ML Process overview, Dataset terminology, Data preprocessing, and Feature extraction/selection/scaling

Total points 13/14

Email *

shinoda.c.i@gmail.com

✓ Select the correct statement about the ML process steps: * 1/1

- ○ Data Collection: Collecting the relevant data to be used in the model, which may include structured, semi-structured or unstructured data.
- ○ Model Training: Building a machine learning model on the prepared data to predict an outcome.
- ○ Model Evaluation: Assessing the performance of the model, typically by measuring its accuracy, precision, recall, and F1 score, as well as other relevant metrics.
- ○ Deployment: Implementing the model in a production environment, which may involve integrating it with other software systems and ensuring its security and scalability.
- ◉ All of them are correct. ✓

✓ Which of the following best describes the difference between the terms "feature", "label", and "sample" in the context of machine learning? *1/1

- (●) Features are the inputs to the model, labels are the outputs, and samples are the training data. ✓
- ( ) Feature and labels mean the same thing, while samples are the training data.
- ( ) Features are the outputs of the model, samples are the inputs, and labels are the predictions.

Feedback

Features are the input variables or characteristics that are used to make predictions, labels are the output variables or targets that the model tries to predict, and samples are the individual instances of data used for training and testing the model.

✓ Consider the following dataset, which describes characteristics of cars. Assuming you are developing a regression model to predict the price of cars, select the true sentences *1/1

Make	Model	Year	Price
Chevrolet	Astra	2010	25,000
Ford	Ka	2012	21,000
...	...	...	...

- ( ) a) The price column should be used as a feature
- ( ) b) Make, Model, and Year should be used as labels when training the model
- (●) c) It is a good practice to split the dataset so you can evaluate your model with unseen data and detect problems such as overfitting ✓

✓ Which of the following scenarios represents an appropriate usage of data cleansing in a dataset that will be used to train a model using ML algorithms? * 1/1

- a) Filling missing values with a fixed value for all the observations, without taking into account the specific characteristics of the missing data
- b) Removing all the rows containing missing values, without analyzing the patterns and frequency of the missing data
- (●) c) Imputing missing values using the mean or median value of the column, after analyzing the patterns and frequency of the missing data ✓

Feedback

c) Correct. This approach involves analyzing the missing data to determine the most appropriate imputation method

✓ Based on the following table, identify the example of “dirty data”: * 1/1

Student ID	Grade	Major	Age	Graduation Year
001	3.4	CS	12	2024
002	2.9	EE	22	2023
003	3.8	CS	20	2025
004	3.1	CS	23	2022
005	3.7	CS	21	2024

- (●) Student 001 - Age is below the minimum expected age for a person that is supposed to graduate from CS in 2024 ✓
- ( ) Student 002 - Age is below the minimum age requirement for the program
- ( ) Student 003 - Major is inconsistent with the Graduation Year
- ( ) Student 005 - Graduation Year is missing

Feedback

dirty data because it could be a value misplacing (12 instead of 21)

✓ Data preprocessing is an important step in machine learning that involves cleaning and transforming data before training models * 1/1

- (●) True ✓
- ( ) False

✓ Regarding data preprocessing, text and categorical attributes may need to be encoded as numbers * 1/1

- (●) True ✓
- ( ) False

✓ When is feature selection useful in machine learning? * 1/1

- ○ When the data has a small number of features
- ○ When the data is composed of textual and numerical data
- ◉ When the data has a large number of irrelevant features ✓
- ○ When the data is already in a format that is suitable for machine learning

Feedback

Removing irrelevant features helps reducing the amount of noise in the data

✓ What is the difference between filter and wrapper methods for feature selection in machine learning? *1/1

- ◉ Filter methods involve evaluating each feature based on a statistical measure, while wrapper methods uses not single features, but the combination of them (subsets of features) ✓
- ○ Filter methods involve selecting features based on their performance on a specific machine learning algorithm, while wrapper methods evaluate each feature based on a statistical measure.
- ○ Filter methods and wrapper methods are the same thing.
- ○ Filter methods and wrapper methods are both embedded methods for feature selection

✗ Scenario: You are working on a machine learning project and have a dataset with thousands of features. You want to reduce the number of features to improve the performance of your model. Which type of feature selection method would be most appropriate? *0/1

- ○ Filter method
- ◉ Wrapper method ✗
- ○ Embedded method

Correct answer

- ◉ Filter method

✓ Fill in the blanks in the following text: *1/1

"Feature extraction is the process of selecting or transforming _____ data into a set of relevant _____ that can be used as input to a machine learning algorithm for training and prediction. This is typically done by applying various mathematical and statistical techniques to the data, such as filtering, dimensionality reduction, or normalization. The resulting set of features should be informative, _____ , and independent of each other, so that they can effectively capture the underlying patterns in the data and generalize well to new examples.

- ◉ high-dimensional; features; discriminative ✓
- ○ low-dimensional; labels; dependent
- ○ raw; features; correlated

✓ Dimensionality reduction techniques can be used to reduce the number of features in a dataset, which can improve the performance of machine learning models. *1/1

- (●) True ✓
- ( ) False

✓ Given the following situation, choose the correct alternative about the min-max scaling: *1/1

Suppose you have a dataset with a numeric feature X that has a minimum value of 10 and a maximum value of 50. Apply min-max scaling to X and calculate the scaled value for X when its original value is 30.

0.5 ✓

Feedback

The range of X is (50 - 10) = 40, and the original value of X is 30. Therefore, the scaled value of X is:

scaled value = (30 - 10) / 40 = 0.5

✓ Select the correct statements about min-max scaling: * 1/1

- [ ] It transforms the values of a variable to have a mean of 0 and a standard deviation of 1
- [x] It can be sensitive to outliers and can distort the distribution of the data ✓
- [ ] It involves dividing each value of a variable by its mean

Feedback

If the range of the variable is very wide or if there are extreme values in the data, the scaling may be biased towards those values, and the resulting distribution may be distorted.