Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

# SEMESTER 1 EXAMINATIONS 2022/2023

**MODULE:** EE514 - Data Analysis and Machine Learning

**PROGRAMME(S):**

| | |
|---|---|
| MECE | MEng Electronic & Computer Engineering |
| MCTY | MSc Electronic and Computer Technology |
| MSAR | MSc in Astrophysics and Relativity |
| CAPT | PhD-track |
| PHPM | MSc |
| MQTY | Qualifier Prog MSc Electronic & Computer |
| MEPT | PhD-track |

**YEAR OF STUDY:** 1,2,3,C

**EXAMINERS:**

| | | |
|---|---|---|
| Dr. Kevin McGuinness | (Internal) | (Ext: 5133) |
| Prof. Roberto Verdone | (External) | External |

**TIME ALLOWED:** 3 hours

**INSTRUCTIONS:** **Answer 4 questions. All questions carry equal marks.**

---

**PLEASE DO NOT TURN OVER THIS PAGE UNTIL INSTRUCTED TO DO SO**
The use of programmable or text storing calculators is expressly forbidden.
Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

---

***There are no additional requirements for this paper.***

***QUESTION 1*** ***[TOTAL MARKS: 25]***

*Data interpretation, storage, and summarization*

**Q 1(a)** **[6 Marks]**

The table below shows a sample of some rows from an online retail dataset:

| InvoiceNo | Stock Code | Quantity | Invoice Date | Unit Price | Seller Rating |
|---|---|---|---|---|---|
| 537126 | 22722 | 12 | 02/12/2010 10:39 | 1.25 | 5 |
| 536404 | 20957 | 1 | 01/12/2010 15:40 | 120.00 | 3 |
| 537190 | 85224 | 5 | 15/09/2011 16:37 | 12.95 | 4 |
| 537812 | 72741 | 1 | 15/09/2011 17:07 | 25.50 | 2 |
| 535659 | 12554 | 2 | 27/11/2011 15:02 | 3.95 | 5 |

For each column in the table, specify which of **Steven's scales of measurement** it could be classified as and explain your reasoning.

**Q 1(b)** **[6 Marks]**

Suppose that you performing some data wrangling on the full dataset from the previous question and discovered there are some **missing values**. Explain what the different categories of missing values are and any special care that would need to be taken in handling these values.

**Q 1(c)** **[6 Marks]**

Write down the formula for the fourth standardized moment and explain what this is commonly used to measure and how it can be estimated from a sample in practice.

**Q 1(d)** **[7 Marks]**

The CSV and HDF5 formats are both commonly used to store tabular data. Describe each of these file formats and explain their advantages and disadvantages. Give an example of a situation where you would choose to use HDF5 over CSV.

***[End Question 1]***

***QUESTION 2*** ***[TOTAL MARKS: 25]***

*Supervised learning*

**Q 2(a)** **[6 Marks]**

In supervised learning, the initial dataset is often divided into three subsets: training, validation, and test. Explain the purpose of each of these subsets. Describe two precautions that should be taken when splitting the data into these subsets.

**Q 2(b)** **[6 Marks]**

The **bias-variance decomposition** tells us that expected squared error can be decomposed as:

$$\mathbb{E}[(y - \hat{y})^2] = \sigma^2 + \mathbf{var}(\hat{y}) + \mathbf{bias}^2(y, \hat{y}),$$

where $y$ is the true target value and $\hat{y}$ is the predicted target value. Explain, with the aid of an example or a diagram, what the meaning of each of the terms in the decomposition is.

**Q 2(c)** **[6 Marks]**

Explain what the difference is between a **parametric** classifier and a **non-parametric** classifier. Give an example of each.

**Q 2(d)** **[7 Marks]**

In the linear discriminant analysis (LDA) model, the probability of a data point $\mathbf{x}$ given the class $y = 0$ is given by:

$$p(\mathbf{x} \mid y = 0) = \frac{1}{Z} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_0)^T \Sigma^{-1} (\mathbf{x} - \mu_0)\right),$$

and the probability $x$ given $y = 1$ is:

$$p(\mathbf{x} \mid y = 1) = \frac{1}{Z} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma^{-1} (\mathbf{x} - \mu_1)\right).$$

where in both cases $Z$ is a normalizing constant that depends only on the covariance matrix $\Sigma$. The decision boundary is the set of all points $\mathbf{x}$ that satisfy

$$p(\mathbf{x} \mid y = 0) = p(\mathbf{x} \mid y = 1).$$

Show that the decision boundary for an LDA model is linear.

***[End Question 2]***

***QUESTION 3*** ***[TOTAL MARKS: 25]***

*Linear and logistic regression*

**Q 3(a)** **[5 Marks]**
Explain what is meant by **regularization** in machine learning and why regularization is used. Describe one commonly used form of regularization.

**Q 3(b)** **[8 Marks]**
The loss function in a ridge regression model can be written as:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - X\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2,$$

where $X \in \mathbb{R}^{N\times(D+1)}$ is a design matrix.

1. If $D = 20$ and $N = 1000$, how many parameters does this model have?
2. Write down the prediction function for the linear model.
3. Show that the minimizer of $\mathcal{L}(\mathbf{w})$ is the solution to the linear system $(X^TX + \lambda I)\mathbf{w} = X^T\mathbf{y}$.
4. Explain what the purpose of $\lambda$ is in ridge regresson and how it could be chosen in practice.

**Q 3(c)** **[6 Marks]**
In logistic regression the target variable is binary and the prediction function can be written as:

$$f(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}).$$

What function is $\sigma(\cdot)$ in logistic regression? What loss function is typically used to fit logistic regression models? Explain why this loss function is used.

**Q 3(d)** **[6 Marks]**
Explain how logistic regression can be extended to handle categorical targets.

***[End Question 3]***

***QUESTION 4*** ***[TOTAL MARKS: 25]***

*Decision trees and ensemble methods*

**Q 4(a)** **[6 Marks]**
Explain, in your own words, the key idea behind ensemble methods.

**Q 4(b)** **[6 Marks]**
Why are decision trees and decision stumps often used as the base classifier in ensemble methods?

**Q 4(c)** **[6 Marks]**
Show that the variance of the average of $n$ uncorrelated predictors $Y_1, Y_2, \ldots, Y_n$, each with variance $\sigma^2$ is

$$\mathbf{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{\sigma^2}{n}.$$

**Q 4(d)** **[7 Marks]**
Explain what is meant by *bagging* in the context of ensemble methods. How does bagging help to reduce variance? Describe how **random forests** extend bagging and explain why bagging is usually less effective than the random forest approach.

***[End Question 4]***

***QUESTION 5*** ***[TOTAL MARKS: 25]***

*Question*

**Q 5(a)** **[5 Marks]**

Describe, with the aid of an example, what $2 \times 2$ **max pooling** layer with stride 2 does. Provide TWO reasons why max pooling layers are often used in deep neural networks.

**Q 5(b)** **[10 Marks]**

Given the following decision function for a three layer fully connected neural network:

$$f(\mathbf{x}) = W_3\ g(W_2\ g(W_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2) + \mathbf{b}_3,$$

where $g = \max\{x, 0\}$ is the activation function, and suppose that:

$$W_1 \in \mathbb{R}^{75\times 15}, \quad W_2 \in \mathbb{R}^{50\times 75}, \quad W_3 \in \mathbb{R}^{1\times 50}$$

1. What must the dimensions of $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ be?
2. What is the dimension of the input of the network $\mathbf{x}$?
3. What is the dimension of the output of the network $f(\mathbf{x})$?
4. How many parameters does this model have?
5. Would this network be more suited to a regression problem or a classification problem? Explain why.

**Q 5(c)** **[10 Marks]**

he Elman RNN can be described by two equations:

$$\mathbf{s}_t = g_s(W_s\mathbf{s}_{t-1} + V_s\mathbf{x}_t + \mathbf{b}_s)$$
$$\mathbf{y}_t = g_y(W_y\mathbf{s}_t + \mathbf{b}_y),$$

where $\mathbf{s}_t$ is the state vector at time $t$, $\mathbf{s}_{t-1}$ is the previous state vector, $\mathbf{x}_t$ is the input, and $\mathbf{y}_t$ is the output.

1. Assume the input $\mathbf{x}_t$ has dimension 100, the output $\mathbf{y}_t$ has dimension 5 and the hidden state $\mathbf{s}_t$ has dimension 50. Write down the sizes of the parameter matrices $W_s, V_s, W_y$ and the bias vectors $\mathbf{b}_s$ and $\mathbf{b}_y$.
2. What is the role of the function $g_s$?
3. If the RNN is to be used for a multiclass sequence classification, what kind of activation function would be an appropriate choice for $g_y$.
4. What kind of issues may be encountered if you were to train this kind of RNN in practice?

***[End Question 5]***

***[END OF EXAM]***