URAL FEDERAL UNIVERSITY

INSTITUTE OF RADIOELECTRONICS AND INFORMATION TECHNOLOGY

DEPARTMENT OF "BIG DATA ANALYSIS AND METHODS OF VIDEOANALYSIS"

END OF SECOND SEMESTER EXAMINATIONS – SPRING 2023

CLASSICAL METHODS OF MACHINE LEARNING

Lecturer: Ebenezer Agbozo (eagbozo@urfu.ru)

Student: Dmitrii Muras, RIM-120981

ANSWER ALL QUESTIONS

For both sections (A & B), answer all your questions in a Python Notebook and push to your GitHub

portfolio for the course.

SECTION A

Select the best answer(s) where necessary

1. Assuming that you have a sufficient data for each of the following problems, which of them

would address Supervised Learning techniques?

a. Determine whether a website displays content for a mature audience.

b. Learn the best way of to split a group of car buyers into categories based on their buying

patterns.

c. Given the medical records from patients suffering a specific illness, learn whether we

split them into different groups for better treatment.

d. Predict next year’s crop yield taking into account data of the past decade.

My comment: Supervised Learning works on existing or given data given to the system so the past year’s crop data will be a material for taking into account and training the system.

2. What is the difference between Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)?

a. RMSE penalizes larger differences between the predictions and the expected results.

b. RMSE is significantly faster to compute than MAE.

c. From both metrics, RMSE is the only one indifferent to the direction of the error

d. From both metrics, MAE is the only one indifferent to the direction of the error

My comment: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are two metrics which are used to evaluate work and performance of the regression model. MAE is the absolute mean of the errors (errors are the differences between the values predicted by our regression model and the actual values of the variable), and RMSE is the root of the mean square of the errors. When we calculate these two metrics, RMSE will penalize us more for the error because of the square in the formula, so it will increase the weight of the error.

3. What is the goal of hyperparameter tuning?

a. To choose a set of optimal samples from the data to train a model.

b. To choose the set of hypotheses that better fit the goal of the model.

c. To choose the optimal parameters for a learning algorithm to train a model.

d. To choose the set of optimal features from the data to train a model.

My comment: Hyperparameters are parameters of an algorithm which determine the performance of the model. The best parameters are the parameters which result to the best accuracy or the least error. So hyperparameter tuning is the process of getting the most optimal parameters for the model.

4. Which of these define Overfitting and Underfitting in simple terms?

a. Overfitting occurs when your model is too complex for your dataset. For example, a very

deep neural network trying to learn a few dozen examples with a couple of features.

b. Underfitting occurs when your model is too simple for your dataset. For example, a

linear regression model trying to learn a large dataset with so many (thousands) of

features.

c. Overfitting occurs when your model is too simple for your dataset. For example, a linear

regression model trying to learn a large dataset with so many (thousands) of features.

d. Underfitting occurs when your model is too complex for your dataset. For example, a

very deep neural network trying to learn a few dozen examples with a couple of

features.

My comment: In Machine learning, overfitting and underfitting occur while training our models. Usually, they show our model’s performance. Overfitting means that our training has focused on the dataset so much that it has missed the point of training entirely. Underfitting means the model has not captured the point or logic of the dataset.

5. Noa is a newly graduated data scientist who works for a school. She is tasked with developing a machine-learning model to predict what college their students will want to apply to at the end of the year. Noa has access to every grade from every previous student, including labels indicating the college they went to. She has several options for building a classification model.

Which of the following should be the best approach to build that model?

a. Noa should use a Decision Tree, a Supervised Learning technique.

b. Noa should use Linear Regression, a Supervised Learning technique.

c. Noa should use Reinforcement Learning.

d. Noa should use Unsupervised Learning.

My comment: Noa has valuable data from the past: information about every grade and every previous student and their college they went to. So, this data will be used to learn and train a model to predict future colleges.

6. Willow overheard her two friends arguing about the best way to handle a few categorical

features on their dataset.

One suggested Label encoding, while the other was pushing for One-Hot encoding. Both are

popular encoding techniques, but Willow didn't know enough to understand the difference. She decided to write a quick summary of both techniques to get everyone on the same page, but the discussion had her confused. She came up with two different explanations for each method, but she wasn't sure which one was correct.

Which of the following statements are correct about these two encoding techniques?

a. One-Hot encoding replaces each label from the categorical feature with a unique integer

based on alphabetical ordering

b. One-Hot encoding creates additional features based on the number of unique values in

the categorical feature.

c. Label encoding replaces each label from the categorical feature with a unique integer

based on alphabetical ordering.

d. Label encoding creates additional features based on the number of unique values in the categorical feature.

My comment: The realization of Label Encoding sorts unique values alphabetically and then assigns them a consecutive number or a unique integer. The implementation of One-Hot encoding is based on the creation of binary features that show belonging to a unique value.

7. Olga is taking an exam for her Master's degree in machine learning.

One of the questions tests her knowledge of Supervised Learning techniques. She needs to

select every problem she can solve using Supervised Learning.

Which of the following problems should Olga select as examples of Supervised Learning?

a. Given a dataset of emails and their classification, build an application to determine whether an email is spam.

b. Given a dataset of audio files and their text transcripts, build an application that turns

any audio snippet into text.

c. Given a dataset of translations between English and Spanish, build an application that

turns any sentence written in English into Spanish.

d. Given a dataset of images of circuit boards and whether they work, build an application

that determines if a picture of a circuit board corresponds to a working board.

My comment: Turning an audio snippet into text and translation from English into Spanish is connected to Natural Language Processing of Machine learning. At the same time, determining of emails and pictures is not connected to it.

8. Your mission is to build a decision tree.

You'll work with a dataset where every feature has a value of 0 or 1. The dataset can have any

number of features.

You want the decision tree to learn a function that outputs how many features in a sample have

a value of 0.

Assuming the dataset has n rows and d features, how many leaf nodes would your decision tree

have?

a. 2ⁿ leaf nodes

b. 2ᵈ leaf nodes

c. 2n leaf nodes

d. 2d leaf nodes

My comment: It comes from the understanding of the decision tree model. It consists of a root node, decision nodes and terminal or leaf nodes. When we want to calculate how many leaf nodes would our decision tree have, we need to raise it to a degree.

9. Sasha knows her k-Nearest Neighbor (KNN) implementation uses a value of K that's too high.

She wants to start experimenting with a lower value.

What should Sasha expect to happen as she decreases K?

a. As Sasha decreases the value of K, she will reduce the algorithm's variance and bias.

b. As Sasha decreases the value of K, she will increase the algorithm's variance and bias.

c. As Sasha decreases the value of K, she will increase the algorithm's variance and reduce

its bias.

d. As Sasha decreases the value of K, she will reduce the algorithm's variance and increase

its bias.

My comment: A low value of K can result to a effect of undertraining the model. At the same time, too high value of K can lead to a risk of overtraining and problems with performance. Also, there are important things such as an algorithm’s variance and bias.

10. A team built a binary classification model. They named the classes A and B.

After finishing training, they evaluated the model on a validation set, and here is the confusion

matrix with the results:

Given the above confusion matrix, what is the f1-score of this binary classification model at

predicting class A?

a. The f1-score of the model at predicting class A is 52%.

b. The f1-score of the model at predicting class A is 80%.

c. The f1-score of the model at predicting class A is 84%.

d. The f1-score of the model at predicting class A is 88%.

My comment: f1-score = 2 \* (Precision \* Recall) / (Precision + Recall)

To calculate Precision -> TP / (TP + FP)

To calculate Recall -> TP / (TP + FN)

TP and NP – true and false positives respectively

TN and FN – true and false negatives respectively

Precision = 52 / (52 + 13) = 0,8

Recall = 52 / (52 + 7) = 0,88

F1-score = 2 \* (0,8 \* 0,88) / (0,8 + 0,88) = 1,408 / 1,68 = 0,838095 or 0,84.

SECTION B

(SELECT ONLY ONE MACHINE LEARNING TASK FROM THE VARIANTS BELOW)

You have been employed as a Senior Data Scientist for a consulting firm, and your job is to extract

knowledge from data, as well as build ML models for use by clients.

Train an ML model, and build a Streamlit App to deploy your model (MLOps)

1. (Open Task – Regression or Classification) – Music Lyrics

Build an NLP-based model with themusic lyrics and popularity database.

The columns/features -Name of the song, Artist Name, Album Name, Popularity on Spotify,

Transcribed Lyrics.

You have to use the Transcribed Lyrics in building the ML Model

Dataset:https://disk.yandex.ru/d/gxnTjHN7OZpHGg

2. (Financial Credit Model Worthiness Model)

As a prominent bank’s Data Scientist, build a model that predicts whether a customer is credit-

worthy. The columns include:- checking\_status, duration, credit\_history, purpose,

credit\_amount, savings\_status, employment, installment\_commitment, personal\_status,

other\_parties, residence\_since, property\_magnitude, age, other\_payment\_plans, housing,

existing\_credits, job, num\_dependents, own\_telephone, foreign\_worker, class

Dataset:https://disk.yandex.ru/d/zyXWN2xc5WZKLg

3. (Steel Factory Energy Prediction Model)

The information gathered is from the DAEWOO Steel Co. Ltd in Gwangyang, South Korea. It

produces several types of coils, steel plates, and iron plates. The information on electricity

consumption is held in a cloud-based system. The information on energy consumption of the

industry is stored on the website of the Korea Electric Power Corporation (pccs.kepco.go.kr),

and the perspectives on daily, monthly, and annual data are calculated and shown. The

attributes below define the dataset:

 Data Variables Type Measurement

 Industry Energy Consumption Continuous kWh

 Lagging Current reactive power Continuous kVarh

 Leading Current reactive power Continuous kVarh

 tCO2(CO2) Continuous ppm

 Lagging Current power factor Continuous %

 Leading Current Power factor Continuous %

 Number of Seconds from midnight Continuous S

 Week status Categorical (Weekend (0) or a Weekday(1))

 Day of week Categorical Sunday, Monday - Saturday

 Load Type Categorical Light Load, Medium Load, Maximum Load

Dataset:https://disk.yandex.ru/d/7fLSLxH2hi0jZA

Disclaimer:

Perform data pre-processing, exploratory data analysis (EDA), model building, and model

evaluation.

Creative Data Science Thinking will be awarded!

A High Model accuracy will be awarded good scores, but the most important part of this

exam task is your ability to be creative with your model.

Good Luck.