

Machine learning

Assignment 1



July 24, 2024

systima nx

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 for Medical Diagnosis :

The Terminologies used:

1.Feature

Patient’s age, BP, Cholesterol, Symptoms are the features of this dataset.

2. Label

The diagnosis of a specific disease

3.Prediction

Predicting the disease of the patient with symptoms.

4.Outlier

The high blood sugar level of the patient compared to the range of all other patients.

5.Test data

The set of patient data used to test the diagnosis model.

6.Training model

The collection of patient records used to train the model to diagnose the disease.

7.Model

The decision tree model used to classify the disease of the patient by all the feature.

8.Validation data

The separate dataset from the training dataset used to tune the model’s parameters and prevent overfitting.

9.Hyperparameter

The maximum depth of the decision tree used for diagnosing the disease.

10. Epoch

One complete pass through the patient data during the training of the neural network for diagnosing the disease.

11.Loss function

Cross-entropy loss used to measure the difference between the predicted disease and actual disease.

12.Learning rate

The step size used for updating the weights of a neural network during training.

13.Overfitting

The model that performs well on training data but poor on test data because it learned even noises and specifics.

14.Underfitting

The model is too simple and fails to diagnose the complex problems and leads to poor performance.

15.Regularization

Adding a penalty to the model for large coefficients to prevent disease prediction models.

16.Cross-validation

Splitting patient data into multiple subsets to test/train the model on different combination to evaluate performance.

17.Feature Engineering

Creating a new feature from existing features, created new feature called risk factor by combining cholesterol and blood pressure features.

18.Dimensionality Reduction

Reducing the number of features in patient data using Principal Component Analysis (PCA).

19.Bias

Systematic errors in models due to inadequate model.

20.Variance

High sensitivity to fluctuations in training data, causing the model to perform inconsistently.

BY :

SWATHI K

22CSR216