7/25/2024

SOWBARANIGA KULANDHASAMY 22CSR202

systima nx

JOINT TECH INTERNSHIP COMMUNITY PROGRAM

MACHINE LEARNING

Assignment 1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Patient ID** | **Age** | **Blood Pressure**  **(mm/Hg)** | **Cholesterol**  **(mg/dL)** | **Symptom 1** | **Symptom 2** | **Symptom 3** | **Diagnosis** |
| 001 | 67 | 140/90 | 220 | Headache | Nausea | Fatigue | Diabetes |
| 002 | 56 | 120/80 | 240 | Cough | Fever | Weight loss | Flu |
| 003 | 34 | 115/75 | 190 | Fatigue | Nausea | Muscle pain | Normal |
| 004 | 89 | 150/95 | 210 | Shortness of breath | Chest Pain | Head ache | Heart disease |
| 005 | 23 | 110/70 | 220 | Dizziness | Nausea | Fatigue | Hypertension |

Sample test data of Medical Diagnosis:

In the above Example Training Data Set,

**1.Feature**

* A feature is data that is used as the input for Machine Learning models to make predictions.
* Example: Patient age, BP, Cholesterol, Symptoms are the features of this dataset.

**2. Label**

* Label is a value or thing we are trying to predict.
* Example: The diagnosis of a specific disease.

**3.Prediction**

* Prediction is the process in which we feed unlabeled data to the trained model and obtain the values of label.
* Example: Predicting the disease of the patient with symptoms.

**4.Outlier**

* Outlier are data points that significantly differ from the majority of the data in the dataset.
* Example: The high blood sugar level of the patient compared to the range of all other patients.

**5.Test Data**

* Test dataset is a dataset for which class label is unknown.
* It is tested using model.
* A test dataset used for assessment of the finally chosen model.
* Example: The set of patient data used to test the diagnosis.

**6.Training Data**

* Training Dataset has attributes and class labels and used for training Machine Learning Algorithm to prepare model.
* Machine can learn when they observed relevant data.
* From that they find relationships, detect patterns, understand complex problems and make decision.
* Example: The collection of patient records used to train the model to diagnose the disease.

**7.Model**

* Model defines the relationship between features and label.
* The decision tree model used to classify the disease of the patient by all the feature.

**8.Validation Data**

* Validation datasets use a sample of data that is withheld from training. That data is then used to evaluate any apparent errors. Machine Learning Engineers can then tune the model’s hyperparameters – which are adjustable parameters used to control the behavior of the model.

**9.Hyperparameter**

* A hyperparameter is a Machine Learning parameter whose value is chosen before a learning algorithm is trained.
* Hyperparameter should not be confused with parameters. In Machine Learning, the label parameter is used to identify variables whose values are learned during training.

**10. Epoch**

* Epoch are defined as the total number of iterations for training the Machine Learning model with all the training data in one cycle.
* One complete pass through the patient data during the training of the neural network for diagnosing the disease.

**11.Loss function**

* Quantifies the error between output of the algorithm and given target value.
* It is also known as Cost function or Error function.
* Cross-entropy loss used to measure the difference between the predicted disease and actual disease.

**12.Learning rate**

* How big/small the steps are gradient descent takes into the direction of the local minimum are determined by the learning rate, which figures out how fast or slow we will move towards the optimal weights.

**13.Overfitting**

* Overfitting means the more we train our model, the more chances of occurring the overfitted model.
* Overfitting is an undesirable Machine Learning behaviour that occurs when the Machine Learning model gives accurate predictions for training data but not for new data.

**14.Underfitting**

* In case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable properties.

**15.Regularization**

* Regularization avoids overfitting by adding a penalty to the model’s loss function.
* Regularization = Loss Function + Penalty

**16.Cross-validation**

* Cross Validation is a technique used in Machine Learning to evaluate the performance of a model on unseen data.
* Cross Validation is an important step in the Machine Learning process and helps to ensure that the model selected for deployment is robust and generalizes well to new data.
* Splitting patient data into multiple subsets to test/train the model on different combination to evaluate performance.

**17.Feature Engineering**

* Feature Engineering is the process of selecting, extracting and transforming the most relevant features from available data to build more accurate and efficient Machine Learning models.

**18.Dimensionality Reduction**

* Dimensionality Reduction is the transformation of data from a high dimensional space into a low dimensional space so, that the low dimensional represents some meaningful properties of the original data, ideally closer to original data.
* Reducing the number of features in patient data using Principal Component Analysis (PCA).

**19.Bias**

* Bias is simply defined as the inability of the model.
* Because of that there is some difference or error occurring between the model’s predicted value and the actual value.

**20.Variance**

* Variance is the measure of spread in data from its mean position.
* In Machine Learning, variance is the amount by which the performance of a predictive model changes when it is trained on different subsets of the training data.