Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Faculty of Engineering, Mathematics and Science
School of Computer Science & Statistics

Instructions to Candidates:

All Questions must be answered. The maximum number of marks per question is stated in square brackets below each question.

Answer Section A in a separate answer book.

When answering open questions, write very clearly. If the examiner cannot **easily** read your answer, you will receive zero marks for your answer.

Answer Section B in a Multiple-Choice Answer Form.

Marks are given as follows.

- Full marks are given for a question when all correct answer(s) and only the correct answer(s) are selected.
- Half the marks are given for a question if:
  - There is only 1 correct answer to the question, and this answer as well as one wrong answer is selected.
  - There are 2 or more correct answers to the question, and
    - all but one correct answers have been selected, or
    - all correct answers and one wrong answer have been selected
- No marks are given otherwise.

Please note that the phrasing of the questions does not indicate how many correct answers there are. For instance, there could be only one correct answer even when the question is “Which of the following statements are true”?

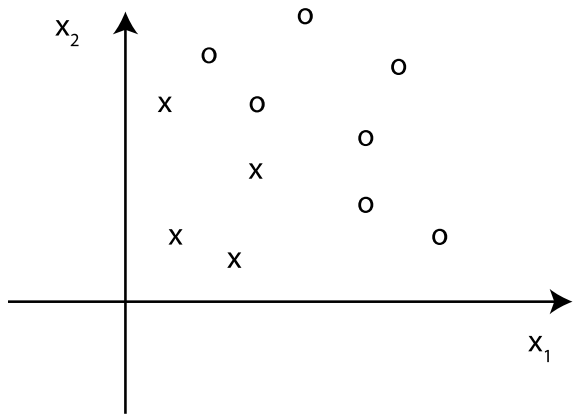
You may not start this examination until you are instructed to do so by the invigilator.

Materials Permitted for this examination:

Black or blue pen, exam paper, non-programmable calculator.

Section A: Open Questions

1. Consider the following two-dimensional classification task, where 'x' indicates class y = 1 points and 'o' class y = −1 points.



(i) (a) Describe the hypothesis/predictive model used in a logistic regression model, and its parameters. [5 marks]

(b) And explain how it can be fitted to this training data. [5 marks]

(ii) Will a logistic regression model fitted to this training data correctly predict the labels for all these training points, or will it make mistakes for some points ? Explain your answer. [5 marks]

(iii) Suppose now that a SVM (the parametric version i.e. not using kernels) is used to classify this data. Which are the support vectors in the above training data ? [5 marks]

(iv) Suppose that the penalty term used in the SVM cost function is $\lambda\theta^{\mathrm{T}}\theta$. As λ is increased what will be the effect on the margin, support vectors and classification accuracy ? Explain your answer with reference to the above training data. [5 marks]

Model Solution

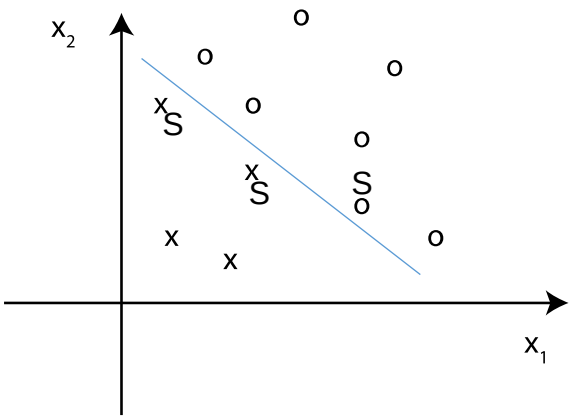
(i)(a) sign($\theta^T x$), where $x=[x_1, x_2]^T$ and θ is a vector of (unknown) parameters.

(b) For logistic regression with m training data points $(x^{(i)}, y^{(i)})$ we select the value of θ that minimises the empirical loss function:

$$\frac{1}{m}\sum_{i=1}^{m}\log(1+e^{-y^{(i)}\theta^T x^{(i)}})/log(2)$$

(ii) The data can be separated by a straight line (it is linearly separable), so logistic regression will be able to make accurate predictions.

(iii) The support vectors are marked by S for the indicated decision boundary below:



These are the points which are equidistant on each side of the decision boundary, and also closest to the boundary.

(iv) The SVM cost function is:

$$\frac{1}{m}\sum_{i=1}^{m}\max(0, 1-y^{(i)}\theta^T x^{(i)}) + \lambda\theta^T\theta$$

In this example the data is linearly separable. It is the ratio of the elements in θ that determine the decision boundary. When $\lambda=0$ we can always make the cost zero by choosing the elements of θ sufficiently large. As λ is initially increased this tends to penalise large elements of θ and so remove this degree of freedom in θ (i.e. the freedom to scale θ without changing the decision boundary) without loss of accuracy or a change in the support vectors. However, as λ becomes very large it will tend to force the elements of θ towards zero regardless

of the first term in the cost function, reducing classification accuracy and changing the support vectors.

2. Consider a 1D dataset with four examples: -3,-1,2,4.

(i) By hand, apply k-means clustering until convergence, assuming the initial cluster centres are -4 and 0. For each iteration give the assignment of examples to clusters and the new values of the cluster centres. [10 marks]

(ii) In general, will the k-means algorithm converge to a local optimum or a global one ? Explain your answer with reference to an example dataset of your choice. [5 marks]

(iii) Explain how the value of the parameter k in k-means clustering can affect over/under fitting of the data. Describe how to use cross-validation to select k. [5 marks]

(iv) Sketch a dataset with at least 5 points on which k-means would work poorly even when k is chosen correctly. Describe why k-means would not work well. [5 marks]

Model Solution

(i) Initial centres are -4 and 0. Assign the points to the nearest centre (where distance is measured e.g. by Euclidean distance): point -3 is assigned to centre -4, points -1, 2 and 4 are assigned to centre 0. Now update the centres to be the average of the assigned points, so the first centre becomes -3/1=-3 and the second centre (-1+2+4)/3=1.66. Assign the points to the nearest of these new centres: points -3 and -1 are assigned to centre -3 and points 2 and 4 to centre 1.66. Update centres to (-3-1)/2=-2 and (2+4)/2=3. Points -3 and -2 are assigned to centre -2 and points 2 and 4 to centre 3, there are now no further changes.

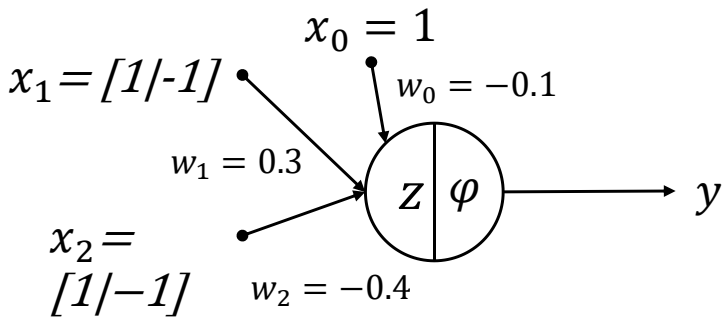
(ii) It will generally converge to a local optimum. For example, it might split points between two clusters when they would be better assigned to a single cluster, depending on the initial choice of cluster centres.

(iii) Selecting k too large will tend to lead to overfitting (there are more clusters than really are present in the data) and selecting k too small to underfitting (there are too few clusters). To use cross-validation to select k, partition the data into e.g. 10 folds. For a range of values of k use each combination of 9 folds (i.e. leaving one out) to run the k-means algorithm and calculate the square cost of the final configuration (the sum of the square of the distances between the data and the centres to which they are assigned). Select the value of k which minimised this cost, taking account of the confidence intervals of the cost indicated by the multiple combinations of folds.

(iv) One example would be where the clusters overlap. This is because k-means assumes that clusters are disjoint.

Section B: Multiple Choice

1. The illustration below shows a perceptron with two inputs x_1 and x_2, which can take the values 1 or -1. x_0 is the bias input with a constant value of 1. The weights of the inputs are given as w_0, w_1 and w_2 with the values as provided in the illustration. The formula for z and the activation function φ are provided below. Which logical operation is the neuron computing, assuming that "1" represents "true" and "-1" represents "false"?



$$\varphi = h(z) = \begin{cases} 1 \; if \; z > 0 \\ -1 \; if \; z \leq 0 \end{cases}$$

$$z = \sum_{i=0}^{n} w_i x_i$$

[7 marks]

A. x_1 AND x_2
B. x_1 OR x_2
C. x_1 XOR x_2
D. x_1 AND (NOT x_2)
E. None of the above

2. A perceptron with a single layer of LTUs (i.e. no hidden layer) shall learn the logical 'OR' operation. It is trained on a dataset with four instances with the input variables x_1 and x_2 and the target variable y, which are all take the values '1' (TRUE) or '-1' (FALSE). The bias input is $x_0 = 1$. The learning rate is η=0.1, and the initial random weights are $w_0 = 0.05$, w_1=0.1 and w_2=0.2. The table below shows the first iterations up to the third epoch. The training has not yet converged, and the weights need to be adjusted. What is the new adjusted weight w_0^{new}, given the data in the table below? As this is a single layer perceptron, just use the simple training algorithm from Frank Rosenblatt and no backpropagation or gradient descent.

Epoch	x0	x1	x2	y	w0	w1	w2	∑	ŷ	Error
1	1	-1	-1	-1	0.05	0.1	0.2	-0.25	-1	0
	1	-1	1	1	0.05	0.1	0.2	0.15	1	0
	1	1	-1	1	0.05	0.1	0.2	-0.05	-1	2
	1	1	1	1	0.25	0.3	0	0.55	1	0
2	1	-1	-1	-1	0.25	0.3	0	-0.05	-1	0
	1	-1	1	1	0.25	0.3	0	-0.05	-1	2
	1	1	-1	1	0.45	0.1	0.2	0.35	1	0
	1	1	1	1	0.45	0.1	0.2	0.75	1	0
3	1	-1	-1	-1	0.45	0.1	0.2	0.15	1	-2
	1	-1	1	1	?					

[7 marks]

A. $w_0^{new} = \mathbf{0.25}$

B. $w_0^{new} = 0.30$

C. $w_0^{new} = 0.35$

D. $w_0^{new} = 0.40$

E. $w_0^{new} = 0.50$

3. The upper part of the illustration shows a term-document matrix with term frequencies for four terms t1…t4 in seven documents d1…d7. The '-' indicates that a document does not contain the corresponding term. Which of the following statements are correct, if you calculated TF-IDF for all terms and documents, and plotted the documents in the vector space with the four dimensions t1, t2, t3, and t4? Use the Euclidean distance to measure distance between documents in the vector space. Feel free to use the lower part of the table to calculate TF-IDF values if needed.

		t1	t2	t3	t4
TF	d1	1	-	12	4,855,654
	d2	97	-	-	48,987
	d3	12	-	894	-
	d4	798	-	208	-
	d5	4,444	200	206	-
	d6	800	202	-	-
	d7	94	-	46	64

		t1	t2	t3	t4
TF-IDF	d1				
	d2				
	d3				
	d4				
	d5				
	d6				
	d7				

[7 marks]

A. Documents d5 and d4 are closer to each other in the vector space than d5 and d6.

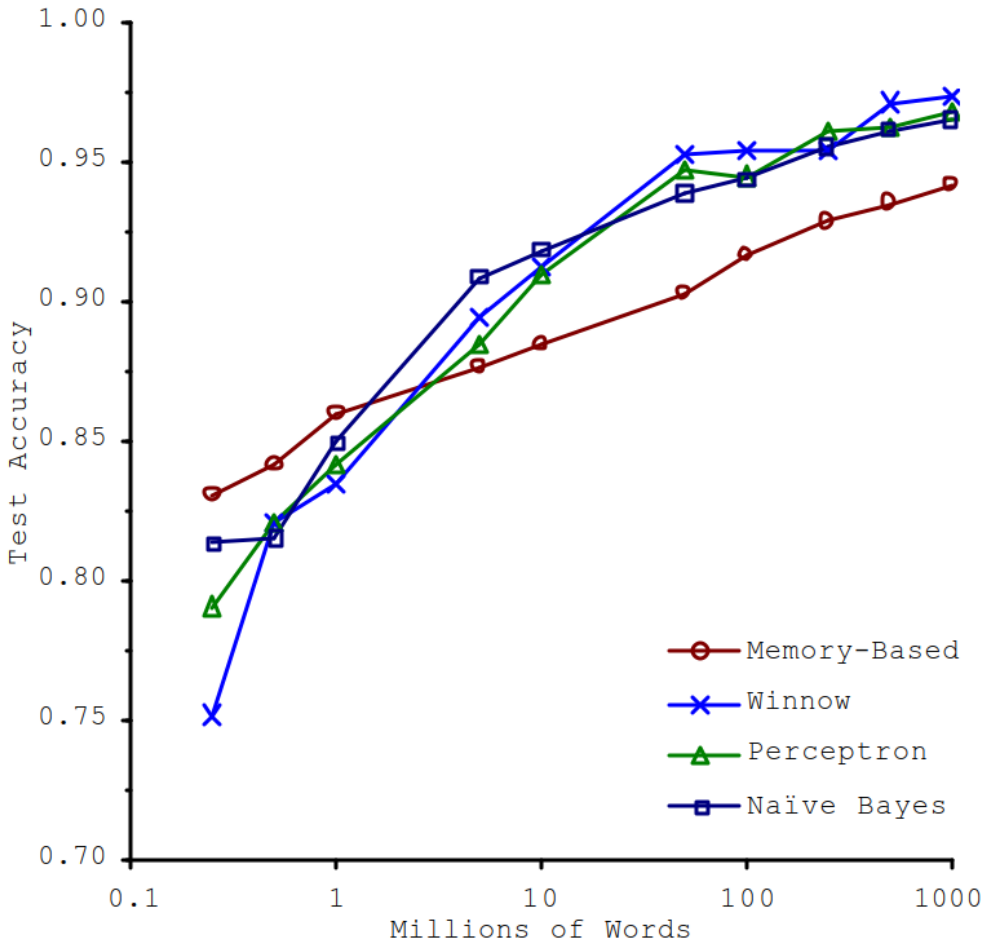
B. Documents d5 and d6 are closer to each other in the vector space than d5 and d4.

C. Documents d5 and d4 are as close to each other in the vector space as are d5 and d6.

D. TF-IDF(t3,d4) = 30.4

E. TF-IDF(t3,d4) = 36.8

4. The chart below shows the effectiveness of machine-learning algorithms for word disambiguation (y-axis), based on the number of words used for the training (x-axis). Which of the following interpretations are correct?



[4 marks]

A. The effectiveness of Memory-Based machine learning increases exponentially with the number of words being used for training.
B. The effectiveness of Memory-Based machine learning increases linearly with the number of words being used for training.
C. All algorithms are around three times more effective when trained with 1,000 million words compared to when trained with 1 million words.
D. As performance of all four algorithms increases, it would make sense to analyse how effective the four algorithms become when trained on even more data.
E. None of the listed answers applies.

5. You develop a machine learning application for a hospital that wants to predict the health status that a patient will have in a few days, i.e. “will be cured” or “will not (yet) be cured”. The input to that system would be some health-related data measured in the past couple of days. To design your system, you have a dataset with 85,988 instances of health data for 2,812 patients, and each patient is identified by a unique ID. For each patient, the dataset contains the following data, recorded on multiple days: patient’s body temperature, the time at which the temperature was taken (either at 8:00 or 10:00 or 17:00), the blood pressure, and 82 more independent variables. For each instance, the dataset contains a class label “1 = cured” or “2 = not yet cured”. However, you were informed that the thermometer that measured the temperature was incorrectly calibrated for 798 patients. For those patients, temperatures are 0.53°C too high. You do not know, which of the patients are affected by the incorrectly calibrated thermometer. Which of the following statements are true?

[5 marks]

A. The 0.53°C error will have no effect on the predictions because it is very small (<5%) compared to the typical temperature of a human (35° - 41°C).
B. Given that humans generally tend to have higher temperatures in the evening than in the morning, it is a poor methodology to measure temperatures at different times of the day. This may negatively impact the performance of predicting the health status.

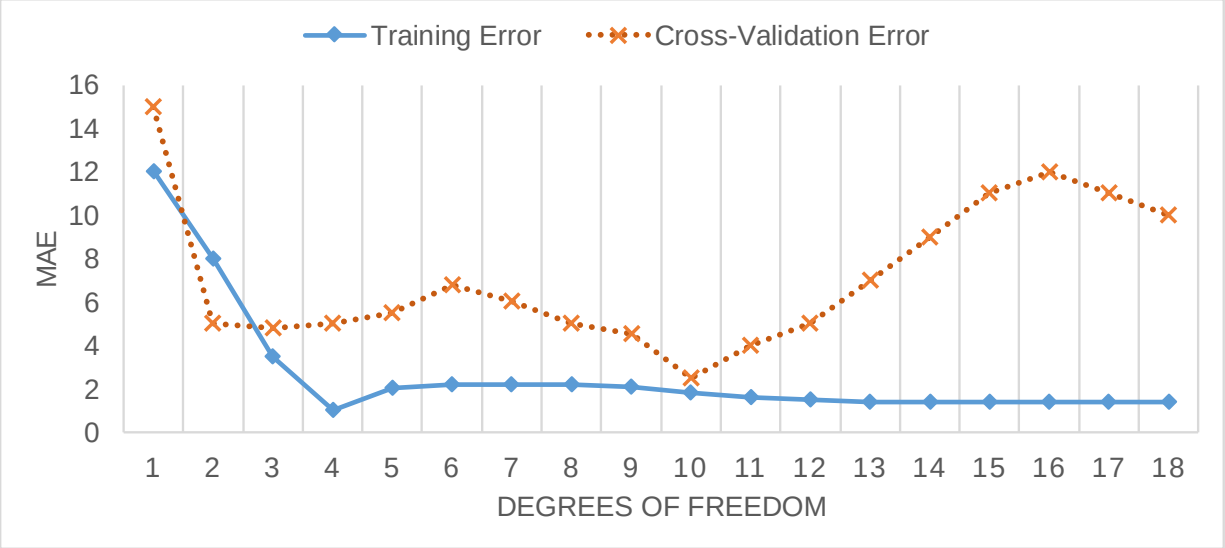
C. Since temperature is only one of many features, the error in temperature is not important and can be ignored.

D. Measuring temperatures at different times of the day is ok, because the time at which the temperature was measured is known, too.

E. None of the above

6. The chart below shows both the training error and testing error (MAE) for a machine-learning model on the y-axis, trained and evaluated with different degrees of freedom (x-axis). The evaluation method was 20-fold cross validation. What is the ideal degree of freedom for that model (assuming that training and testing was performed on a large dataset and all results are statistically significant)?

[2 mark]



A. 2 degrees of freedom

B. 3 degrees of freedom

C. 4 degrees of freedom

D. 10 degrees of freedom

E. 16 degrees of freedom

7. The table below shows the job experience and income for ten employees as well as some additional statistics (min, max, mean etc.). What is the “mean” for the standardized income for all ten persons?

	Job Experience (in Years)	Income
Person 1	36	97,080 €
Person 2	3	22,514 €
Person 3	22	88,233 €
Person 4	6	51,251 €
Person 5	28	65,978 €
Person 6	34	78,582 €
Person 7	14	86,429 €
Person 8	24	85,934 €
Person 9	21	35,139 €
Person 10	21	94,090 €

Min	3	22,514 €
Max	36	97,080 €
Mean	20.90	70,523 €
Median	21.50	82,258 €
Std Dev	10.25	24,655 €

[2 marks]

A. -1

B. 0

C. 1

D. 1.47

E. 20.90

8. A group of researchers developed a novel machine learning algorithm that predicts customers' gender ("male" or "female"), based on some independent variables. The researchers based their work on a dataset with 10,000,000 customers of which 4,996,540 were female and the remaining were males. The dataset was split into a random 80/20 split for training and testing. The training set had 8,000,000 customers of which 3,981,232 were female. The researchers report that their algorithm achieved an Area Under the Curve (AUC) of 0.42, while a state-of-the-art algorithm only achieved an AUC of 0.38, and the difference was statistically significant with $p<0.01$. Which of the following statements are correct?

[5 marks]

A. Even though the AUC is rather low for both algorithms, the novel algorithm seems to have a promising performance.
B. Using AUC is not appropriate with the given data. AUC can only be used for measuring the performance of a binary classifier ("positive", "negative") and not for classifying "equal" classes such as "male" and "female".
C. Such low values for AUC do not make sense. Something must be wrong in the training or evaluation.
D. The methodology seems flawed because the test set had (relatively) less female customers than the training set.
E. The researchers misinterpreted the results. A lower AUC is better than a higher AUC, hence the novel algorithm performed worse and not better than the state-of-the-art.

9. Which statements about traditional A/B testing and interleaving are true in the scenario of a recommender system?

[3 marks]

A. A traditional A/B test selects randomly an algorithm to give recommendations. This means, users see either recommendations based on algorithm A *or* B.

B. There is no difference between the two methods.

C. Interleaving calculates for each request recommendations with two (or more) algorithms, and then merges the algorithm´s results into one result set that is shown to users. This means, all users see recommendations based on both algorithm A *and* algorithm B.

D. In A/B testing, every user receives two variations (A and B) and the system measures, which variation receives better feedback. For instance, the system shows two lists of recommendations, one created with algorithm A, one created with algorithm B.

E. None of the above

10. In k-fold cross validation, the dataset is split into k folds. To calculate the final metrics such as accuracy or precision for all folds, there are two options to average the results. Given the confusion matrix below, what are the two possible final values for *accuracy*? Note that the third fold has one instance more than the other two folds.

	Number of Instances	TP	TN	FP	FN
Fold 1	10	6		4	-
Fold 2	10	7	3	-	-
Fold 3	11	6	2	1	2

[6 marks]

A. 0.776 for both

B. 0.778 for both

C. 0.600 and 0.730

D. 0.730 and 0.774

E. **0.776 and 0.774**