IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2019-2020

MSc in Artifical Intelligence
MSc in Computing Science (Specialist)
for Internal Students of the Imperial College of Science, Technology and Medicine

PAPER C553

INTRODUCTION TO MACHINE LEARNING

Wednesday 11th December 2019, 10:00
Duration: 90 minutes

Answer TWO questions

Paper contains 3 questions
Calculators required

1a Given the 2D points $A = (3,6), B = (6,6), C = (6,3), D = (-3,9), E = (-6,-3), F = (-3,-3)$

Starting from initial clusters Cluster1 = A which contains only the point A and Cluster2 = D which contains only the point D, run the K-means clustering algorithm and report the final clusters. Use L1 distance as the distance between points which is given by
$d((x_1, y_1), (x_2, y_2)) = \|x_1 - x_2\| + \|y_1 - y_2\|$

Finally, draw the points on a 2-D grid and check if the clusters make sense.

b In the part above, we considered two clusters. Is this the optimal number of clusters to solve this problem? Explain the methodology you would use to determine the optimal number of clusters.

c Now, consider this set of 1D points: D={1, 2, 3, 5, 8, 10, 11}
We have two alternative Gaussian Mixture Models (GMM) that we want to compare. GMM_1 and GMM_2 are defined as follows:
$GMM_1 = \{\mu_1 = 4, \sigma_1 = 1, \pi_1 = 0.5, \mu_2 = 6, \sigma_2 = 0.5, \pi_2 = 0.5\}$
$GMM_2 = \{\mu_1 = 2, \sigma_1 = 0.1, \pi_1 = 0.9, \mu_2 = 8, \sigma_2 = 0.5, \pi_2 = 0.1\}$

Which of the two GMMs is the best representation of the dataset D? Explain the methodology used to reach your conclusion and write out the intermediate and the final results.

To help you, you can refer to the following table providing the value of $\mathcal{N}(x|\mu, \sigma)$:

						x					
μ	σ	1	2	3	4	5	6	7	8	9	10	11
2	0.1	0.000	3.990	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.5	0.108	0.798	0.108	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	1	0.242	0.399	0.242	0.054	0.004	0.000	0.000	0.000	0.000	0.000	0.000
4	0.1	0.000	0.000	0.000	3.990	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.5	0.000	0.000	0.108	0.798	0.108	0.000	0.000	0.000	0.000	0.000	0.000
4	1	0.004	0.054	0.242	0.399	0.242	0.054	0.004	0.000	0.000	0.000	0.000
6	0.1	0.000	0.000	0.000	0.000	0.000	3.990	0.000	0.000	0.000	0.000	0.000
6	0.5	0.000	0.000	0.000	0.000	0.108	0.798	0.108	0.000	0.000	0.000	0.000
6	1	0.000	0.000	0.004	0.054	0.242	0.399	0.242	0.054	0.004	0.000	0.000
8	0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	3.990	0.000	0.000	0.000
8	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.108	0.798	0.108	0.000	0.000
8	1	0.000	0.000	0.000	0.000	0.004	0.054	0.242	0.399	0.242	0.054	0.004

d Assume that it is possible to produce a better GMM than the two models listed above. Explain the algorithm that you will use to find this GMM (the number of mixture components is fixed to two), and provide the equations associated with each step of the algorithm.

The four parts carry, respectively, 30%, 20%, 30%, and 20% of the marks.

2a Consider the following set of training examples regarding a decision process to know whether or not to purchase an item:

	Quality	Brand	Price	Review	Buy?
1	High	Famous	Cheap	Bad	No
2	Poor	Famous	Cheap	Bad	No
3	High	Mainstream	Expensive	Good	No
4	Poor	Mainstream	Cheap	Good	Yes
5	Standard	Mainstream	Cheap	Good	Yes
6	Standard	Unknown	Expensive	Good	No
7	Poor	Mainstream	Cheap	Bad	No
8	High	Famous	Cheap	Good	Yes
9	Poor	Mainstream	Expensive	Bad	No
10	Standard	Mainstream	Cheap	Good	Yes

Apply the algorithm seen in class (ID3) to create a decision tree. Write out the intermediate and the final results. Draw the produced decision tree.

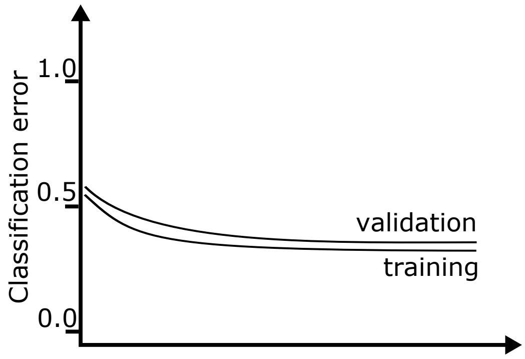
b You want to tune the hyper-parameters of a neural network by using a genetic algorithm. The network is a multi-layer perceptron used for a classification task and the tunable hyper-parameters are:
- number of hidden layers. Possible values: {1,2,3,4}
- number of neurons per hidden layer. Each hidden layer can have a different number of neurons. Possible values: {1,3,5,10,15,20,30,50}
- activation function (for all neurons). Possible values: {tanh, sigmoid, relu, linear}
- learning rate. Possible values:{1, 0.5, 0.1, 0.01}
Give a suitable genotype, phenotype, and function used to develop a genotype into a phenotype, that you would use to solve the problem described above. Explain your answer in a clear and compact manner.

c Choose the genetic operators that can be used to solve this problem (given the genotype you defined in the previous part). Explain your answer in a clear and compact manner.

d You are given a large dataset associated with the considered classification task. Define the fitness function that can be used to solve this hyper-parameter problem and explain how you evaluate the final performance of the network. Explain your answer in a clear and compact manner.

The four parts carry, respectively, 30%, 20%, 20%, and 30% of the marks.

3a You are training a neural network on a binary classification task that is known to be solvable by neural networks with classification error below 10%. After a sufficient number of epochs, the training and validation classification error plots of your network look as follows:



What is the problem and what can you do to overcome this situation? (We assume that there is no implementation bug).

b What is the ”vanishing gradient” problem? Why is it a problem? Why is ReLU a solution to this problem?

c The classification problem considered here is a multi-class problem, with four classes. What does this impose on the network architecture and what is the most appropriate loss function in this case? Give the definition of this loss function.

d Consider the following 40 predictions from the network. The ”id” column is the index of the sample (from 1 to 40), the ”L” column corresponds to the true label of the sample and the ”P” column refers to the prediction (output) of the network.

id	L	P		id	L	P		id	L	P		id	L	P
1	c1	c1		11	c2	c1		21	c3	c4		31	c4	c4
2	c1	c2		12	c2	c2		22	c3	c1		32	c4	c4
3	c1	c1		13	c2	c2		23	c3	c2		33	c4	c4
4	c1	c1		14	c2	c4		24	c3	c4		34	c4	c3
5	c1	c4		15	c2	c2		25	c3	c4		35	c4	c4
6	c1	c1		16	c2	c2		26	c3	c2		36	c4	c4
7	c1	c3		17	c2	c3		27	c3	c4		37	c4	c1
8	c1	c1		18	c2	c2		28	c3	c1		38	c4	c2
9	c1	c2		19	c2	c2		29	c3	c4		39	c4	c3
10	c1	c1		20	c2	c2		30	c3	c1		40	c4	c4

Compute the confusion matrix, accuracy, recall, precision, F1, and UAR metrics. What comment can you make?

e You are now developing a new gradient descent algorithm as an alternative to Stochastic Gradient Descent (SGD), and you want to evaluate its benefits in terms of performance. Given a common neural network architecture, you execute 10 training sequences with both your new algorithm and SGD (all starting from random initial conditions, with different seeds for the random number generator). The median performance of the networks trained with your algorithm is 13% better than those trained with SGD. You use the Wilcoxon rank sum test which gives you a p-value=0.07 (for a pre-defined threshold of 0.05). What can you conclude from your evaluation and what can be done to improve it?

The five parts carry equal marks.