MSc Data Science and Artificial Intelligence admission level check
M2 intake

The MSc Data Science and Artificial Intelligence (DSAI) is a master's degree of excellence, requiring a solid grounding in programming, algorithms, statistics and probability.

To self-assess your level in these subjects, please complete the exercises on Python and Statistics. We recommend that you take the time to do these exercises, and above all that you do them yourself, without any help or use of Chat GPT.

If you struggle to complete these exercises, it may indicate that following the MSc DSAI program will be challenging and may lead to difficulties in succeeding. Therefore, please take this assessment seriously and submit your solutions within this document in a single PDF format. You fill find the exercices in the following pages.

EXERCISES IN STATISTICS

How did you find the exercices ?

Please tick a box.

Very easy	Easy	Medium	Hard	Very hard
☐	☐	☐	☐	☐

Comment if any :

EXERCISES IN MACHINE LEARNING

How did you find the exercices

Please tick a box.

Very easy	Easy	Medium	Hard	Very hard
☐	☐	☐	☐	☐

Comment if any :

EXERCISES IN DEEP LEARNING

How did you find the exercices ?
Please tick a box.

Very easy	Easy	Medium	Hard	Very hard
☐	☐	☐	☐	☐

Comment if any :

Admission M2 - MSc DSAI

Michel RIVEILL

December 2025

1 Statistics

Exercise 1 Let consider the following regression model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i, \;\; i = 1, \ldots, n$$

with $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}$, $p \in \mathbb{N}^*$ (x_i^p meaning x_i at the power p) and ε_i independent and identically distributed random variables, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

1. Write the model in matrix form $= X\beta + \varepsilon$
2. Give the distribution of
3. Is it still possible to fit a multiple regression model? (explain your answer)

Let assume that we have some real data represented in Figure 1.

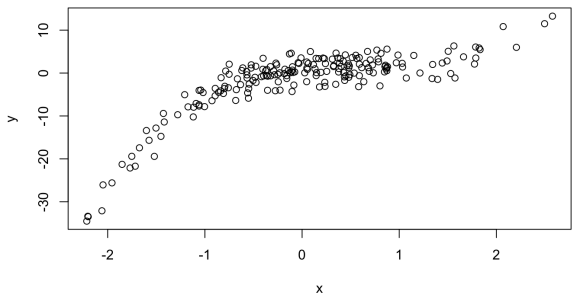

Figure 1: Scatter plot of the data

We now fit the model on $n = 200$ data with $p = 6$.
Then R give us the following summary

```
Call:
lm(formula = y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5), data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1389 -1.4593  0.1208  1.4576  4.6349

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.59138    0.20743   2.851  0.00483 **
x            3.29630    0.38984   8.455 6.61e-15 ***
I(x^2)      -2.81103    0.26406 -10.645  < 2e-16 ***
I(x^3)       0.63126    0.27186   2.322  0.02127 *
I(x^4)      -0.05581    0.04416  -1.264  0.20780
I(x^5)       0.10027    0.03294   3.044  0.00266 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.05 on 194 degrees of freedom
Multiple R-squared:  0.9313, Adjusted R-squared:  0.9295
F-statistic: 525.7 on 5 and 194 DF,  p-value: < 2.2e-16
```

4. Explain how are obtained the different values related to the coefficient of variable x (for each value give a detailed explanation)

   ```
   Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   x            3.29630    0.38984   8.455 6.61e-15 ***
   ```

5. Give a 95% confidence interval on β_1. (justify why it is not mandatory to use Student distribution here)

6. Based on the summary :

   (a) Give the value of s

   (b) Give the value of R^2

   (c) What test is performed in the last line and what is your conclusion?

   ```
   F-statistic: 525.7 on 5 and 194 DF,  p-value: < 2.2e-16
   ```

7. What is the advantage of using the adjusted R^2 compared with using the R^2?

The value of $(X'X)^{-1}$ is the following:

	(Intercept)	x	I(x^2)	I(x^3)	I(x^4)	I(x^5)
(Intercept)	1.024112e-02	1.276427e-03	-0.0088421548	-1.053081e-03	1.103871e-03	9.738012e-05
x	1.276427e-03	3.617186e-02	-0.0015977162	-2.220144e-02	9.748356e-06	2.388357e-03
I(x^2)	-8.842155e-03	-1.597716e-03	0.0165956557	1.733767e-03	-2.474542e-03	-1.303479e-04
I(x^3)	-1.053081e-03	-2.220144e-02	0.0017337667	1.759120e-02	4.018354e-06	-2.056922e-03
I(x^4)	1.103871e-03	9.748356e-06	-0.0024745419	4.018354e-06	4.641252e-04	-2.398747e-05
I(x^5)	9.738012e-05	2.388357e-03	-0.0001303479	-2.056922e-03	-2.398747e-05	2.582698e-04

8. Based on this matrix and (on previous information ...) give the estimated covariance matrix of the random vector composed with $\hat{\beta}_1$ and $\hat{\beta}_2$
9. Thus deduce the estimated variance of $\beta_1 - \beta_2$
10. Thus propose a way to test $H_0 : \beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$ and give your conclusion
11. If you want to test if the order of the consider polynomial is $p = 3$ versus $p = 5$ what test could you perform ? (explain just the method, don't give the result of the test.

Exercise 2 Let consider the following dataset where we aim at predicting the class (die or live) based on the age, the sex, and the fact that the people take steroid or not. Here is a summary of the data

```
 class          age            sex      steriod
die : 32   Min.   : 7.00   female: 16   no :76
live:122   1st Qu.:32.00   male  :138   yes:78
           Median :39.00
           Mean   :41.27
           3rd Qu.:50.00
           Max.   :78.00
```

We fit a logistic regression model, and we obtain the following summary

```
Call:
glm(formula = class ~ age + sex + steriod, family = "binomial",
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1324   0.0001   0.5150   0.7322   1.2147

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   20.40283 1543.00813   0.013  0.98945
age           -0.04692    0.01773  -2.646  0.00815 **
sexmale      -17.48818 1543.00786  -0.011  0.99096
steriodyes     0.65788    0.42650   1.543  0.12295
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157.39  on 153  degrees of freedom
Residual deviance: 138.86  on 150  degrees of freedom
```

```
AIC: 146.86

Number of Fisher Scoring iterations: 17
```

1. Based on the fitted model give the probability of death for this person, and also give its predicted class

   ```
   age    sex steriod
    30 female      no
   ```

2. All being equal, are men more at risk that woman?
3. Give the estimated odds-ratio for the risk of death between male and women.
4. Based on the output, give the value of the likelihood
5. Would you keep this fitted model? (justify your answer)

We now fit the model only why the age variable and get the following results

```
Call:
glm(formula = class ~ age, family = "binomial", data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0192   0.4589   0.5717   0.7069   1.2410

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.17880    0.75622   4.204 2.63e-05 ***
age         -0.04266    0.01623  -2.629  0.00857 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157.39  on 153  degrees of freedom
Residual deviance: 150.18  on 152  degrees of freedom
AIC: 154.18

Number of Fisher Scoring iterations: 4
```

6. Based on the available information what test could you perform to choose between the first and the second model? (give the value of the test statistic considered)
7. From statistical information criterion point of view, which model would you keep?

2 Machine Learning

Support Vector Machine Consider a supervised learning problem in which the training examples are points in a 2-dimensional space. The positive examples are (1,1) and (-1,-1). The negative examples are (1,-1) and (-1,1).

Question 1. Are the positive examples linearly separable from the negative examples in the original space? Illustrate your answer with a figure.

Question 2. Consider the feature transformation

$$\phi(x) = [1, x_1, x_2, x_1 x_2]$$

, where x1 and x2 are, respectively, the first and second coordinates of a generic example x. The prediction function is

$$y(x) = w^T * \phi(x)$$

in this feature space. Give the coefficients, w of a maximum-margin decision surface separating the positive examples from the negative examples. (You should be able to do this by inspection, without any significant computation).

Using the new coordinates, draw a figure to validate your w vector.

Question 3. Give the expression of the corresponding kernel K.

Question 4: Write the code that implements such SVM classification using sklearn.

Question 5. Considering another dataset of 10000 samples with 2 classes but with only 5 positively labeled samples and 5 negatively labeled samples (other samples are not labeled). How to perform active learning on such dataset? Describe in detail the necessary steps.

Decision Tree We use the following data set:

Train

	class	sex	age
1	First	female	38.0
2	Third	female	26.0
38	Third	female	18.0
23	First	male	28.0
34	First	male	28.0
37	Third	male	21.0

Test

	class	sex	age
3	First	female	35.0
4	Third	male	35.0
6	First	male	54.0
35	First	male	42.0
43	Second	female	3.0
67	Third	male	19.0

Figure 2: Train and test dataset

We have chosen to create 3 decision trees of depth 1. Each decision tree uses a single feature.

To do this, we use the following code:

```
1  dt = {}
2
3  fig, axis = plt.subplots(nrows=1, ncols=3, figsize=(10,4), dpi=300)
4
5  for i, f  in enumerate(X_train.columns):
6      str = ""
7      if X_train[f].dtypes!="object" and X_train[f].dtypes!="category":
8          # It's a numerical feature
9          preprocessor = Pipeline(steps=[
10             ('imputer', 'passthrough'),
11             ('scaler', 'passthrough'),
12         ])
13     else:
14         # It's a categorical feature
15         preprocessor = Pipeline(steps=[
16             ('imputer', 'passthrough'),
17             ('encoder', 'passthrough')),
18         ])
19
20     pipeline = Pipeline(steps=[
21         ('preprocessor', preprocessor),
22         ('classifier', DecisionTreeClassifier(criterion='gini', max_depth=1))
23     ])
24
25     dt[f] = pipeline.fit(X_train[f].to_numpy().reshape(-1,1),
26                          y_train)
27
28     if X_train[f].dtypes=="object" or X_train[f].dtypes=="category":
29         str = dt[f]['preprocessor']['encoder'].categories_[0]
30
31     tree.plot_tree(dt[f]['classifier'], feature_names=[f], class_names=['0', '1'], ax = axis[i])
32     axis[i].set_title(f"feature {f}\n{str}", fontsize = 12)
```

Figure 3: Code

Question 1: Considering the use of a library conforming to the decision tree model. Is this code correct? Justify your answer, particularly with regard to preprocessing and decision tree parameters (lines 25 to 32 are correct).

If you think that the sklearn library requires one or more adaptations, please specify.

Question 2: Could you complete the following pictures (use response sheet):

Question 3: for each tree, predict and predict_proba for the test samples.

Question 4: Explaining the different steps of the calculation, give the predicted class for the two test observations, if we use the "hard" vote and the "soft" vote.

3 Deep Learning

Question 1 You are training a large feedforward neural network (100 layers) on a binary classification task, using a sigmoid activation in the final layer, and a mixture of tanh and ReLU activations for all other layers. You notice your weights to a subset of your layers stop updating after the first epoch of training, even though your network has not yet converged. Deeper analysis reveals the

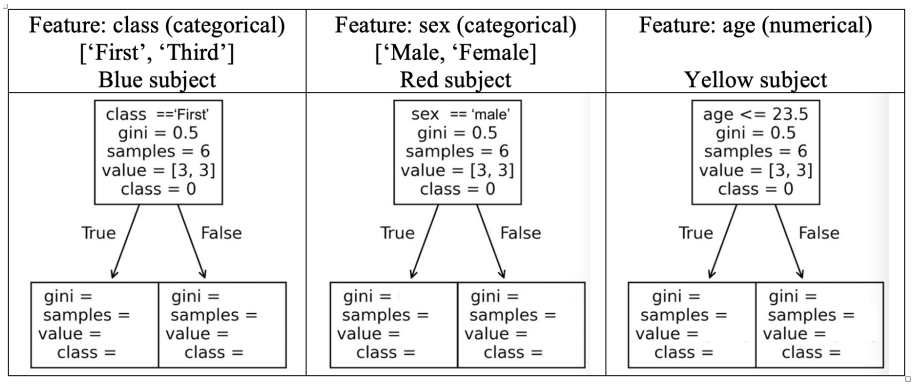

Figure 4: Trees

gradients to these layers completely, or almost completely, go to zero very early on in training. Which of the following fixes could help? (You also note that your loss is still within a reasonable order of magnitude). i. Increase the size of your training set ii. Switch the ReLU activations with leaky ReLUs everywhere iii. Add Batch Normalization before every activation iv. Increase the learning rate What's the problem? What is the best way of dealing with it? Do you know of any other techniques for resolving the problem?

Question 2 What properties are expected of an activation function? Which of the following would you consider to be valid activation functions (elementwise non-linearities) to train a neural net in practice?

$$f(x) = min(2, x)$$

$$f(x) = 0.9x + 1$$

$$f(x) = \begin{cases} min(x, 0.1x) & \text{if } x \geq 0 \\ max(x, 0.1x) & \text{if } x < 0 \end{cases}$$

$$f(x) = \begin{cases} max(x, 0.1x) & \text{if } x \geq 0 \\ min(x, 0.1x) & \text{if } x < 0 \end{cases}$$

Question 3 A friend recommends you use the following non-linearity activation function on your convolutional neural network with the Adam optimizer.

$$f(x) = \begin{cases} max(x, 0.1x) & \text{if } x >= 0 \\ min(x, 0.1x) & \text{if } x < 0 \end{cases}$$

Would you follow their advice? Why or why not?

Question 4 You are designing a deep learning system to diagnose chest cancer through X-ray images. What do you think might be the most appropriate evaluation metric and why: Accuracy, Precision, Recall, F1 score.

Question 5 You are training on a GAN to generate cool octopus images. But you are only able to curate a small number of images of actual octopus. You decide that image augmentation might be a good idea to improve GAN training. You try using three common augmentation techniques: adding blur to the image, changing the color of pixels and flipping the left-right axis of the image. Which of these do you think might be a good idea to help the Generator output better octopus images and why?

Question 6 You are training a single-layer, feedforward neural network with a softmax activation function in the final layer to classify among 99 classes, with a cross-entropy loss training objective. Recall, the cross-entropy loss function: $L(y, \hat{y}) = -y.log(\hat{(}y)$ where y is the one-hot encoded label, and ŷ is the predicted probability distribution over labels.

You decide to independently sample your initial weights from a Gaussian of mean 0, standard deviation 0.0001. You can assume perfect class balance in dataset. You accidentally set a 0 learning rate. What would you expect your average loss after the first training epoch to be? Provide a brief explanation (1-2 sentences) as to why this is so. You don't have to calculate the exact numerical value - feel free to leave your answer as a fraction.

Question 7 Consider the convolutional neural network defined by the layers in the left column below. Fill in the shape of the output volume and the number of parameters at each layer. You can write the activation shapes in the format (H, W, C), where H, W, C are the height, width and channel dimensions, respectively. Unless specified, assume padding 1, stride 1 where appropriate. Notation:

- CONVx-N denotes a convolutional layer with N filters with height and width equal to x.
- POOL-n denotes a n×n max-pooling layer with stride of n and 0 padding.
- FLATTEN flattens its inputs, identical to torch.nn.flatten / tf.layers.flatten
- FC-N denotes a fully-connected layer with N neurons

Question 6 A recurrent neural network (RNN) takes an input vector x_t and a state vector h_{t-1} and returns a new state vector h_t and an output vector $y_t$with:

$$h_t = f(w_1.x_t + w_2.h_{t-1} + b_2)$$

$$y_t = g(w_3.h_t + b_3)$$

Layer	Output shape	Number of parameters
Input	32 × 32 × 3	0
CONV3-8		
Leaky ReLU		
POOL-2		
BATCHNORM		
CONV3-16		
Leaky ReLU		
POOL-2		
FLATTEN		
FC-10		

Figure 5: CNN

where f and g are activation functions.

The following diagram depicts a single RNN unit, where x_t, h_{t-1}, $h_t$and y_t are all scalars, as a state machine:

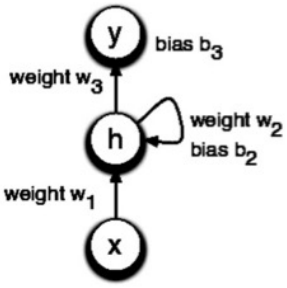

Figure 6: RNN

In this part:

- $h_0 = 0$
- $if x >= 0 then f(x) = 1 else f(x) = 0$
- $g(x) = x$

and the RNN initially outputs 0, but as soon as it receives an input of 1, it switches to outputting 1 for all subsequent time steps. For instance:

- the input 1010101 produces the output 1111111,
- the input 0101010 produces the output 0111111,
- the input 0001010 produces the output 0001111.

Several values of weight and bias are possible for the network to adopt such an operation, but in one of them, the hidden unit has a value of $h_{t-1} = 0$ until there is an input $x_t = 1$, after which it switches to an activation of 1 forever. The output unit always predicts the same thing as the hidden unit, i.e. $y_t = h_t$.

Calculate the different weights and biases w_1, w_2, w_3, b_2, b_3.