K. D. P. R. Jayawardena

2021s18852

Assignment 1

1.

a.) Training set: Training dataset is the sample of data used to fit the model. It is the actual dataset that we use to train the model. Training is the phase where a machine learning model learns from the labeled training data. The model adjusts its internal parameters to minimize the difference between its predictions and the actual labels in the training set.

b.) Validation set: A validation set is a set of data used to train artificial intelligence with the goal of finding and optimizing the best model to solve a given problem. Validation sets are also known as dev sets.

c.) Test set: Test dataset is the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset. It is used once a model is completely trained.

2.

a.) Underfitting: When a model has not learned the patterns in the training data well and is unable to generalize well on the new data, it is known as underfitting. An underfit model has poor performance on the training data and will result in unreliable predictions. Underfitting occurs due to high bias and low variance. Underfitting can occur due to higher bias of the model, model being too simple and size of the training dataset used is not enough. Underfitting can be tackled by increasing model complexity, reduce noise in the data and by increasing the duration of training the data.

b.) Overfitting: When a model performs very well for training data but has poor performance with test data, it is known as overfitting. In this case, the machine learning model learns the details and noise in the training data such that it negatively affects the performance of the model on test data. Overfitting can happen due to higher variance, size of training dataset used isn't enough and model being too complex.

3.

a.) Data splitting means partitioning of a dataset into different subsets such as training, validation and test sets. It ensures the creation of data models and processes that use data models are accurate.

b.) The train_test_split() method is used to split our data into train and test sets. First, we need to divide our data into features (X) and labels (y). The dataframe gets divided into X_train, X_test, y_train, and y_test. X_train and y_train sets are used for training and fitting the model. The X_test and y_test sets are used for testing the model if it's predicting the right outputs/labels.

we can explicitly test the size of the train and test sets. It is suggested to keep our train sets larger than the test sets.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

- X: The features (independent variables)
- y: The labels (dependent variable)
- test_size: The proportion of the dataset to be used as the test set
- random_state: A seed for the random number generator to ensure reproducibility
- X_train, X_test: The training and testing features, respectively.
- y_train, y_test: The corresponding labels for the training and testing sets.

c.) test_size is the number that defines the size of the test set. It's very similar to train_size. You should provide either train_size or test_size. test_size = 0.2 means that 20% of the data will be used for testing, and the remaining 80% will be used for training. If neither is given, then the default share of the dataset that will be used for testing is 0.25, or 25 percent.

d.)

```
from sklearn.model_selection import train_test_split
import numpy as np

# Example data (features and labels)
X = np.array([[1, 2], [3, 4], [5, 6], [7, 8], [9, 10]])
y = np.array([1, 0, 1, 0, 1])

# Split the data into training and testing sets (80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Print the results
print("X_train:")
print(X_train)

print("X_test:")
print(X_test)

print("y_train:")
print(y_train)

print("y_test:")
print(y_test)
```

Output is as following:

```
X_train:
[[ 9 10]
 [ 5  6]
 [ 1  2]
 [ 7  8]]
X_test:
[[3 4]]
y_train:
[1 1 1 0]
y_test:
[0]
```

4.

a.)

I. CRIM - Per capita crime rate by town
II. ZN - The proportion of residential land zoned for lots over 25,000 sq. ft.
III. INDUS - The proportion of non-retail business acres per town.
IV. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
V. NOX - Nitric oxides concentration (parts per 10 million)
VI. RM - Average number of rooms per dwelling
VII. AGE - Proportion of owner-occupied units built before 1940
VIII. DIS - Weighted distances to five Boston employment centers
IX. RAD - Index of accessibility to radial highways
X. TAX - Full-value property-tax rate per $10,000
XI. PTRATIO - Pupil-teacher ratio by town
XII. B - $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
XIII. LSTAT - Percentage lower status of the population
XIV. MEDV - Median value of owner-occupied homes in $1000's

b.)

c.)