Example Exam. The final exam will have equal marks for Section A and B, and Section B will have 11 questions.

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 may 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.

a. Explain what a confidence interval is. [5 marks]

b. Describe how you would use cross-validation to estimate a confidence interval for the predictions made by a machine learning algorithm. [5 marks]

c. When using k-fold cross-validation, what are the trade-offs involved in selecting a smaller/larger value of k ? [5 marks]

Suppose you train a SVM using training data which has 3 input features.

d. Discuss how the number of features used can affect underfitting/over-fitting of the data. [5 marks]

e. Describe how cross-validation can be used to decide which input features to retain in the model. [5 marks]

2. Suppose you train a linear regression model.

a. Describe the hypothesis/predictive model used and its parameters. [5 marks]

b. Explain how it can be fitted to the training data. [5 marks]

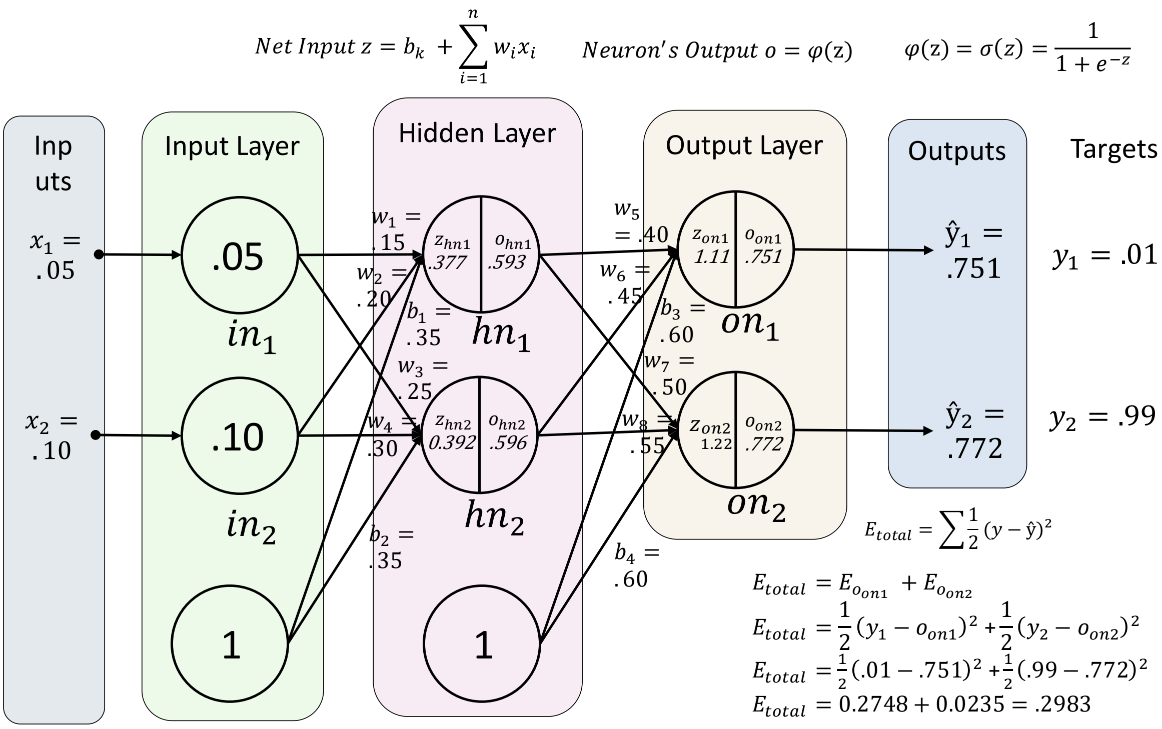
c. You now add a regularisation/penalty term $\lambda\theta^T\theta$ to the cost function, giving a ridge regression model. How does increasing/decreasing the penalty weighting λ affect the model parameters ? [5 marks]

d. Describe the probabilistic interpretation of a linear regression model, in particular (i) the assumptions made as to the observation noise and (ii) how the ridge regression cost function is derived from Bayes Rule. [10 marks]

Section B: Multiple Choice

1. The image below shows a neural network with one hidden layer, some random weights, the inputs x1 and x2 and two output neurons. Using backpropagation with a learning-rate η=0.2, what would be the new w_5?

[7 marks]



$$Net\ Input\ z = b_k + \sum_{i=1}^{n} w_i x_i \qquad Neuron's\ Output\ o = \varphi(\mathrm{z}) \qquad \varphi(\mathrm{z}) = \sigma(z) = \frac{1}{1+e^{-z}}$$

$$E_{total} = \sum \frac{1}{2}(y-\hat{y})^2$$

$$E_{total} = E_{o_{on1}} + E_{o_{on2}}$$

$$E_{total} = \frac{1}{2}(y_1 - o_{on1})^2 + \frac{1}{2}(y_2 - o_{on2})^2$$

$$E_{total} = \frac{1}{2}(.01 - .751)^2 + \frac{1}{2}(.99 - .772)^2$$

$$E_{total} = 0.2748 + 0.0235 = .2983$$

A. $w_5 = 0.01$

B. $w_5 = 0.384$

C. $w_5 = 0.397$

D. $w_5 = 0.40$

E. $w_5 = 0.666$

2. What is the term “out-of-sample” data referring to?

[2 marks]

A. All historic data that is available in e.g. a company´s database

B. The historic data that is available in e.g. a company´s database but was neither used for training nor testing the machine-learning model.

C. Outliers in the dataset.

D. Data that was explicitly removed from the dataset (e.g. instances with missing values)

E. None of the listed options

3. People with low income tend to provide their postal address in surveys less often than people with higher income. This means, for low-incomes, you will more often have missing addresses in your machine-learning dataset. What kind of missing data is this?

[2 marks]

A. The data is “Not Missing at Random” (NMAR) / “Missing Not at Random” (MNAR).

B. The data is “Missing at Random” (MAR) / “Missing Conditionally at Random” (MCAR).

C. The data is “Missing Completely at Random” (MCAR)

D. The data is “Missing for a Reason” (MFAR)

E. None of the above

4. The table below shows the job experience and income for ten employees as well as some additional statistics (min, max, mean etc.).

	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 €

If you normalized the job experience of all the persons in the table, what would be the normalized job experience of person 2?

[3 marks]

A. -1

B. 0

C. 1

D. 1.47

E. 12

5. IMDB is a website on which users can rate movies on a scale between 1 and 10. Your goal is to develop a machine learning model that predicts a movie's average rating, based on the following input features: average rating of a particular movie, number of users who rated the movie, release year, budget, names of the actors, number of actors' Facebook likes, and the name(s) of the director(s). The target variable is the average rating. You have access to a dataset with 5,000 movies that were released between 2010 and 2016. 500 of these movies have not received any ratings, the others have received at least one rating. You train and test two models A and B and find that their RMSEs are 1.02 (Model A) and 0.82 (Model B) respectively. Which of the following statements are true?

[2 mark]

A. Something must be wrong (RMSE cannot be larger than 1).

B. RMSE is an inappropriate metric in this scenario.

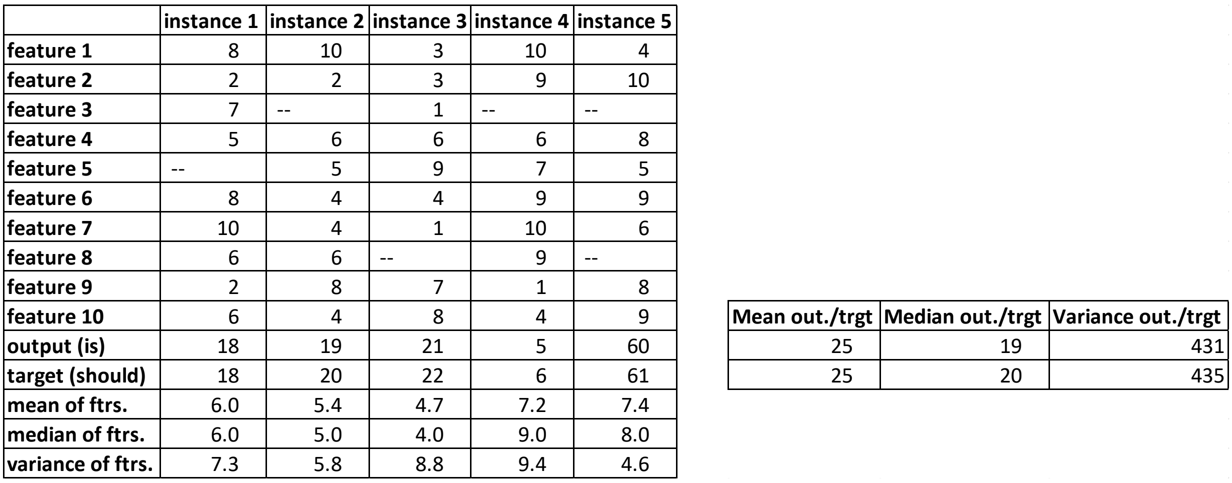
C. Model A has a larger RMSE than Model B, and hence performs worse than model B.

D. The gold standard for movie recommendations is the MovieLens dataset. Hence, even if ratings for IMDB should be predicted, the evaluation should be based on the gold standard (MovieLens).

E. Model A has a larger RMSE than the baseline, and hence performs better than Model B.

6. You trained a supervised machine learning algorithm on a dataset with 50 instances. Each instance has 10 numerical features and one numerical target. You evaluate the model on a test dataset with 5 instances. The table below shows the instances and features, as well as the target value for each instance, the output predicted by the machine learning model, and a few summary statistics (mean, median, …). What is the RMSE of the model?

[6 mark]



	instance 1	instance 2	instance 3	instance 4	instance 5
feature 1	8	10	3	10	4
feature 2	2	2	3	9	10
feature 3	7	--	1	--	--
feature 4	5	6	6	6	8
feature 5	--	5	9	7	5
feature 6	8	4	4	9	9
feature 7	10	4	1	10	6
feature 8	6	6	--	9	--
feature 9	2	8	7	1	8
feature 10	6	4	8	4	9
output (is)	18	19	21	5	60
target (should)	18	20	22	6	61
mean of ftrs.	6.0	5.4	4.7	7.2	7.4
median of ftrs.	6.0	5.0	4.0	9.0	8.0
variance of ftrs.	7.3	5.8	8.8	9.4	4.6

Mean out./trgt	Median out./trgt	Variance out./trgt
25	19	431
25	20	435

A. $RMSE \approx 0.01$

B. $RMSE \approx 0.3$

C. $RMSE \approx 0.6$

D. $RMSE \approx 0.9$

E. $RMSE \approx 3$

7. You got a new job as machine learning engineer in a medical clinic that screens patients for skin cancer. The clinic wants to use machine learning to predict whether a patient will survive or not, once the patient is diagnosed with cancer. The clinic has historic data from 20 years and 1,579,999 screened patients of which 11,059 were diagnosed with cancer (0.7%). Of the 11,059 cancer patients, 54.78% survived, and 41.77 % died within 5 years after the diagnosis. For 3.45% of the patients it is not known whether they survived or not. The clinic has comprehensive data about the patients such as age, skin type, gender, height, weight, the date when the cancer was diagnosed, when and how often they received chemo therapy and how many days after the diagnosis the patients died (or "-1" if the patient survived). The senior machine-learning engineer suggests treating the problem as classification problem ("survive" or "not survive" within 5 years). She has run some tests on the data already and found that a Support Vector Machine performed best, achieving an accuracy of 89.83% in predicting whether a patient will survive or not. The next best algorithm only had an accuracy of 69.48%. Which of the following statements are true?

[6 marks]

A. The accuracy of 89.83% is rather meaningless because only 0.7% of all screened patients were diagnosed with cancer.
B. Something must be wrong. A support vector machine cannot be used for classification.
C. Using data from the past 20 years to predict survival chances may be problematic given the advances in cancer diagnosis and treatment in the past two decades.
D. Whenever possible, a problem should be treated as regression problem rather than a classification problem. This means, the clinic should predict the number of years (or days) after which a patient would die, instead of predicting the binary class "survive" or "not survive".
E. A Support-Vector-Machine cannot properly be trained and tested with only 11,059 instances.

8. You have four documents in a corpus, defined by the following term-frequency vectors

$$d_1 = \{TF(t_1) = 3, TF(t_2) = 5, TF(t_3) = 10\}$$

$$d_2 = \{TF(t_1) = 0, TF(t_2) = 2, TF(t_3) = 1\}$$

$$d_3 = \{TF(t_1) = 18, TF(t_2) = 0, TF(t_3) = 100\}$$

$$d_4 = \{TF(t_1) = 0, TF(t_2) = 5, TF(t_3) = 18\}$$

What is $\text{TFIDF}(d_1, t_3)$?

[5 marks]

A. $\text{TFIDF}(d_1, t_3) = 0.5$

B. $\text{TFIDF}(d_1, t_3) = 1.1$

C. $\text{TFIDF}(d_1, t_3) = 12.7$

D. $\text{TFIDF}(d_1, t_3) = 63$

E. None of the above.

9. You and a friend read the following news article from the New York Times:

> *Every year, scores of cats fall from open windows in New York City. From June 4 through Nov. 4, 1984, for instance, 132 such victims were admitted to the Animal Medical Center on 62d Street in Manhattan. Most of the cats landed on concrete. Most survived. Experts believe they were able to do so because of the laws of physics, superior balance and what might be called the flying-squirrel tactic. In a study for the medical center, Dr. Wayne Whitney and Dr. Cheryl Mehlhaff recorded the distance of the fall for 129 of the 132 cats. The falls ranged from 2 to 32 stories, with an average distance of 5.5 stories. Two cats fell together. About a quarter fell during daylight hours, and about 40 percent at night. For the rest, the time of the fall was unknown. Three cats were seen falling by their owners. Two were described as having fallen while turning on a narrow ledge, and the third had lunged for an insect. Seventeen of the cats were put to sleep by their owners, in most cases not because of life-threatening injuries but because the owners said they could not afford medical treatment. Of the remaining 115, 8 died from shock and chest injuries. Even more surprising, the longer the fall, the greater the chance of survival. Only one of 22 cats that plunged from above 7 stories died, and there was only one fracture among the 13 that fell more than 9 stories. The cat that fell 32 stories on concrete, Sabrina, suffered*

> *a mild lung puncture and a chipped tooth. She was released from the hospital after 48 hours. The cat's ability to twist around while falling and land on its feet is well known. But why did cats from higher floors fare better than those on lower ones? One explanation is that the speed of the fall does not increase beyond a certain point, Dr. Mehlhaff and Dr. Whitney said in the December 1987 issue of The Journal of the American Veterinary Medical Association. This point, "terminal velocity," is reached relatively quickly in the case of cats. Terminal velocity for a cat is 60 miles per hour; for an adult human, 120 m.p.h. Until a cat reaches terminal velocity, the two speculated, the cat reacts to acceleration by reflexively extending its legs, making it more prone to injury. But after terminal velocity is reached, they said, the cat might relax and stretch its legs out like a flying squirrel, increasing air resistance and helping to distribute the impact more evenly. "Cats may be behaving like well-trained paratroopers," Dr. Jared Diamond, who teaches physiology at the University of California at Los Angeles Medical School, wrote in the August issue of the magazine Natural History.*

Your friend tells you that the clinic gave him access to their data, and he had trained a neural network to predict the survival chance of cats falling from a building. He used a 10-fold cross validation, and found that his neural network achieved an accuracy of 91% in predicting whether a cat would survive or not. Your friend claims that if he got data (time of the day, ground material, height, reason for fall, …) for a cat that would currently fall from a building, his model should achieve a similar accuracy. What would be your response? For your answer, ignore the rather low sample size.

[3 marks]

A. The friend is probably right, and the model should be able to predict the survival chance of an actually falling cat with a high accuracy.
B. The ground truth is flawed, and the machine learning model will probably perform quite worse with out-of-sample cats.
C. The gold standard is flawed, and the machine learning model will probably perform quite worse with out-of-sample cats.
D. Accuracy is not an appropriate metric in this scenario.
E. Using 10-fold cross validation seems sensible in the given scenario.