Fall-2022

Q-1

A. What is random tree? Explain how attribute bagging technique address over-fetting Problems

⇒ A Random Tree is a decision tree algoritum where each node is split based on a random subset of features rather that the entire feature set. It can be seen as a variation of the standand decision tree, designed to reduce overfitting and variance.

Bagging is a technique to reduce over fetting in machine learning models, epecially for high-variance algorithms like decision tree. During the construction of each tree, only a random subset of attributes is considered at each spit rather than all feature.

① Each tree gets a different random subset of features, which leads to de-correlated models.

② By averaging, the variance and sensitivity to the noise in the data are reduced, leading to more generalized performance.

B. compare between RandomForest and Bagging algorithms.

Aspect	Random Forest	Bagging
Basic concept	Random Forest is a decision tree where random subset of feature used during tree creation.	In Bagging, multiple models are trained on random bootstrapped samples of data.
Feature subset selection	At each split, Random forest considers only a random subset of feature	In Bagging Approse, use all available features at each split but different data
Overfitting hadding	Less prone to overfitting due to randomness in feature selection, leading to less correlated trees	Help reduce offen fetting but Not too much as Random forest since all feature used
Training speed	Slower compared to bagging, as random feature selection adds complexity	Faster training compared to Random forest, as all feature are used, reducing complexity.

③ Random Forest vs AdaBoost ?

Factors	Random Forest	AdaBoost
Learning process	Parallel: Trees are build independently and simultaneously.	sequential: Tree are build one after another, each correcting the errors of previous one.
Model weight	All trees have equal weight in the final prediction	Trees are weighted based on their performance, better-performing trees have more influence.
time complexity	Fully grown decision tree with many split	simple decision stump with only one split.
strength of models	strong individual models	weak indivitual models, combined for better performance.
Main Focus	~~Building~~ Reducing Overfitting by building diverse tree.	Reduing error by focusing on the mistakes of previous trees.

(A) what is the purpose of ensemble method ? Discuss differene between different techniques.

=> The ensemble method in machaine learnings is a technique that combines predictions from multiple models to improve overall performance, such as accuracy, robustness, or stability, compared to individual model.

Different Ensemble Technique.

(1) Bagging ⟶ Random Forest.

(2) Bosting

(3) voting

Techique	Main goal	How models are combined	Strengths	weakness
Bagging	Reduce varian	independent model (voting/ avg.)	Reduce Variance, works on over fitting	Doesn't reduce Bias
Boosting	Reduce bias and error	sequential model	Handle bias and complex pattern well	Prone to over-fitting, slow training, noice
Voting	Improve prediction accuracy	Majority voting or averaging	simple to us, good with independent models	Doesn't optimize for best combination, lack model interaction

5 What are the basic difference between noice and outliers in ML?

⇒ Noise and outliers refer to irregular data point.

Noise refers to random errors or variations in data that do not present the underlying pattern. It's happened in data collection, measurement or other sources. Noice can degrade the performance of model. It's reduce by data preprocessing, smoothing, or filtering.

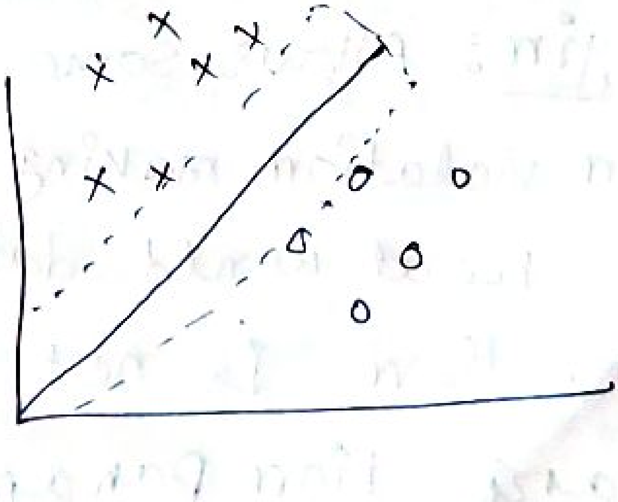
Outliers are data points that are significantly different from the majority of the data. These point maybe valid but unusual observations or errors. Outliers are isolated and do not follow the general pattern of the data. Outliers ~~hav~~ heavily skew models.

⑥ What is SVM ?

⇒ Support Vector Machine (SVM) is a powerful superwised machine learning ~~model~~ Algorithm used for classification task. It's also little bit use in Regression problem. SVM is know for it's effectiveness in handiling high-dimensional data and cases where the classes are not sperate linearly separable.

⑦ Linear Vs Non-linear SVM ?

Linear SVM: When the data is linearly separable, SVM finds a straight-line hyperplane that can perfectly divide the classes.



Non-Linear SVM: when the data is not linearly separable, SVM uses a technique called the kernel trick to project the data into a higher-dimensional space, where it becomes linearly separable.

⑧ Hard Margin Vs Soft Margin?

Hard Margin: Assumes that the data is perfectly separable by hyperplane, meaning no missclassification are allowed. Rarely used in practice since most real data contains noice.

Soft Margin: Allows some misclassification or margin violation, making it more applicable on real world data-set where perfet sepation is not possible. The regularization parameter C.

control trade-off between maximizing the margin and minimizing classification errors

⑨ what is Kernel Tricks?

⇒For non-linearly separable data, SVM uses the kernel tricks to map the data from its orginal space into a higher dimentional space, where linear separable canbe applied.

⑩ How C value works in soft margin

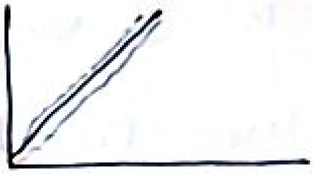
In the real world it's often not possible to perfectly separate classes without some missclassification. soft margin allows some data points to be within or on the ~~wro~~ wrong side of the margin.

C value controls how much penalty the model should assign to misclassified points. It determines the balance between

- ⊡ Maximizing the margin (Generalization)
- ⊡ Minimizing classification errors.(correctly classifying training points).

High value of C: with a large C, the svm become more sensitive to individual data points and will attempt to classify all data point correctly. As a result, It may create a very narrow margin around the ~~de~~ decision boundary.

If $c = 1000$(high)



it's ignore missclassified data points heavily.

low value of C: with a small C, the model is more relaxed about misclassification and focuses on maximizing the margin between classes, allowing some training points to fall on the wrong side of decision boundary.

if $C = 0.01$ (low)

It's tolenate mone ennons.

Slide

Possitive effect: when models are individually acunate, the ensemble can connect and improve overall pertormance.

No effect: When models are too similar (lack diversity) and highly accunate, the ensemble add little value, resulting no significant improvement.

Negative effect:- when an ensemble includes a model with low accunacy, even if other models are diverse, the overall pertormance can suffer, leading to a decline in accunacy

Q) What are the requirements to acheive an improved model?

⇒ Base model should be

accurate and Diverse.

Q) How to introduce diversity among model?

(i) By using Different Algorithms.

(ii) same algorithm but training the model on different subsets of data.

(iii) combination of both.

Q) Two model to construct Random Forest:

Forest-RI (random input selection): Randomly select, at each node, F attributes as candidates

Forest-RC (random linear combination): Creates new attributes on features that are linear combination of existing attributes.

⑩ Between Random forest and Adaboost, which method is more susceptible to outliers/Noice in dataset? Briefly explain your answer.