Do not open or start the test before the announcement is made to start.

Student Name: ___________________________ ID: _____________

Class Test

Machine Learning

and

Machine Learning (extended)

17th Nov 2017

Total time: 40 minutes

Total marks: 20

Instructions

This is a closed book and closed notes exam.

Mobiles, smartwatches, and similar electronic devices must not be used.

Simple calculator is allowed.

Write your answers in the free space after each question.

No extra sheet or paper should be used, any rough work should be done on the back pages.

1. Let's consider the following two-dimensional data which belongs to two clusters as shown:

   Cluster 1: (6,7), (8,6)

   Cluster 2: (1,1), (2,4), (3,3)

   Given the above information, can we determine which cluster a new point (4,5) belongs to by using k-means clustering algorithm? Show all the working/reasoning to support your answer. Describe any assumptions you may consider. [3 marks]

2. Given the total training loss estimate function below between actual target labels t_n and model predicted target labels $\boldsymbol{w}^T\boldsymbol{x}_n$ where $\boldsymbol{x}_n$ denotes the attributes and $\boldsymbol{w}$ denotes model parameters:

$$\mathcal{L} = \sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2 = (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})$$

   Derive the expression to estimate parameters $\widehat{\boldsymbol{w}}$ for a linear model that minimizes this total loss. [3 marks]

3. Consider the objects (V,W,X,Y,Z) described by the distance matrix below:

	V	W	X	Y	Z
V	0	4	1	2	4
W		0	4	4	3
X			0	4	5
Y				0	5
Z					0

Using hierarchical agglomerative clustering with min/single link, what will be the cluster formation at dendrogram junction heights 1, 2, 3, 4, and 5? Show all the working and the dendrogram. [3 marks]

4. Given the confusion matrix below, compute the sensitivity and specificity for disease 1. Note that sensitivity estimates the proportion of diseased samples that are classified as diseased, while specificity estimates the proportion of healthy samples that are classified as healthy. Show all the working. [3 marks]

		Actual class label		
Predicted class label		Disease 1	Disease 2	Healthy
	Disease 1	19	3	2
	Disease 2	4	15	3
	Healthy	4	1	21

5. Consider the following data set with two input attributes (i.e. the x and y coordinates of the points) and one binary output t. We want to use k-NN classifier with Euclidean distance to predict target t. Compute the average leave-one-out cross-validation error of a 3-NN classifier on this data set. Show all the working. [4 marks]

x	y	t
-3	-2	+
-1	-2	+
-3	0	+
-1	0	+
3	1	+
-2	-1	–
-2	1	–
2	-2	–
2	0	–
3	-1	–
4	-2	–
4	0	–

6. Let's consider the mean and covariance of two class dataset below:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 2 \\ 2 & 3 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & -2 \\ -2 & 4 \end{bmatrix}$$

Make a rough drawing of how the point cloud of each class would appear if generated with a multivariate Gaussian probability density function. Comment about whether this data can be separated by a linear boundary. What will be the shape of the decision boundary with a maximum likelihood Gaussian classifier, without naïve assumption? [4 marks]