Higher Institute of Technological Studies of Bizerte

AY: 2022-2023
Exam | Machine Learning
03/01/23 (09:00→10:30)

L3-S5: Dept. of Electrical Engineering
Teacher: A. Mhamdi
Time Limit: $1\frac{1}{2}$ h

This document contains **6** pages numbered from **1/6** to **6/6**. As soon as it is handed over to you, make sure that it is complete. The **3** tasks are independent and can be treated in the order that suits you.

The following rules apply:

❶ **A handwritten double-sided A4** sheet is permitted.

❷ **Any electronic material**, except basic calculator, is prohibited.

❸ **Mysterious or unsupported answers** will not receive full credit.

❹ **Label all relevant aspects** of the graph, if you are asked to draw one.

❺ **Hand in your answer sheets** at the end of the exam.

Task N°1 ⏳ 45mn | ($13\frac{1}{2}$ points)

(a) A well-posed learning problem has been defined by Tom Mitchell as: « A computer program is said to learn from experience $\mathcal{E}$ with respect to some task $\mathcal{T}$ and some performance measure $\mathcal{P}$, if its performance on $\mathcal{T}$, as measured by $\mathcal{P}$, improves with experience $\mathcal{E}$. »

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam.

i. ($\frac{1}{2}$ point) What is the experience $\mathcal{E}$ in this setting?

- ◯ Classifying emails as spam or not spam.
- √ Watching you label emails as spam or not spam.
- ◯ The number (or fraction) of emails correctly classified as spam/not spam.

ii. ($\frac{1}{2}$ point) What is the task $\mathcal{T}$ in this setting?

- √ Classifying emails as spam or not spam.
- ◯ Watching you label emails as spam or not spam.
- ◯ The number (or fraction) of emails correctly classified as spam/not spam.

iii. ($\frac{1}{2}$ point) What is the performance measure $\mathcal{P}$ in this setting?

- ◯ Classifying emails as spam or not spam.
- ◯ Watching you label emails as spam or not spam.

√ The number (or fraction) of emails correctly classified as spam/not spam.

(b) (½ point) k-NN is a linear classifier.
○ True √ False

(c) (½ point) In the k-NN algorithm, do we need to specify the number of neighbors?
√ Yes ○ No

(d) (½ point) What is the default parameter for the number of neighbors k?
○ $k = 2$ ○ $k = 3$ √ $k = 5$ ○ $k = 10$

(e) (½ point) What is the purpose of splitting the datasets into the training set and the test set?
○ To prevent underfitting
√ To prevent overfitting
○ To prevent equalfitting
○ To make sure that all of the columns in our datasets are at the same scale before we apply the model

(f) (½ point) Logistic Regression is a linear classifier.
√ True ○ False

(g) (½ point) Your Machine Learning (ML) system is attempting to describe a hidden structure from unlabeled data. How would you describe this machine learning method?
○ supervised learning
√ unsupervised learning
○ reinforcement learning
○ semi-unsupervised learning

(h) (½ point) You want to create a supervised machine learning system that identifies pictures of kittens on social media. To do this, you have collected more than 100000 images of kittens. What is this collection of images called?
√ training data ○ linear regression ○ big data ○ test data

(i) (½ point) Which statement about K-means clustering is true?
○ To be accurate, you want your centroids outside of the cluster.
○ The number of clusters are always randomly selected.
√ In K-means clustering, the initial centroids are sometimes randomly selected.
○ K-means clustering is often used in supervised machine learning.

(j) (½ point) Which choice is best for binary classification?
○ K-Means

- ◯ Principal Component Analysis (PCA)
- ◯ Linear regression
- √ Logistic regression

(k) (½ point) Logistic Regression returns probabilities.

√ True ◯ False

(l) (½ point) We should use simple linear regression to predict the winner of a football game.

◯ True √ False

(m) (½ point) In multiple linear regression, is the dependent variable represented by y?

√ Yes ◯ No

(n) (½ point) Using the formula for logistic regression, the line is seen as the best fit *(similar to linear regression)*

√ Yes ◯ No

(o) (½ point) What is the worst choice of split ratio *(Training Set: Test Set)*?

◯ 80 : 20 ◯ 75 : 25 √ 50 : 50

(p) (½ point) Which of these is not one of the steps in the ML process?

- ◯ Data pre-processing
- √ Dashboard creation
- ◯ Modeling
- ◯ Evaluation

(q) (½ point) In *Python*, what is the class used to create a k-NN classifier?

- ◯ `KNearestNeighborsClassifier`
- ◯ `KNearestNeighbors`
- √ `KNeighborsClassifier`
- ◯ `KNN`

(r) (½ point) What is feature scaling?

- √ Scaling column values so that they are comparable
- ◯ Scaling row values so that they are comparable
- ◯ Scaling only 1 column
- ◯ None of the above

(s) (½ point) What is the dependent variable?

- √ Something we are trying to explain or to predict
- ◯ All used variables
- ◯ The variable that is not needed in the datasets
- ◯ None of the above

(t) (½ point) You are part of a data science team that is working for a national fast-food chain. You create a simple report that shows trend: Customers who visit the store more often and buy smaller meals spend more than customers who visit less frequently and buy larger meals. What is the most likely diagram that your team created?

- ○ multi-class classification diagram
- √ linear regression and scatter plots
- ○ K-means cluster diagram

(u) (½ point) You work for an organization that sells a spam filtering service to large companies. Your organization wants to transition its product to use machine learning. It currently a list of 25000 keywords. If a message contains more than few of these keywords, then it is identified as spam. What would be one advantage of transitioning to Machine Learning (ML)?

- ○ The product would look for new patterns in spam messages.
- ○ The product could go through the keyword list much more quickly.
- ○ The product could have a much longer keyword list.
- √ The product could find spam messages using far fewer keywords.

(v) (½ point) Which of the following formulas is not a simple linear regression model?

- ○ Salary = a×Experience
- ○ Salary = a×Experience+b
- √ Salary = a×Experience+b×Age

(w) (½ point) What is the constant in simple linear regression?

- √ It is where the line crosses the vertical axis
- ○ It is where the line crosses each axis
- ○ It is the final model output

(x) (½ point) The Ordinary Least Squares (OLS) method is used in simple linear regression. True/ False?

√ True ○ False

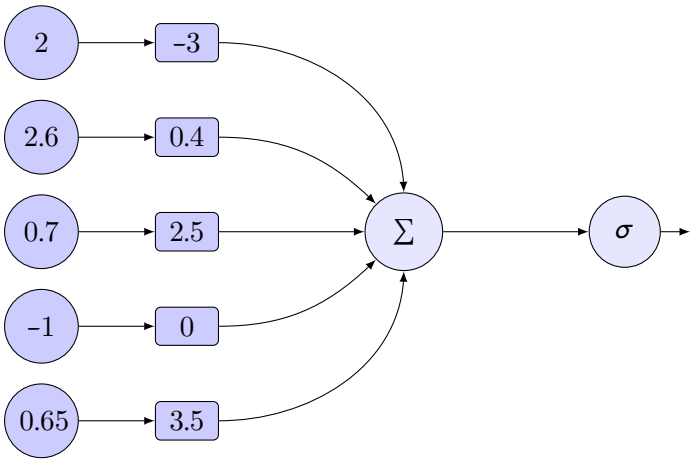
(y) (½ point) Which of the following formulas is not a multiple linear regression model?

- ○ Salary = a×Experience+b×Age+c
- √ Salary = a×Experience+b×Age^2+c
- ○ Salary = a×Experience+b×Age+c×Level+d

Task N°2

⏳ 30mn | (4 points)

(a) (2 points) Compute the output being fired by the following neuron.
σ designates the sigmoid function.



$$y = \sigma(2 \times -3 + 2.6 \times 0.4 + 0.7 \times 2.5 + 0.65 \times 3.5) = 0.28$$

(b) (2 points) Determine how many trainable parameters are in `clf` model, given its following neural network architecture

```
[ ]: clf = Sequential()
     clf.add(Dense(units=8,activation="relu",input_dim=15))
     clf.add(Dense(units=4,activation="relu"))
     clf.add(Dense(units=4,activation="relu"))
     clf.add(Dense(units=1,activation="sigmoid"))
```


$$\text{Params \#} = 15 \times 8 + 8 + 8 \times 4 + 4 + 4 \times 4 + 4 + 4 \times 1 + 1 = 189$$

Task N°3

⏳ 15mn | (2½ points)

We have data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here is four training samples.

x_1	x_2	**Class**
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

We denote by x_1 and x_2 the acid durability in seconds and the strength in kg/m^2, respectively. Now the factory produces a new paper tissue that passes laboratory test with $x_1 = 3$ and $x_2 = 7$.

(a) ($\frac{1}{2}$ point) Without another expensive survey, can you propose a method to determine to which class belongs this new tissue?

k-Nearest Neighbors

(b) (1 point) Calculate the euclidean distance between this query instance and all the training samples.

The coordinate of the query instance is $(3,\ 7)$.

x_1	x_2	$\sqrt{(x_1 - 3)^2 + (x_2 - 7)^2}$
7	7	4
7	4	5
3	4	3
1	4	$\sqrt{13}$

(c) (1 point) Using majority voting mechanism, is the new paper tissue with $x_1 = 3$ and $x_2 = 7$ good or bad?

The three nearest neighbors are $(7,\ 7)$; $(3,\ 4)$ and $(1,\ 4)$. Therefore, using simple majority rule, we have:

$$\mathcal{S}(\text{Good}) = +2 \quad \text{and} \quad \mathcal{S}(\text{Bad}) = +1$$

We conclude then that a new paper tissue that passes laboratory test with $x_1 = 3$ and $x_2 = 7$ is included in **Good** category.