Name :- Afnan Attar   PRN :- F19112003 Class :- BE COMPII
Subject :- ML

Home Assignment 1

Q1) How does machine learning work?

Ans 1. Machine learning is a form of artificial intelligence (AI) that teaches computers to think in a similar way to how humans do: Learning and improving upon past experiences.

2. It works by exploring data and identifying patterns, and involves minimal human intervention.

3. Almost any tasks that can be completed with a data-defined pattern or set of rules can be automated with machine learning.

4. This allows companies to transform process that were previously only possible for humans to perform.

5. Machine learning's foundation is purely statistical mathematics, it is the basis for its working.

Q2) Explain reinforcement learning with suitable example.

Ans 1. Reinforcement learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing actions and seeing the results of actions.

2. Here agent learns automatically using feedbacks unlike labelled data.

3. It solves a specific type of problem where decision making is sequential, and the goal is long-term such as game-playing, robotics, etc.

4. Example:

i) Suppose an agent is presented with a maze environment

and his goal is to find the diamond;

ii) The agent continues doing these three things (take action, change state/remain in the same state and get feedback) and by doing these actions he learns and explores the environment.

iii) Hence the agent learns what leads to a positive feedback rewards and what leads to negative feedback penalty.

Q3) Explain geometric and probabilistic models.

Ans I] Geometric Models :-

1. It is a technique of combining machine learning and computer vison to solve visual tasks.

2. These models define similarity by considering the geometry of instance space.

3. Here features can be described in two dimensions (x and y-axis) or three dimensions (x, y, z-axis).

4. Geometric model that is constructed directly in instance space is called as linear model and a model that uses distance as a metric to represent the similarities between instances is known as distance based models.

II] Probabilistic Models :-

1. A probability model is based on theory of probability or the fact that randomness plays role in future events.

2. Hence according to the probability of an event occuring we make predictions and take certain decisions.

3. Naïve Bayes theorem is the foundation of probabilistic models :-

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

Likelihood ↑ (P(X|Y))

Prior probability ↗ (P(Y))

Posterior Probability ↑ (P(Y|X))

Predictor Prior probability ↑ (P(X))

Q4) Explain the different feature engineering techniques.

Ans A] Imputation :-
1. This deals with handling missing values in data.
2. Imputation can on categorical or Numerical variables.

B) Discretization :-
1. Discretization involves taking set of values of data and grouping sets of them together in some logical fashion.
2. We can group by intervals, frequencies or decision tree.

c] Categorical Encoding :-
1. It is used to encode categorical features into numerical values which are usually simpler.
2. Categorical variables are converted into 1's and 0's.

d] Feature Splitting:
Splitting features into parts can sometimes improve the value of the features towards target to be learned.

e] Handling Outliers:
Outliers are unusually highly or low values in the dataset which are unlikely to occur in normal scenarios. We can remove, replace etc. on them.

f] Variable transformation :-
Variable transformation technique could help with normalizing skewed data.

g] Scaling:
Feature scaling is done owing to the sensitivity of some machine learning algorithms to the scale of input values.

h] Creating features:
Feature creation involves deriving new features from existing ones.

Q5) Explain different dimensionality reduction techniques.

Ans   Various Methods used for dimensionality reduction include:

I] Principal Component Analysis (PCA):

1. PCA works on a condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.

2. It helps in data compression and hence reduced storage space which in turn reduces computation time.

II] Linear Discrimant Analysis (LDA):

1. It is commonly used for supervised classification problem.

2. It is used for modelling differences in groups i.e seperating two or more classes, it can project the features in higher dimension space into a lower dimension space.

III] Generalized Discriminant Analysis (GDA):

1. It deals with nonlinear discriminant analysis using kernel function operator.

2. The underlying theory is close to the support vector machines in so far as the GDA method provides a mapping of the input vectors into a high dimensional feature space. Similar to LDA, GDA is used to find a projection for the features into lower-dimensional space.

Q6) Explain various feature selection techniques.

Ans   There are mainly 3 types of feature selection techniques:-

i) Wrapper Methods:

1. In wrapper methodology, selection of features is done by considering it as search problem, in which different combinations are made, evaluated and compared with other combinations.

2. On the basis of the output of the model, features are added or subtracted and with new feature set model is

trained again.

ii) **Filter Methods:-**

1. In this features are selected on the basis of statistical measures.

2. The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking.

3. Techniques like information gain, chi-square tests and so on are used in filter method.

iii) **Embedded Methods:-**

1. Embedded methods combine the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost.

2. Some techniques of embedded methods are :-

i) **Regularization :-**
Regularization adds a penalty term to different parameters of machine learning model for avoiding overfitting in the model.

ii) **Random Forest Importance :-**
Different tree-based methods of feature selection help us with feature importance to provide a way of selecting features.