UNIVERSITY OF WARWICK

LEVEL 7 Open Book Assessment [2 hours]

Department of Computer Science

CS909-1 Data Mining (MSc)

Instructions
1. Read all instructions carefully – and read through the entire paper at least once before you start writing.
2. There are **6** questions. You should **attempt 4 questions in total with at least ONE question from Section B.** You should not submit answers to more than the required number of questions.
3. All questions will carry the same number of marks unless otherwise stated.
4. You should handwrite your answers either with paper and pen or using an electronic device with a stylus (unless you have special arrangements for exams which allow the use of a computer). Start each question on a new page and clearly mark each page with the page number, your student id and the question number. Handwritten notes must be scanned or photographed and all individual solutions should (if you possibly can) be collated into a single PDF with pages in the correct order. You must upload two files to the AEP: your PDF of solutions and a completed cover sheet. You must click **FINISH ASSESSMENT** to complete the submission process. After you have done so you will not be able to upload anything further.
5. Please ensure that all your handwritten answers are written legibly, preferably in dark blue or black ink. If you use a pencil ensure that it is not too faint to be captured by a scan or photograph.
6. Please check the legibility of your final submission before uploading. It is your responsibility to ensure that your work can be read.
7. You are allowed to access module materials, notes, resources, references and the internet during the assessment.
8. You should not try to communicate with any other candidate during the assessment period or seek assistance from anyone else in completing your answers. The Computer Science Department expects the conduct of all

students taking this assessment to conform to the stated requirements. Measures will be in operation to check for possible misconduct. These will include the use of similarity detection tools and the right to require live interviews with selected students following the assessment.
9. By starting this assessment you are declaring yourself fit to undertake it. You are expected to make a reasonable attempt at the assessment by answering the questions in the paper.
Please note that: - You must have completed and uploaded your assessment before the 24 hour assessment window closes. - You have an additional 45 minutes beyond the stated length of the paper to allow for downloading and uploading the assessment, your files and technical delays. - For further details you should refer to the AEP documentation.

Use the AEP to seek advice immediately if during the assessment period:

- you cannot access the online assessment;
- you believe you have been given access to the wrong online assessment.

Please note that technical support is only available between 9AM and 5PM (BST).

Invigilator support will be also be available (via the AEP) between 9AM and 5PM (BST).

Notify Dcs.exams@warwick.ac.uk as soon as possible if you cannot complete your assessment because:

- you lose your internet connection;
- your device fails;
- you become unwell and are unable to continue;
- you are affected by circumstances beyond your control (e.g. fire alarm).

Please note that this is for notification purposes, it is not a help line.

Your assessment starts below.

Section A

1. (Linear Classifiers)

(a) **What** is the typical life cycle of a project in data mining and machine learning? **Describe** different steps involved in the process. [5]

(b) **What** is meant by linear separability in machine learning? **Give examples**, by a scatter plot, of a linearly separable dataset and a linearly non-separable dataset. [5]

(c) If we represent an example with d-features by a vector **x**, what is the discriminant function f(**x;w**) of a linear classifier with a weight vector **w** for this example? **What** is the role of the bias term in the discriminant function? [5]

(d) **Show** whether the following dataset is linearly separable or not. **Write** the set of classification constraints for each example and show whether those constraints can be satisfied by a linear classifier or not. **Verify** your reasoning by plotting the data and the linear discriminant. [5]

Index	Example Features	Label
1	$\begin{bmatrix}0\\0\end{bmatrix}$	-1
2	$\begin{bmatrix}0\\1\end{bmatrix}$	1
3	$\begin{bmatrix}1\\0\end{bmatrix}$	1
4	$\begin{bmatrix}1\\1\end{bmatrix}$	1
5	$\begin{bmatrix}-2\\-2\end{bmatrix}$	1

(e) **What** is meant by a loss function between actual and target outputs? **Give** an example each of a loss function for binary classification, regression and one-class classification by explaining how they measures error through a **plot** of the loss function with the target output value set to -1. **Give** the complete mathematical equations for the loss functions with clear explanation on what is meant by each variable. [5]

2. (Performance Evaluation)

(a) **What** is stratified cross-validation? If you have a total of 100 examples for validation with 30 positive and 70 negative examples, **what** is the total number of positive and negative examples in each fold of 5-fold stratified cross-validation? **Why** would you use stratified cross-validation as opposed to non-stratified cross-validation? [5]

(b) **What** are the limitations of accuracy as a performance metric? **Give reasoning** behind these limitations and a **suitable alternative metric** that does not suffer from these limitations. [5]

(c) What are the advantages and limitations of using Area Under the Receiver Operating Characteristic Curve (ROC)? **Give reasoning** behind these limitations and suggest **an alternate metric** that does not suffer from these limitations. [5]

(d) Given the following confusion matrix, calculate the true positive rate (TPR), false positive rate (FPR), sensitivity and precision. [5]

		Predicted Label	
		Positive	Negative
True Label	Positive	40	60
	Negative	80	100

(e) **Draw** the ROC curves for the following classifiers and **calculate** the area under both of them. **Give reasoning** behind your answers. [5]

1. A perfect classifier (one that does not generate any errors)
2. A uniformly random classifiers (one that generates random labels for an example)

3. (Dimensionality Reduction and Kernels)

(a) Assume that you have been assigned to develop a machine learning model to predict whether a given image contains a cat or not. The size of the image is 64x64 and you have 10,000 labelled images for training. **What** is the maximum number of principal components for this data? **How** would you go about selecting the number of principal components for dimensionality reduction for this data? [5]

(b) **What** is a kernel function k(**a,b**) between two examples **a** and **b**? **What** purpose does it serve in Support Vector Machines? **Give** examples of at least two kernel functions by writing their **mathematical formulae** and **plotting** their values as a function of one-dimensional feature vector **a** with **b=1** . [5]

(c) **What** is the role of the parameter "C" in support vector machines? **What** happens when the value of C is set very high? [5]

(d) Given the following data points, **plot** the direction of the first principal component on a scatter plot of the data as a vector (arrow). **Give a justification** as to why this would be the case. **Also, write the direction vector as a unit vector**. [5]

Index	Example Features
1	$\begin{bmatrix}0\\0\end{bmatrix}$
2	$\begin{bmatrix}1\\1\end{bmatrix}$
3	$\begin{bmatrix}2\\2\end{bmatrix}$
4	$\begin{bmatrix}3\\3\end{bmatrix}$
5	$\begin{bmatrix}4\\4\end{bmatrix}$

(e) Given the following data points, **plot** the direction of the first principal component on a scatter plot of the data as a vector (arrow). **Give a justification** as to why this would be the case. **Also, write the direction vector as a unit vector**. [5]

Index	Example Features
1	$\begin{bmatrix}0\\0\end{bmatrix}$
2	$\begin{bmatrix}-1.5\\-1\end{bmatrix}$
3	$\begin{bmatrix}-2\\-2.5\end{bmatrix}$
4	$\begin{bmatrix}-3\\-3\end{bmatrix}$
5	$\begin{bmatrix}-4\\-4.5\end{bmatrix}$

4. (Optimization)

a. Given a set of d-dimensional examples $\boldsymbol{x_i} \in \mathbb{R}^{\boldsymbol{d}}, i = 1 \ldots N$ with associated labels y_i, **derive the analytical solution** for the optimal weight parameters for a linear model $f(\boldsymbol{x}; \boldsymbol{w})$ with the following objective function: [10]

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \left(1 + y_i - f(\boldsymbol{x_i}; \boldsymbol{w})\right)^2$$

b. **How can you solve this objective function using gradient descen**t? Specifically, calculate the gradient of the objective function with respect to the weight parameters and write at least one step of the gradient-based weight update equation. [10]

c. What is the role of batch normalization and drop out in neural networks? [5]

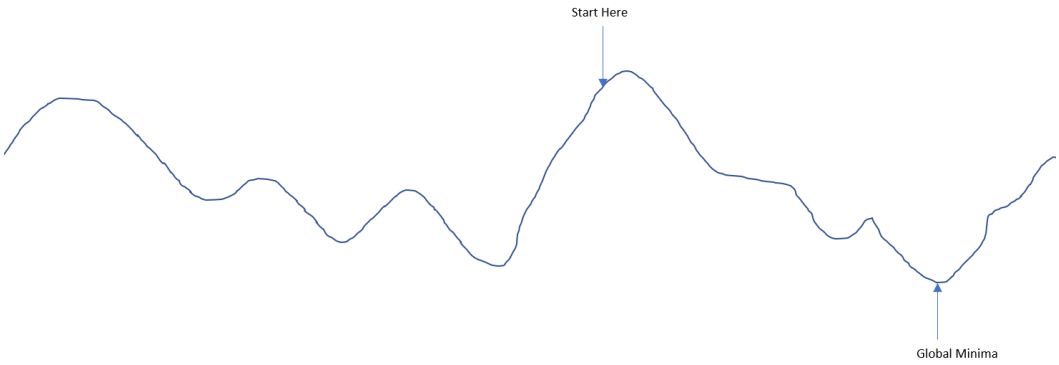
Section B

Answer at least ONE question from Q5 or Q6.

---

5. (SRM and Neural Networks)

a. Consider the following optimization problem with multiple minima. Will the gradient descent algorithm be able to find the global minima of this function by starting at the start location indicated in the figure? **Explain** your reasoning. [5]



b. What is the principle of Structural Risk Minimization? How is it used in the development of machine learning models. [5]

c. How is the number of nearest neighbours in k-nearest neighbour classification related to regularization? [5]

d. **What** is the problem of exploding and vanishing gradients in neural networks? **Describe** the source of this problem in terms of the weight update equation of a multilayer perceptron and **how** it impacts learning as well as how we can counter it. [5]

e. **What** is meant by Term Frequency-Inverse Document Frequency (TF-IDF)? **Calculate** the term frequency of each letter in the following sentence: “a quick brown fox jumps over the lazy dog”. [5]

6. (Application)

You are given the following dataset and are tasked with predicting the popularity of different pizza made by a pizza shop.

Index	Dough Thickness	Topping	Size	Popularity
1	Medium	Chicken	Large	4
2	Medium	Pepperoni	Large	4
3	Thin	Chicken	Large	2
4	Thin	Veggie	Large	3
5	Thick	None	Small	1
6	Thick	Pineapple	Medium	0
So on				

Answer the following questions:

a. **How** would you represent these variables (Dough Thickness, Topping and Size) as features? Give a justification of your design choice. [5]
b. **Would** you model this as a classification problem or a regression problem? Explain the reasoning behind your choice. [5]
c. What additional features do you think can be relevant for this problem? [5]
d. How will you chose a particular machine learning model for this problem? [5]
e. How will you ensure that your predictor generalizes well on data that is not part of your training dataset? [5]