



COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

END Semester Examination

Programme: Final Year B.Tech

Semester: VIII

Course Name: AI & ML for Mechanical Engineers

Course Code: ME(DE)-22016

Branch: Mechanical Engineering

Academic Year: 2022-23

Duration: 3 hours

Max Marks: 60

- | | Marks |
|---|-------|
| Q.1 A decision tree classifier is to be trained to predict the likelihood of a vehicle rolling over during a sharp turn based on input features: vehicle speed during the turn (mph), angle of the turn (degrees), type of vehicle (e.g., sedan, SUV, pickup truck), weight (lbs), road surface conditions (e.g., dry, wet, icy), tire tread, driver behavior (e.g., aggressive, cautious). | 6 |
| <ul style="list-style-type: none">• Observe the dataset; Identify discrete & continuous attributes along with their features.• Calculate information gain of attribute ‘Vehicle Type’ & ‘Road Surface’.• Compare them and comment on which one of these two is suitable for the best split. | |

Speed (mph)	Turn Angle	Vehicle Type	Weight (lbs)	Road Surface	Tire Tread	Driver Behavior	Rollover
35	50°	SUV	4000	Wet	Good	Aggressive	Yes
25	30°	Sedan	3200	Dry	Good	Cautious	No
40	70°	Pickup	5000	Icy	Poor	Aggressive	Yes
30	45°	SUV	4500	Wet	Good	Cautious	No
50	80°	Pickup	5500	Dry	Good	Aggressive	Yes
20	20°	Sedan	2800	Dry	Good	Cautious	No
30	60°	SUV	4200	Wet	Good	Aggressive	Yes
40	75°	Pickup	5100	Icy	Poor	Cautious	Yes
35	55°	SUV	4100	Wet	Good	Cautious	No
45	90°	Pickup	5800	Dry	Good	Aggressive	Yes

Answer

Discrete attributes:

- Vehicle Type: SUV, Sedan, Pickup
- Road Surface: Wet, Dry, Icy
- Tire Tread: Good, Poor
- Driver Behavior: Aggressive, Cautious

Continuous attributes:

- Speed (mph), Turn Angle, Weight (lbs)



COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

Entropy before split = $-(6/10) \log_2(6/10) + -(4/10) \log_2(4/10) = 0.971$

Vehicle Type

- SUV: 4 instances, 2 rollovers, 2 non-rollovers
- Sedan: 2 instances, 0 rollovers, 2 non-rollovers
- Pickup: 4 instances, 4 rollovers, 0 non-rollover

Entropy after split:

- SUV: $-(2/4) \log_2(2/4) + -(2/4) \log_2(2/4) = 1.0$
- Sedan = 0
- Pickup = 0

Information gain = $0.971 - (4/10 * 1) = 0.57$

Road Surface

- Wet: 4 instances, 2 rollovers, 2 non-rollovers
- Dry: 4 instances, 2 rollovers, 2 non-rollovers
- Icy: 2 instances, 2 rollovers, 0 non-rollovers

Entropy after split:

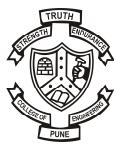
- SUV: $-(2/4) \log_2(2/4) + -(2/4) \log_2(2/4) = 1.0$
- Sedan: $-(2/4) \log_2(2/4) + -(2/4) \log_2(2/4) = 1.0$
- Pickup = 0

Information gain = $0.971 - (4/10 * 1) - (4/10 * 1) = 0.17$

In the given scenario, the information gain for Road Surface is 0.17, and the information gain for Vehicle Type is 0.57. A higher information gain means that the attribute provides more useful information for splitting the data. Therefore, **Vehicle Type is better suited for the best split in this scenario** because it has a higher information gain than Road Surface. It means that using Vehicle Type as the splitting attribute will result in a more effective classification.

Q.2 Use PCA to find the principal components of a heat transfer problem on a heat exchanger with the training data consisting of 4 attributes i.e. inlet temperature of fluid A, inlet temperature of fluid B, flow rate of fluid A and overall heat transfer coefficient. 6

Sample	Inlet temperature of fluid A (°C)	Inlet temperature of fluid B (°C)	Flow rate of fluid A	Overall HT coefficient
0	300	200	20	0.1
1	350	150	22	0.2
2	325	180	21	0.3
3	310	190	25	0.4
4	275	225	19	0.5
5	290	210	23	0.6



COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

Answer

Explained variance: [0.67791955 0.23154196 0.08608349 0.004455]

Principal components:

[[-0.46064822 0.50682283 -0.53559456 -0.48863388]

[0.87958437 -0.19865672 -0.35507762 -0.2593064]

[0.08865614 0.83483462 0.52916349 -0.12230958]

[0.0424782 0.07994237 0.56438126 -0.82047375]]

PC1 = - 0.461*(Inlet temperature of fluid A) + 0.507*(Inlet temperature of fluid B) - 0.536*(Flow rate of fluid A) - 0.489*(Overall HT coefficient)

PC2 = 0.879*(Inlet temperature of fluid A) - 0.198*(Inlet temperature of fluid B) - 0.355*(Flow rate of fluid A) - 0.259*(Overall HT coefficient)

PC3 = 0.088*(Inlet temperature of fluid A) + 0.834*(Inlet temperature of fluid B) - 0.529*(Flow rate of fluid A) - 0.122*(Overall HT coefficient)

PC4 = 0.042*(Inlet temperature of fluid A) + 0.079*(Inlet temperature of fluid B) - 0.564*(Flow rate of fluid A) - 0.820*(Overall HT coefficient)

- Q.3 In dataset given below, there are 3 attributes – ‘Location’, ‘Type of Solar Panel’, and ‘Efficiency’ and 6 samples of different solar energy systems. Using Bayes classifier, calculate the probability of a solar panel being a Monocrystalline panel given that it has an efficiency of 0.18 and is installed at location California. 6

Sample	Type of Solar Panel	Efficiency	Location
1	Monocrystalline	0.18	California
2	Polycrystalline	0.16	Texas
3	Thin Film	0.14	Florida
4	Monocrystalline	0.18	New York
5	Polycrystalline	0.16	Ontario
6	Thin Film	0.14	Quebec

Answer

To calculate the probability of a solar panel being Monocrystalline given that it has an efficiency of 0.18 and is installed in California, we can use Bayes' theorem:

$$P(M|E,C) = P(E|M,C) * P(M|C) * P(E|C) / 1$$

where, M = Monocrystalline E = Efficiency of 0.18 C = California

We need to calculate the following probabilities:

$P(E|M,C)$ = Probability of an efficiency of 0.18 given that the panel is Monocrystalline and installed in California
 $P(M|C)$ = Probability of a Monocrystalline panel given that it is installed in California
 $P(E|C)$ = Probability of an efficiency of 0.18 given that it is installed in California.



COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

Let's calculate these probabilities using the given data:

Total number of solar panels installed in California = 2
Number of Monocrystalline panels installed in California = 1
Number of panels with an efficiency of 0.18 installed in California = 1
Total number of Monocrystalline panels = 2
Total number of panels with an efficiency of 0.18 = 2
Total number of panels in the dataset = 6

Using the above information, we can calculate the probabilities:

$$P(E|M,C) = 1/1 = 1 \quad P(M|C) = 1/2 \quad P(E|C) = 2/6$$

Now, we can calculate the probability of a solar panel being Monocrystalline given that it has an efficiency of 0.18 and is installed in California:

$$P(M|E,C) = P(E|M,C) * P(M|C) * P(E|C) / 1$$

$$P(M|E,C) = 1 * (1/2) * (2/6) / 1$$

$$P(M|E,C) = 0.1677$$

Therefore, the probability of a solar panel being Monocrystalline given that it has an efficiency of 0.18 and is installed in California is 0.1677 or 16.77%.

Q.4 A confusion matrix for a cutting tool faults classification is shown below. Observe this matrix and calculate. 6

- Correctly classified samples (considering all classes and per class)
- Incorrectly classified samples (considering all classes and per class)
- Overall accuracy of classification
- True positive, True negative, False positive, False negative, Precision, F1 Score

Actual					
Normal	Inner Race Fault	Outer Race Fault	Ball Fault		Predicted
140	00	00	00	Normal	
00	10	05	00	Inner Race Fault	
00	00	20	03	Outer Race Fault	
00	00	02	08	Ball Fault	

Answer

Correctly classified samples for

Normal condition = 140

Inner Race Fault = 10

Outer Race Fault = 20

Ball Fault = 08

Total correctly classified samples = $140 + 10 + 20 + 08 = 178$

Incorrectly classified samples for

Normal condition = 00



COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

Inner Race Fault = 05

Outer Race Fault = 03

Ball Fault = 02

Total incorrectly classified samples = $00 + 05 + 03 + 02 = 10$

Overall accuracy = sum of diagonal elements / total samples = $178 / 188 = 94.68\%$

True positive = 140

True negative = $10 + 20 + 08 = 38$

False positive = 00

False negative = $05 + 03 + 02 = 10$

Precision = $140 / (140 + 0) = 1$

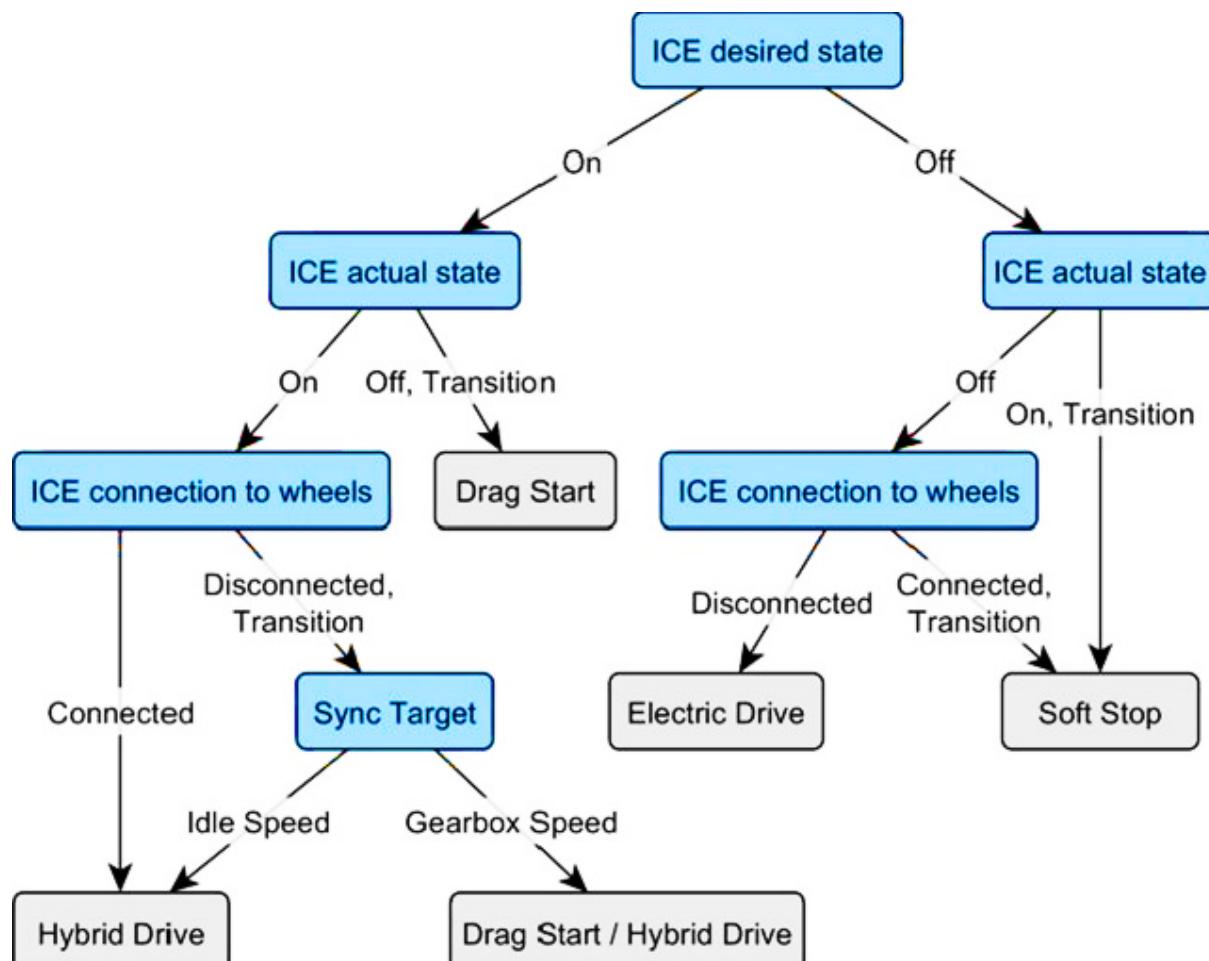
F1 Score = $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$

F1 Score = $2 * (1.0 * 0.9333) / (1.0 + 0.9333) = 0.9655$

Q.5 Answer the following questions.

30

- A. How does following decision tree algorithm work for controlling combustion engine? Why is partial overlapping of drag start and hybrid drive observed in bottom center decision node? Also, identify various nodes, attributes and features.





COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

Answer

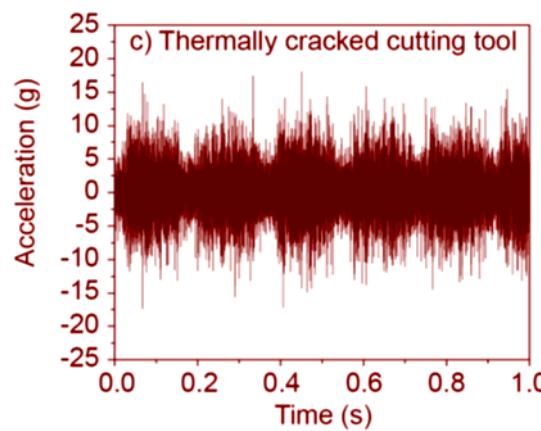
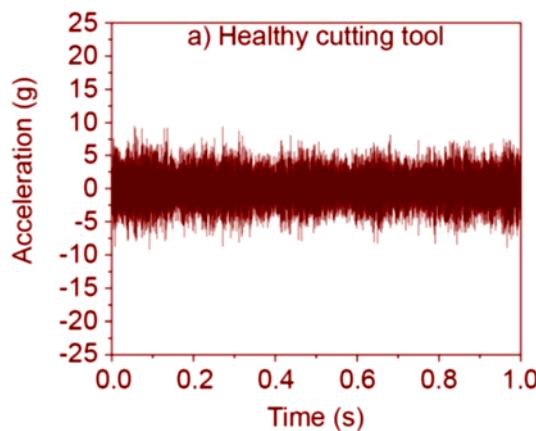
- The control logic of an axle-split hybrid vehicle characterized by a front wheel drive via an internal combustion engine and a rear wheel drive via an electric motor.
- The function Engine control is hierarchically superimposed on the function Combustion start.
- The Engine Control function determines the operating state of the Internal Combustion Engine (ICE) in the hybrid vehicle.
- Depending on the requirements of the vehicle strategy and the current context (engine stopped, transmission on/off, speed synchronized), the combustion engine is operated and switched on or off.
- The four input dimensions span the space of the operating situations and are supplemented by the output dimension Operation with regard to possible output signals of the control.
- Four modes were defined in the second step, which in this case are identical to the four alternatives of the output dimension Operation: Combustion start (Drag Start), Combustion stop (Soft Stop), Combustion off (Electric Drive) and Combustion running (Hybrid Drive).
- When defining the four modes, an overlap between the modes Drag Start and Hybrid Drive was found which represents a hysteresis in the system - If the system was already in Hybrid Drive before, the system remains in this mode, and otherwise it switches to the mode Drag Start.

Root node: ICE desired state

Branch node: ICE actual state, ICE connected wheels, Sync target

Leaf node: Drag start, soft stop, hybrid drive, electric drive, drag start / hybrid drive

- B. Following graphs represent change in vibration signal (in terms of acceleration) with respect to time that depicts 2 conditions i.e. healthy and faulty cutting tool. 5





COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

In order to develop machine learning based classification model, which statistical features will you extract from these vibration plots so as to depict difference between two. Draw a flowchart showing step by step procedure involved in developing classification model.

Answer

Some of the commonly used features are Mean, Standard deviation, Root mean square (RMS), Kurtosis, Skewness, Fast Fourier Transform (FFT) coefficients.

Once the relevant features have been extracted, we can follow the below steps to develop the machine learning-based classification model:

- Data collection: Collect the vibration data from healthy and faulty cutting tools.
- Data pre-processing: Clean and preprocess the data to remove any noise or outliers.
- Feature extraction: Extract relevant statistical features from the vibration plots.
- Data splitting: Split the data into training and testing sets.
- Model selection: Choose the appropriate machine learning algorithm such as Random Forest, Support Vector Machine (SVM), or Neural Networks based on the size and complexity of the data.
- Model training: Train the selected model on the training data.
- Model evaluation: Evaluate the performance of the trained model on the testing data using metrics such as accuracy, precision, recall, F1 score, and confusion matrix.
- Model tuning: Optimize the hyperparameters of the model to improve its performance.
- Model deployment: Deploy the final model on the new data for prediction.

C. Compare bagging and boosting ensemble techniques. Does the Random forest tree algorithm fall into any of these categories? Justify. 5

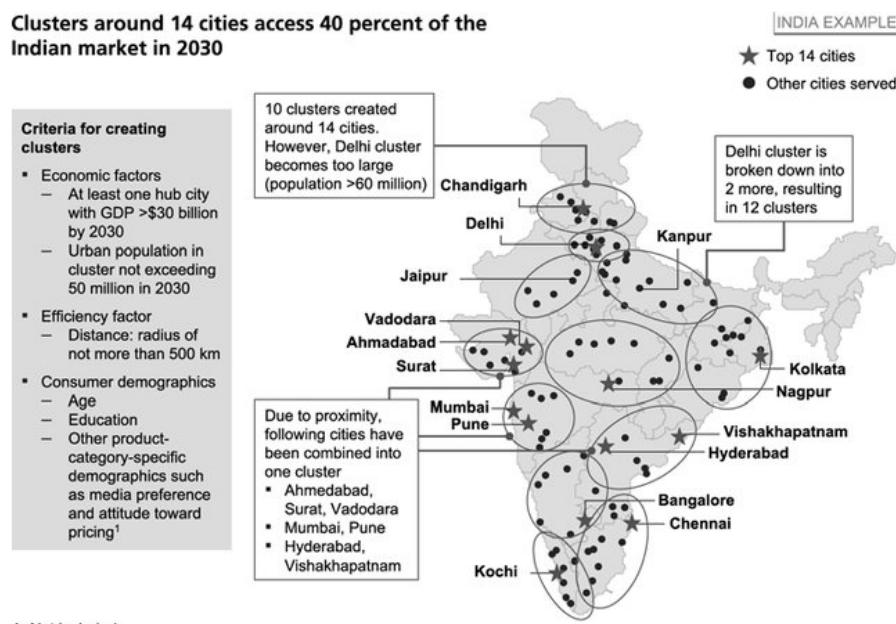
Answer

Bagging	Boosting
<ul style="list-style-type: none">• The original dataset is divided into multiple subsets, selecting observations with replacement.	<ul style="list-style-type: none">• The new subset contains the components misclassified by the previous model.
<ul style="list-style-type: none">• This method combines predictions that belong to the same type.	<ul style="list-style-type: none">• This method combines predictions that belong to the different types.
<ul style="list-style-type: none">• Bagging decreases variance.	<ul style="list-style-type: none">• Boosting decreases bias.
<ul style="list-style-type: none">• Base classifiers are trained parallelly.	<ul style="list-style-type: none">• Base classifiers are trained sequentially.
<ul style="list-style-type: none">• The models are created independently.	<ul style="list-style-type: none">• The model creation is dependent on the previous ones.



Random Forest is a decision tree-based algorithm that uses the bagging technique. It creates multiple decision trees on different subsets of the training data and combines them to make the final prediction. In Random Forest, each decision tree is trained on a randomly selected subset of the features and a randomly selected subset of the training data. This helps in reducing the variance of the individual decision trees and improves the model's performance. Therefore, Random Forest falls under the category of bagging ensemble techniques.

D. What is the scope of machine learning in this picture?



1 Not included.

Answer

The economies of agglomeration underline the laws and trends that determine the number, size, distribution, cluster or density of urban settlements. Human activities produce two main types of goods and services. One is lower order goods and services such as daily goods from grocery stores. The other is higher order goods and services such as sophisticated appliances from specialty stores. People are willing to travel only short distances to get certain lower order goods and services and to go further to get higher order goods and services. The people consumption preferences and behavior lead to the formation of urban systems comprising of urban centres of various sizes. Big cities offer a greater variety of higher order goods and services. There are few large cities (the sentence is hanging). There are more small towns and villages offering lower order goods and services. The emergence of central cities or places results in the clustering of a hierarchy of cities. Specific types of industries and businesses tend to cluster together to achieve maximum competitiveness. They form vertical and horizontal linkages with other industries that supply their inputs or market and sell their products or services.



COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

The formation of city clusters depends on a number of local factors such as topography, climate, transportation, technological facilities, and the personal preferences of consumers. In the long historical development, transportation and infrastructure played an important role in leading to the agglomeration of enterprises and business activities. Firms tend to aggregate in development nodes that were in turn linked to other nodes to form clusters. India, given its relatively early stage of urbanization, is an economy where 14 major clusters of cities capture significant shares of the country's population and GDP. These 14 urban agglomerations would cover 17 percent of the country's total population and 40 percent of India's total GDP in 2030. The scope of unsupervised machine learning in this context could be to identify patterns and clusters in the data related to the formation and distribution of urban settlements and the clustering of businesses and industries. Unsupervised learning algorithms can be used to identify similar patterns in the data, cluster them together, and analyze the factors that contribute to the clustering. This can help in understanding the local factors such as topography, climate, transportation, and technological facilities that lead to the agglomeration of businesses and industries. Furthermore, it can also help in identifying the different types of goods and services that are demanded in different areas and the distances people are willing to travel to obtain them. This information can then be used to develop targeted policies and strategies for urban development and economic growth. Therefore, unsupervised machine learning can be a useful tool in analyzing and understanding the complex systems of urbanization and economic agglomeration.

F. Match the terms with respect to their role in reinforcement learning.

5

Term	Role
1. Agent	a. A situation in which an agent is present or surrounded by.
2. Environment	b. An entity that can perceive/explore the environment and act upon it
3. Action	c. A situation returned by the environment after each action taken by the agent.
4. State	d. The moves taken by an agent within the environment.
5. Reward	e. A strategy applied by the agent for the next action based on the current state.
6. Policy	f. A feedback returned to the agent from the environment to evaluate the action of the agent.
7. Value	g. An expected long-term return with the discount factor and opposite to the short-term reward.
8. Q-value	h. Takes one additional parameter as a current action



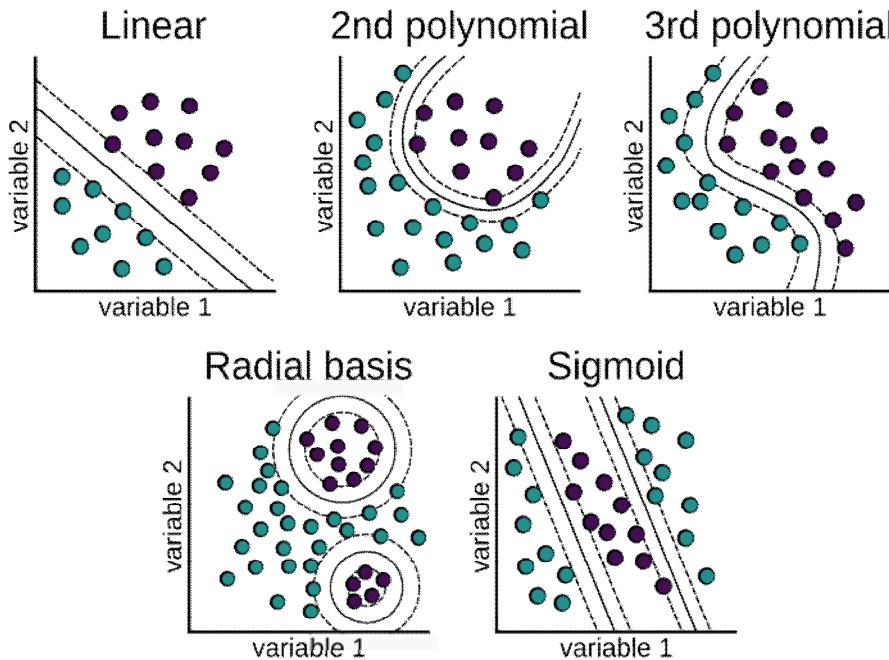
COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

Answer

Term	1	2	3	4	5	6	7	8
Role	b	a	d	c	f	e	g	h

- E. What's the significance of these kernelized models in SVM based classification/ regression? 5



Answer

Kernelized models in SVM-based classification/regression are significant because they allow for the nonlinear separation of data points in a high-dimensional space. In a linear SVM model, the decision boundary is a hyperplane that separates the data points. However, in real-world applications, the data is often nonlinearly separable. This is where kernelized models come in.

A kernel is a function that maps the data points from the input space to a high-dimensional feature space, where the data points are more easily separable. The SVM then finds a hyperplane in this feature space that maximally separates the data points. The kernel trick allows for this computation to be done efficiently, without explicitly computing the high-dimensional feature space.

The following are the most commonly used kernels in SVM-based classification/regression:

- **Linear Kernel:** The linear kernel computes the dot product between the input vectors in the original space, which corresponds to a linear decision boundary in the feature space.
- **Polynomial Kernel:** The polynomial kernel maps the input vectors to a higher-dimensional space using a polynomial function of a certain degree, allowing for curved decision boundaries. The 2nd, 3rd, and higher-degree kernels are commonly used.



- Radial Basis Function (RBF) Kernel: The RBF kernel maps the input vectors to an infinite-dimensional space using a Gaussian function, allowing for complex decision boundaries that can be non-linear and non-monotonic.
- Sigmoid Kernel: The sigmoid kernel maps the input vectors to a high-dimensional space using a sigmoid function, which can be used to model non-monotonic decision boundaries.

Q.6 Answer in one line. **(Any 6)**

6

A. [Which ML based technique will you recommend for Dahi Handi task during Janmashtami?](#)
Justify your answer.



Answer

Dahi Handi is a traditional festival in India where a group of people form a human pyramid to reach and break a pot filled with curd. This task requires a high level of coordination and teamwork, making it a challenging problem for machine learning techniques.

First solution

Reinforcement learning (RL) could be a suitable ML-based technique for the Dahi Handi task, as it could learn to optimize the strategy for forming a stable pyramid and breaking the pot of curd, taking into account a range of complex factors. It could learn to optimize the strategy for forming the pyramid, taking into account factors such as the strength and skill of each team member, the height of the pyramid, and the stability of the structure. It could also learn to adapt to different conditions, i.e. changes in the size of the pot or the no. of team members.



Second solution

The goal of using unsupervised learning (k-means or hierarchical clustering) in this context would be to gain insights into the coordination and movements of the team members during the Dahi Handi task. By analyzing the data using unsupervised learning techniques, we could identify patterns and clusters of team members who work well together and potentially use this information to optimize the team formation for future tasks. Another approach could be to use dimensionality reduction techniques, such as principal component analysis (PCA), to identify the most important features or movements of the team members during the task. The reduced dimensionality data could then be used as input to clustering algorithms to group the team members into clusters.

B. [How does deep learning differ from shallow learning?](#)

Answer

Deep learning models are characterized by their complexity, the use of multiple layers to extract increasingly complex features from the data, and the ability to learn automatically from large and complex datasets. Shallow learning models, in contrast, are typically simpler and more interpretable, but may not perform as well on very large or complex datasets.

Justification for understanding (not expected in exam)

Deep learning and shallow learning differ mainly in the architecture and complexity of the models used to solve a problem. Shallow learning, also known as traditional machine learning, typically involves models with a limited number of layers and a small number of parameters. These models are trained on relatively small datasets, and the training process often involves hand-crafted features that are designed to capture the important characteristics of the data.

In contrast, deep learning models consist of multiple layers of interconnected neurons that can learn to extract and represent increasingly complex features of the data. These models typically require much larger datasets to train effectively, and the feature engineering is often done automatically as part of the training process.

Another key difference between shallow and deep learning is the type of problems they are best suited for. Shallow learning is well suited for problems with a relatively small number of features and a well-defined set of classes or labels. These models can be trained quickly and can achieve high accuracy on many classification and regression tasks.

Deep learning, on the other hand, is better suited for problems with very large or complex datasets, such as image or speech recognition, natural language processing, and robotics. These models can learn to recognize patterns and features in the data that may not be apparent to humans and can achieve state-of-the-art performance on a wide range of tasks.



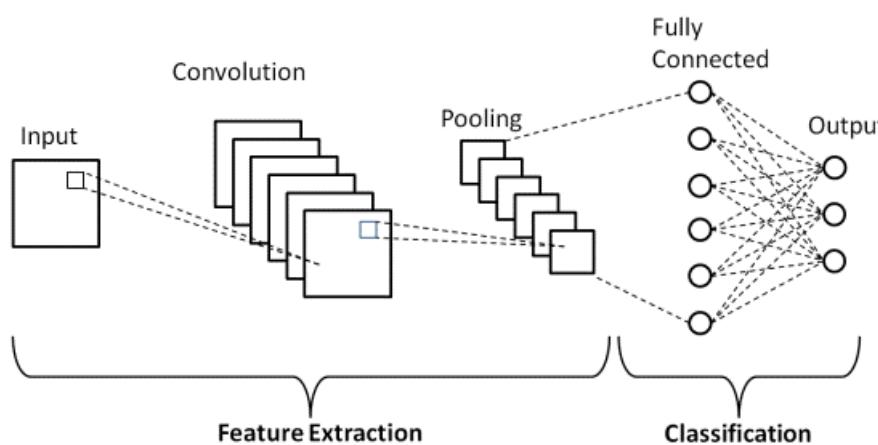
- C. How do neural networks work in deep learning?

Answer

In deep learning, neural networks can have many layers, which allow them to model increasingly complex relationships between inputs and outputs. These networks are called deep neural networks and can be used for a wide range of tasks, such as image recognition, natural language processing, and speech recognition. They use interconnected neurons to process input data and produce an output prediction. Through the process of training, the network learns to adjust its weights and biases to minimize the error between the predicted output and the actual output. With the ability to build deep neural networks, deep learning has become a powerful tool for solving complex machine learning problems.

- D. Draw a general labeled architecture of a Convolutional Neural Network.

Answer



- E. What's the use of Markov Decision process in reinforcement learning?

Answer

Markov Decision Process (MDP) provides a formal way of modeling a decision-making process, where an agent interacts with an environment in order to maximize a cumulative reward. It allows the agent to learn an optimal policy by interacting with a stochastic environment, where the outcome of an action is not always certain.

Justification for understanding (not expected in exam)

MDP models the environment as a set of states, and each state has a probability distribution of possible next states based on the action taken by the agent. By modeling the environment as an MDP, the reinforcement learning algorithm can learn an optimal policy, which is a set of rules that the agent follows to take actions based on the current state in order to maximize the expected cumulative reward.

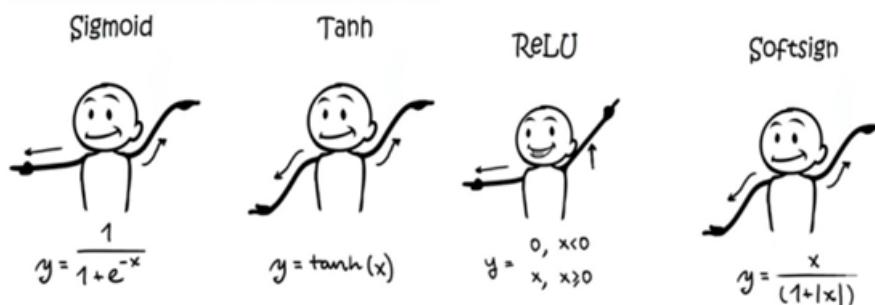


COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

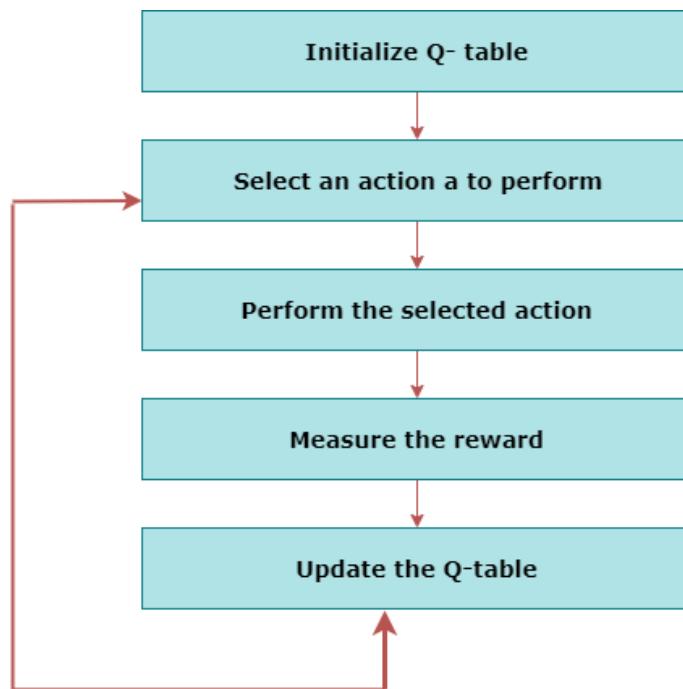
F. Identify activation functions.

Answer



G. Draw a flowchart showing steps in Q-learning.

Answer



H. What is the difference between KNN and K means?

Answer (any one strong and relevant difference will be awarded full marks)

- K-NN is a Supervised while K-means is an unsupervised Learning.
- K-NN is a classification or regression machine learning algorithm while K-means is a clustering machine learning algorithm.
- K-NN is a lazy learner while K-Means is an eager learner. An eager learner has a model fitting that means a training step but a lazy learner does not have a training phase.
- K-NN performs much better if all of the data have the same scale but this is not true for K-means.



- K-means is a clustering algorithm that tries to partition a set of points into K sets (clusters) such that the points in each cluster tend to be near each other. It is unsupervised because the points have no external classification.
 - K-nearest neighbors is a classification (or regression) algorithm that in order to determine the classification of a point, combines the classification of the K nearest points. It is supervised because you are trying to classify a point based on the known classification of other points.
- I. What is the maximum margin hyperplane for a dataset with 10 data points in a two-dimensional space using SVM?

Answer

Here we have 10 data points in a two-dimensional space, the maximum margin hyperplane will have **at most 8 support vectors** ($n-2$, where n is the number of data points).

Justification for understanding (not expected in exam)

To find the maximum margin hyperplane for a dataset with 10 data points in a two-dimensional space using SVM, we need to perform the following steps:

- Choose a kernel function. For this example, we will use a linear kernel.
- Train the SVM model on the data. The SVM algorithm will find the hyperplane that maximizes the margin between the two classes of data points.
- Identify the support vectors. These are the data points that are closest to the hyperplane and define the margin.
- Calculate the margin. The margin is the distance between the hyperplane and the closest support vectors on either side.
- Normalize the weights and bias of the hyperplane.

Assuming that the data is linearly separable, the maximum margin hyperplane is the hyperplane that separates the data with the largest possible margin. In a two-dimensional space, the maximum margin hyperplane is a line.

- Since we have 10 data points in a two-dimensional space, the maximum margin hyperplane will have at most 8 support vectors ($n-2$, where n is the number of data points).
- To calculate the maximum margin hyperplane, we need to perform the above steps. The details of the calculation are beyond the scope of this response, but the basic idea is to use optimization techniques to find the weights and bias that define the hyperplane.
- In practice, the SVM algorithm implemented in a machine learning library such as Scikit-learn can be used to find the maximum margin hyperplane for a given dataset.



- J. Given a dataset with 100 data points and 3 features, how many hyperplanes are there in a linear SVM with binary classification?

Answer

In the case of a linear SVM with binary classification on a dataset with 100 data points and 3 features, there is exactly **one hyperplane** that separates the data into two classes.

Justification for understanding (not expected in exam)

In a linear SVM with binary classification on a dataset with 3 features, the goal is to find a hyperplane that separates the data into two classes. A hyperplane in a 3-dimensional space is a 2-dimensional subspace, which can be defined by an equation of the form,

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + b = 0$$

where w_1, w_2, w_3 are the weights of the hyperplane corresponding to the three features x_1, x_2 , and x_3 , and b is a bias term.

The number of hyperplanes that can be defined in a 3-dimensional space is infinite, since there are an infinite number of possible combinations of weights and bias terms. However, in the context of an SVM, the goal is to find the hyperplane that maximizes the margin between the two classes of data points.

Assuming that the data is linearly separable, there is a unique hyperplane that maximizes the margin, and the SVM algorithm will find this hyperplane by solving an optimization problem. Therefore, in the case of a linear SVM with binary classification on a dataset with 3 features, there is exactly one hyperplane that separates the data into two classes.

- K. What is the maximum number of support vectors in a binary SVM with 100 data points?

Answer

- In the case of a binary SVM with 100 data points, **the maximum number of support vectors is therefore 99**, assuming that the data is perfectly separable with a linear decision boundary.

Justification for understanding (not expected in exam)

- A well-behaved dataset with a clear margin between the two classes, the number of support vectors is typically small.
- In fact, for a linearly separable dataset, the maximum number of support vectors is equal to the number of data points minus one.
- In the case of a binary SVM with 100 data points, the maximum number of support vectors is therefore 99, assuming that the data is perfectly separable with a linear decision boundary.
- However, in practice, the number of support vectors will depend on the complexity of the dataset and the regularization parameters used in the SVM training process.



COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra)

- L. Consider a dataset with three classes: red, blue, and green. Draw a scatter plot of the data points in a two-dimensional feature space, where the x-axis represents feature 1 and the y-axis represents feature 2. Suppose we want to train a support vector machine (SVM) to classify this data using both the one-vs-all and the one-vs-one multi-class strategies. Draw a diagram showing the decision boundaries and the support vectors for both strategies. Label the decision regions for each class and indicate which strategy is more suitable for this dataset.

