no	entailment	12
all other combinations			0

a) Construct a decision tree classifier from the training data. (3p)

b) What is the accuracy of the classifier on the training data? (1p)

c) What is the precision and recall of the classifier for the entailment class? (1p)

d) Suppose you prune the tree after the first stump. What is now the accuracy for the classifier and the precision and recall for the entailment class? (1p)

e) Make a plot showing the points and the decision boundary/boundaries for the decision tree classifier. (2p)

f) What is the best accuracy a logistic regression classifier can achieve on this data set? Sketch a decision boundary for such a classifier in the same figure. (2p)

11) Markov Property (4p)

Explain the Markov property with reference to the following formulas:
Formula 1:

$$Pr(r_t = r', s_{t+1} = s' | s_t, a_t).$$

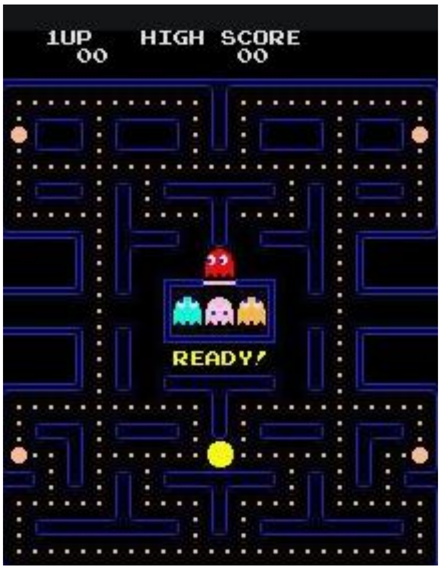
Formula 2:

$$Pr(r_t = r', s_{t+1} = s' | s_t, a_t, r_{t-1}, s_{t-1}, a_{t-1}, \ldots r_1, s_1, a_1, r_0, s_0, a_0)$$

(1p)

Do the following problems fulfill the Markov Property? Briefly explain why. If not, specify some of the information we lose by modelling them as Markov Decision Processes. Specify any additional assumptions you make. (1p for each)

a) Playing chess, assuming the state is a description of the positions of all chess pieces on the board.
b) Driving a car assuming the state is an image of what is currently in front of the car and actions are accelerate, break, turn left or turn right.
c) Playing the computer game Pac-Man assuming the state is an image of the current situation in the game.



12) Unsupervised Learning (20p)

Alice and Bob, your colleagues from the astrophysics department, have given you a collection of astronomical data[1] describing exoplanets in different star systems. Each exoplanet is described by the distance from its orbiting star in AU (astronomical units), its mass (as multiples of the Earth), and the degree of light reflection (as an albedo integer). See table below.

a) Visualization (2p)

	AU from star	Mass	Albedo
HD 209458 b	2	3	7
HD 189733 b	5	3	3
51 Pegasi b	7	2	5
PSR B1257+12 B	3	5	6
PSR B1257+12 C	5	4	5
OGLE-TR-56 b	7	4	3
Fomalhaut b	3	3	8
2M1207 b	4	3	7

Some of these data (about half) have been collected using the *transit detection method* and others (about half) using an *infrared detection method*. Alice and Bob know that these two methods are sensitive to exoplanets with different features, but they do not know which sample has been collected with which method.

Alice argues that looking at *AU from the star* and *albedo* may help them infer which observations were performed with which techniques; Bob holds that looking at *AU from the star* and *mass* may provide a better perspective to group the exoplanets by their discovery method.

Plot the data first according to Alice's hypothesis and then Bob's hypothesis. Which hypothesis seems more likely?

[1] Exoplanet names are real. All the other details are made up.

b) k-means (5p)

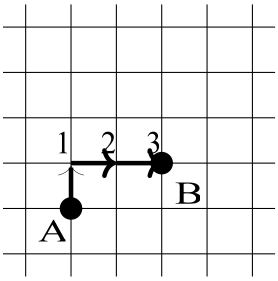
To give more grounding to your conclusions, you decide to run the *k-means algorithm* on your data using two clusters, one for each detection method. Let us call one cluster the blue cluster, and the other one the red cluster.

Run three iterations of k-means (assignment, recomputation of the centroids) on Alice's and Bob's data.

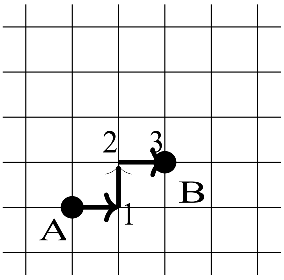
Initialize the blue cluster at (3,2), and the red cluster at (8,4).

In the assignment phase, use as a distance function the Manhattan distance D_{Man} [x_i,x_j], that gives you the number of straight segments necessary to get from one point to the other, for example:

If a point is the same distance from the center of both clusters, assign it to the blue cluster.



(a) Manhattan distance



(b) Notice that the distance is independent from the path.

In the recomputation of the centroids, round the values of the nearest integer:

5.3 → 5

2.5 → 3

3.8 → 4

How do your results agree with your conclusions from the visualization exercise?

c) Quantitative evaluation (5p)

Bob and Alice look very interested in your results: it seems that clustering based on a given pair of features is better than clustering on another set of features. However, they are uneasy accepting a solution based on an intuitive visualization. They ask if your results may be given a quantitative explanation.

You think that an easy way would be to compute the *separation* between the clusters, that is computing the distance between the blue point that is closest to the red cluster and the red point that is closer to the blue cluster. This measure would quantify the gap between the two clusters.

Compute the separation for the clustering of Alice's data and Bob's data.

This measure would quantify the gap between the two clusters.

What would you conclude from the computation of cluster separation?

Yet, you feel this measure is not very robust.

What problem could you imagine having when using cluster separation?

You ask around your colleagues, and Yoshua explains to you that there are two important measures to evaluate clustering: the *inter-cluster distance*, measuring how separate two clusters are, and the *intra-cluster distance*, measuring how compact a cluster is.

Compute the inter-cluster distance for the clustering of Alice's data and Bob's data (do not round to integers).

To compute inter-cluster distance simply compute the distance between the centroid of the red and blue cluster: $D_{inter}[c_{blue}, c_{red}] = D_{Man}[t_{blue}, t_{red}]$, where *c* is a cluster and *t* is a centroid.

Compute the intra-cluster distance for the clustering of Alice's data and Bob's data (do not round to integers).

Differently from the inter-cluster distance, the intra-cluster must be computed for each cluster individually. For each cluster, red or blue, compute the average distance of all the cluster points from the cluster center. For the blue cluster: $D_{intra}[c_{blue}] = \frac{1}{N_{blue}} \sum_{x \in c_{blue}} D[x, t_{blue}]$; similarly for the red cluster. Average then the intra-cluster distance of the blue and red cluster to get the overall intra-cluster distance for Alice and Bob: $\frac{1}{2}(D_{intra}[c_{blue}] + D_{intra}[c_{red}])$.

A good cluster is a cluster that clumps its point tightly close to each other, and that is far removed from other cluster. It is natural to assess the goodness of your clustering as the ratio between inter-cluster distance (which you want to be big) and intra-cluster distance (which you want to be small).

Compute the ratio of inter-cluster distance and intra-cluster distance for the clustering of Alice's data and Bob's data (do not round to integers). How does this confirm/reject your previous conclusions?

d) Processing new data (3p)

Alice and Bob are happy with your solution and decide to adopt the clustering you argued being the best one. From now on, we will use only the clustering that you proved being the best. Now new data has come in:

	AU from star	Mass	Albedo
Beta Pictoris c	9	3	6
K2-282c	6	5	7
Kepler-1658b	2	2	8

Start from the chosen clustering, plot the new data points and assign them to the correct cluster.

e) Outliers (2p)

There is a further recording, coming from another institution, that Alice and Bob would like to process:

	AU from star	Mass	Albedo
Luyten 98-59 d	22	3	3

Alice and Bob are not certain about the quality of this recording and ask your opinion.

Use the chosen clustering, plot the new data point. What do you think about this observation?

For your own interest, you decide to analyze how this new data point will affect the clustering process.

Restart from the original data set of eight data points; place the centroids of the two clusters as you computed them at the end of Section 1.2; add the new data on Luyten 98-59 d and run two iterations of the k-means algorithm. What happens to the clusters?

f) Rescaling (3p)

After discussing with other colleagues at a conference, Alice and Bob became suspicious that the recordings of albedo may be wrong. Following the suggestion of Eve, they are thinking about reducing by half all the recorded values of albedo. They present this possible change to you, and ask your opinion.

How would you interpret the change that they have proposed?

In particular, they are concerned whether this change would affect your results.

Apply the transformation to the original data. Then run two iterations of k-means with the same initialization used in Section 1.2 on Alice's data. What do you observe?

When halving the observed values of albedo, always round down to the closest integer:

3.5 → 3

Is k-means insensitive to the suggested transformation? If not, how would you tackle this inconsistency?

13) Particle Swarm Optimization (PSO) and Developmental Systems (3p – 1 per question)

a) Describe what happens when the position of all particles in PSO are set to the same value of a local optimum (Think about fitness landscapes). How could you adjust PSO to get out of the local optimum to potentially find the global optimum without resetting the particles to random positions initially? Describe your approach in up to 100 words.

b) An L-System is a parallel rewrite method. Describe how from an alphabet {h,j}, the axiom h will be rewritten when using the rewrite rules: h->jjh and j->h. Write down three iterations/recursions.

c) When visualizing a string of an L-System, it is useful to implement a bracketed L-System. Describe what '+', '-', '[' and ']' in such L-Systems are used for.