Name:

Harshit Wadhwani

Subject Code:

18AI88

University Seat Number:

1DT20AI021

Date:

11-05-2024

Question1

Machine learning is the branch of study that gives the ability to computers to learn. A program is said to lean from experience E with respect to some task T and some performance P if its performace on T imporaves with experience E.

3examples are -
Deteting tumors in brain scans, in this each pixel in the imag is classified using CNNs.
Creating a chatbot, this involves many NLP components including Natural Language Understanding and question answering.
Summarizing long documents automatically this branch of NLP called text summarization again using the tools.

Question2

Less quantity of training Data, it takes a lot of data for most machine learning algorithms to work properly For simple problems you typically need thousands of examples and for more complex problem you require millions of examples. Training Data should be representative of the new cases that you want to generalize to.
It is crucial to use a traing set that is representative of cases you want to generalize too. If the sample is too small, you will have sampling. If the sample are large samples can be non representative if the sampling method is flawed This is called sampling bias.

Bad Quality Data, if the training data is full of errors, outliers and noise, it will make it harder

for the system to detect the underlying patterns, so your system is less likely to perform well. It is often well worth the efforts to spend time cleaning up your training data. Most data scientist spend a significant part of their time doing just that example - If some instances are clearly outliers it may help to simply discard them or try to fix the errors manually. If some instances are missing a few features, you must decide whether you want to ignore this attribute altogether, ignore these instances, fill in the missing values or train one model with the feature and one model without it. The following are a couple of examples of when you want to clean up training data If some Instances are clearly outliers, it may help to simply discard them or try to fix the errors manually If some instances are missing a few feature you must decide whether you want to ignore altogether fill in the missing values or train one model with the feature.

Irrelevant features - the system will only be capable of learning if the training data contains enough relevant features and not too many irrelevant ones. A critical part in ML projects is coming up with good set of features to train on, this process is also called feature engineering and involves the following - feature selection, feature extraction, creating new features by gathering new data.

Overfitting - overfitting the training data In machine learning overfitting means that the model performs well on the training data, but it does not generalize well. overfitting happens when the model is too complex comparing to the amount and noisiness of training data some solution to this are using a simple model with fewer parameters

reducing the number of attributes in data, gathering more training data and reducing the noise.

Underfitting the training data., It is opposite of overfitting. It occurs when the model is too simple to learn the underlying structure of data. main option of fixing this problem are selecting a more powerful model with more parameter, feed better feature to the algorithm and reduce constraints on the model.

Question3

Voting classifier - suppose there are a few classifier and each one has 80% accuracy. there might be logistic regression classifier, svm classifier, KNN classifier and so on. a simple way to create a better classifier is to aggregate the predictions of each classifier and predict the class with most votes. This majority vote classifier is called Hard voting. This voting classifier achieves a higher accuracy than the best classifier in the ensemble. even if each classifier is a weak learner the ensemble can still be strong, provided they are sufficiently diversed.

Bagging - one way to get a diverse set of classifier is using different training algorithms another approach is to use the same training algorithm for every predictor and train them on different random subsets of traing sets. when sampling is performed with replacement this method is called bagging and without replacement is called pasting. both allow training instances to be sampled.

Several times across multiple predictors but only bagging allows them to be sampled several times for same predictor. once predictors are trained, ensemble can make prediction by aggregating prediction of all predictors. The aggregation function is typically statistical for classification or average for regression. out of box evaluation with bagging, some instances are sampled several times for a given predictor while others may not be sampled at all. By default a bagging classifier samples M training instances with replacement where M is the size of training set. Since the predictor never sees the oob instances during training. It can be evaluated on this without needing a separate validation set.

Question4

Supervised Learning:
These systems are trained with labeled data, where each input is associated with a corresponding target output. During training, the model learns the mapping between input and output, enabling it to make predictions or decisions when new data is presented.
Unsupervised learning:
Unsupervised algorithm deals with unlabelled data. Where the model attempts to find patterns or structures without explicit guidance. Common tasks include clustering similar data points together or reducing dimensionality of data.
Reinforcement Learning:
In this, an agent learns to interact with the environment by taking action and receiving feedback in the form of rewards or penalties. The goal is learning the optimal strategy or policy that maximizes

Question5

Supervised learning example includes classifying email as spam or not spam email. whereas unsupervised include using customer purchase history data to group them into segments.