Assignment - 1C

Bagging ~~vs~~ and Boosting

Bagging and boosting are two type of Ensemble ~~language~~ learning. These two decrease the variance of a single estimate as they combine several estimates from different models. So the result may be a model with higher stability

Bagging:-

Bootstrap Aggregation, also known as bagging is a machine learning. essemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithm used in ~~statical~~ statistical classification and regression. It decrease the variance and helps to avoid overfitting. It is usally applied to decision tree method. Bagging is a special case of the model averaging approch.

Description of the technique

Suppose a set D of d tuples, at each iteration i, a training set Di of d tuple is selected via row sampling with a replacement method (i.e there can be repetitive element from different d tuples) from D (i.e. bootstrap). Then a classifer model Mi is learned for each traning set D<i. Each classifer Mi returns its class prediction. The bagged classifier M* counts the votes and assign the class with most vote to x (unknown sample)

Implemention steps of Bagging

step1: Mutiple subset are created from the orignal ~~at~~ data set with equal tuple; selecting observation with replacement.

Step2: A base model is ~~learned~~ created on each of these subset

Step3: Each model is learned in parallel with each training set and independ of each other.

Step4: The final prediction are determined by combining the prediction from all the model.

Example of bagging

The Random Forest model uses Bagging, where decision tree models with the higher varience are present. It make random ~~forest~~ feature selection to grow trees make a random forest

---

Boosting

Boosting is an ensemble modeling technique designed to create a strong classifer by combining multiple weak classifiers. The process involves building model aims to correct the error made by the previous one.

- Intially a model is built using the training data.
- Subsequent models are then trained to address the mistake of their predecessors.
- boosting agent assigns weights to the data points in the orignal dataset

- Heigher weight : Instance that were misclassified by the ~~per~~ previous model receive higher weight.
- lower weights: Instance that were correctly classified receive lower weights
- Training on weighted data: The ~~st~~ subsequent model learns from the weighted dataset, focusing its attention on hard to learn examples (those with higher weights)
- This ~~ite~~ iterative process continues until:
  The entire training dataset is accurately predicted, or
  A predefined maximum number of models is reached.

Algorithm:

1 Initialise the dataset and assign equal weight on each of the data poi
2 Provide this as input to the model and identify the wrongly classified data points
3 Increase the weight of the wrongly classified data point ~~correctly~~ and decrease the weight of correctly classified data points. And then normalize the weights of all data points
4 if (got required results)
  Goto step 5
  else
  Goto step 2
5 End

Similarities between bagging and boosting

1. Both are ~~ens~~ ensemble method to get N learners from 1 learner
2. Both generate several training data set by random sampling
3. In Bagging, the final decision is made by averaging predictions or majority voting. In boosting, new model focus on correcting previous errors, and the final decision is made by a weighted majority vote.
4. Both are good at reducing variance and provide higher stabilit

Difference between Bagging and boosting

SN	Bagging	SN	Boosting
1	The simplost way of combining predictions that belong to the same type.	1	A way of combining prediction that belong to the different types.
2	Aims to decrease variance, not bias	2	Aims to decrease ~~variance~~ bias, not variance
3	Each model receive equal weight	3	Models are weighted according to their performance.
4	Each model is build indendently	4	New model are influenced by the performance of perviously build model.
5	Different training data subset are selected using row sampling with replacement and random sampling method from the entire training dataset	5	Iteratively train models, with each ~~row~~ new model focusing on correcting the errors (missclasification or high residual) of the ~~pe~~ previous model
6	Bagging tries to solve the over-fitting problem	6	Boosting tries to reduce bias.
7	In this base classifier are trained parallely	7	In this base classfier are trained sequentially
8)	Example: The Random Forest model uses bagging	8	The Adaboost uses boosting techniques.

K-Mean (Numerical)

~~Ex~~ Question

A1 (2,10), A2 (2,5), A3 (8,4), B1 (5,8), B2 (7,5), B3 (6,4), C1 (1,2), C2 (4,9)

Initial centroid (Iteration 1)

A = (2, 10)

B = (5,8)

C = (1,2)

Point	~~From~~ x	y	dis A	dis B	dis C	cluster	new cluster
A1	2	10	0.00	3.606	8.062		A
A2	2	5	5.00	4.243	3.162		C
A3	8	4	8.485	5.000	7.280		B
B1	5	8	3.606	~~3.606~~ 0.000	7.211		B
B2	7	5	7.071	~~4.123~~ 3.606	6.708		B
B3	6	4	7.211	4.123	5.385		B
C1	1	2	8.062	7.211	0.000		C
C2	4	9	2.236	1.414	7.616		B

Recompute centroid

A = mean ({(2,10)}) = (2.0, 10.0)

B = mean ({(8,4), (5,8), (7,5), (6,4), (4,9)}) = ~~(6.0)~~ (6.0, 6.0)

C = mean ({(2,5), (1,2)}) = (1.5, 3.5)

from distance formula = $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ and

centroid from = $\frac{x}{\Sigma x}, \frac{y}{\Sigma y}$

A = (2,10)
B = (6,6)
C = (1.5, 3.5)

A = (3, 9.5)
B = (6.5, 5.25)
C = (1.5, 3.5)

Point	x	y	disA	dis B	dis C	cluster	New cluster
A1	2	10	0.00	~~4.123~~ 5.658	6.519	A	A
A2	2	5	4.5	4.123	1.581	C	C
A3	8	4	8.485	2.828	6.519	B	B
B1	5	8	3.605	2.236	5.700	B	B
B2	7	5	7.071	1.414	5.700	B	B
B3	6	4	7.211	2.000	4.527	B	B
C1	1	2	8.062	6.403	1.581	C	C
C2	4	9	2.236	3.605	6.041	B	A

A = ({(2,10) ~~(2,5) (8,4)~~ (4,9)}) = ~~(4, 4.75)~~ (3, 9.5)
B = ({(8,4), (5,8), (7,5), (6,4), (4,9)}) = ~~(6.0, 6.0)~~ (6.5, 5.25)
C = ({(2,5), (1,2)}) = (1.5, 3.5)

Point	x	y	dis A	dis B	dis C	cluster	New cluster
A1	2	10	1.11	6.54	6.52	A	A
A2	2	5	4.60	4.51	1.58	C	C
A3	8	4	7.43	1.95	6.52	B	B
B1	5	8	2.50	3.13	5.70	B	A
B2	7	5	6.02	0.56	5.70	B	B
B3	6	4	6.26	1.35	4.53	B	B
C1	1	2	7.76	6.39	1.58	C	C
C2	4	9	1.12	4.51	6.04	A	A

New centroid

A = ({(2,10), (5,8), (4,9)}) = (2+5+4/3, 10+8+9/3) = (3.66, 9)
B = ({(8,4), (7,5), (6,4)}) = (8+7+6/3, 4+5+4/3) = (7, 4.33)
C = ({(2,5), (1,2)}) = (2+1/2, 5+2/2) = (1.5, 3.5)

Point	x	y	dis A	dis B	dis C	cluster	new cluster
A1	2	10	1.94	7.56	6.52	A	A
A2	2	5	4.83	5.04	1.58	C	C
A3	8	4	6.62	1.05	6.52	B	B
B1	5	8	1.67	4.18	5.70	A	A
B2	7	5	5.21	0.67	5.70	B	B
B3	6	4	5.52	1.05	4.53	B	B
C1	1	2	7.49	6.44	1.58	C	C
C2	4	9	0.33	5.55	6.04	A	A

since the cluster and new cluster are equal it is the optimal solution.

Where

$A = \{(3.66, 9)\}$

$B = \{(7, 4.33)\}$

$C = \{(1.5, 3.5)\}$

Machine learning model

Machine learning model is computational program that learns patters from data and makes decision or prediction on new unseen data. It is created by training a machine learning alogrithm on a dataset and optimizing its to a minimize errors.

Machine learning models can be boardly catgorized into four primary paradigm based on the nature of data and learning data.

1 Supervised learning Models
2 Unsupervised learning model
3 Semi- Supervied learning model
4 Reinforcement learning model

1) Supervised learning model

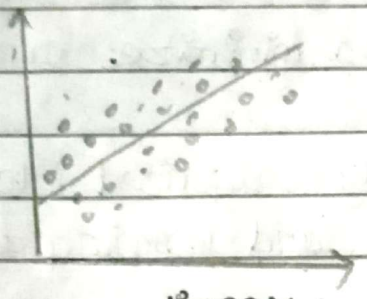
Supervised learning models learn from labeled data, where each input has a known output. The goal is to map input feature to the correct target value using a mathmatical model.

A Regression:- Regression models predict continuous numerical value rather than categories. Some of its alogrithm are

1 linear regression
2 Polynomical regression
c Decision Tree regression
d Random forest regression
e Support vector Pregression

Linear Regression

linear regression analysis is used to create a model that describes the relationships between a dependent variable and one more independent variable.



linear

For eg:- we want to predict a student's exam score based on how many hours they studied. we observe that as students study more hours, their scores go up In the example of predicting exam score based on hours studied. Here

- Independent variable (input): Hours studied because it's the factor we control or observe
- Dependent variable (output): Exam score because it depends on how many hour were studied.

Advantages

1) linear regression is computationally efficient and can handle large datasets effectively.
2) It can be trained on large dataset making it suitable for real-time application
3) linear regression is ~~computationally~~ relatively ~~efficient~~ robust to ~~and~~ ~~can~~ outlier compared to other machine learning algorithm

limitation

1) linear regression is ~~susception~~ suscepitible to both overfitting and underfitting.

2) It provide limited explantory power for complex relationship between variable.

3) It can't perform well if the relationship is not linear.

B classification classification models assign input data to predefined categories. Some of its algorithm are:-

a) Support vector machine (SVM)

b) Decision tree

c) Random forest

d) Naives ~~Bias~~ Bayes

e) K-Nearest Neighbour (KNN)

a) Support vector machine (SVM):-

It is a powerful supervised machine learning algorithm primary used for classification problems, but they can also be applied to regression. It tries to find the best boundary known as hyperplane that seperates different classes in~~to~~ the data. It is useful when you want to do binary classification like spam vs not spam or cat vs dog.

Advantages

1 High-dimension performance

2 Non-linear Capability

3 Outlier Resiliance

4 Binary and multiclass support

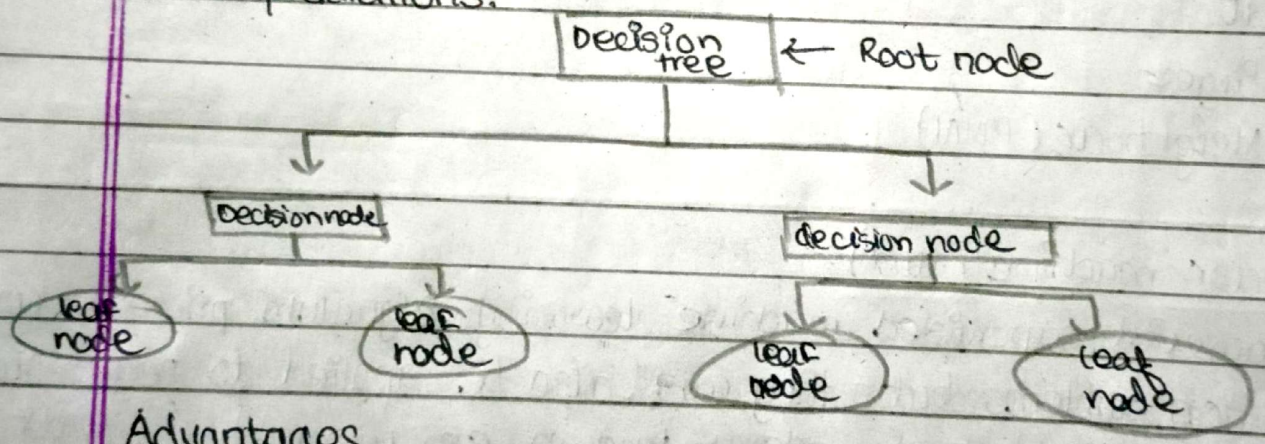
5 Memory efficiency

Disadvantage

1. Slow training
2. Parameter tuning difficulty
3. Noise Sensitivity
4. limited Interpreability
5. Feature Scaling sensitivity :

4) Decision tree

A decision tree helps us to make decision by mapping out different choices and their possible outcomes. It used in machine learning for task like classification and predicitions.



Advantages

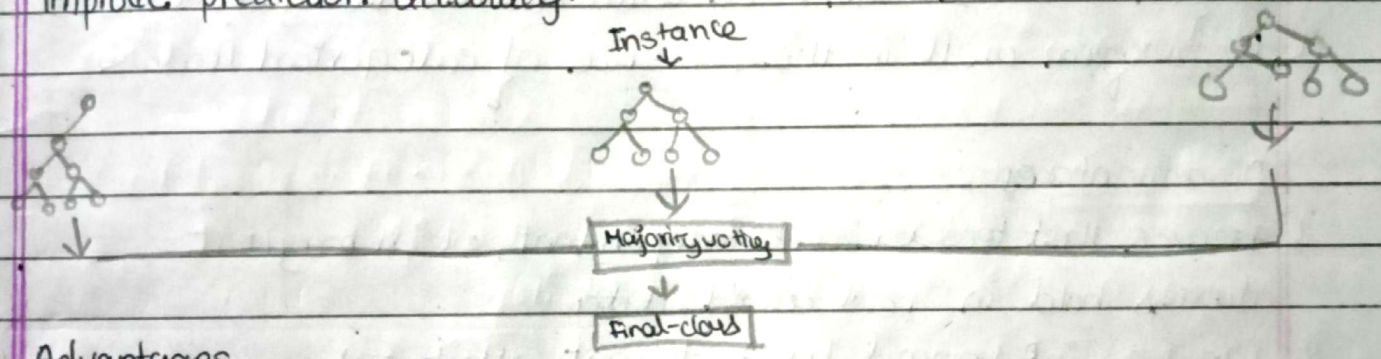
1. Easy to understand
2. Versatility.
3. No Need for feature scaling
4. Handel Non-linear relationships
5. Interpretability
6. Handles Missing data

Disadvantage

1 Overfitting
2 Instability
3 Bias toward feature with Many categories
4 Difficulty in Capture Complex Instruction
5 Computationally Expensive for large dataset

c Random Forest

Random forest algorithm is an esemble learning method for classification and regression task consisting of many decision trees (to 'forest') to improve prediction accuracy.



Advantages

1 Random forest provide very accurate ~~predicti~~ prediction even with large dataset
2 It can handel missing data well without compromising with accuracy
3 It doesn't require normalization or standization on dataset
4 when we combine multiple decision tree, it reduces the risk of overfitting problem

limitation

1 It can be computationally expensive espacially with large number of tree
2 It's hard to interpret the model compared to simpler models like decision trees.

d Naive Bayes

It is a machine learning classification algorithm that predict the category of a data point using probability. It perform well in many real-word application such as spam filtering, document categorisation and sentiment analysis.

Advantage

1 Easy to implement and computionally efficient
2 Effitive in case with a large number of features.
3 Perform well ~~in~~ the even with limited traing data.
4 It perform well in the presence of categorical feature

Disadvantages

1 Assume that feature are independent, which may not always hold in real-world data
2. Can be influenced by irrelavent attributes
3 May assign zero probability to unseen events, leading to poor generalization.

CamScanner

e. k-Nearest Neighbour (KNN) Algorithm

It is a simplest yet powerful machine learning algorithm used for classification and regression task. It works ~~for~~ by finding the "k" nearest neighbour of a new data point and classifing or predicting its value based on the majority or average of those ~~measure~~ neighbours.

Advantages

1. Simple to use
2. No training step
3. Few parameter
4. Versatile

Disadvantage

1. Slow with large data
2. Struggle with many features
3. Can overfit

Unsupervied learning Models

Unsupervied learning model work with unlabeled data, discovering hidden patterns, calters or structures without predefined outputs.

Clustering:- Group similar data into ~~k~~ cluster ~~using centroid~~ based on feature similarity. Some of its algorithm are:-

1 K-Mean
2 Hierarchical Clusting

1 K-Mean

K-Mean clustering group similar data point into cluster without needing labeled data. It is used to ~~co~~ uncover hidden pattern when the goal is to orginize data based on similarity

limitation:-

a choosing the right number of cluster (k)
b Sensitive to Initial centroids
c Non-spherical cluster
d outliers.

2 Hierarchical clustering

It is an unsupervised learning method used to group similar ~~group~~ data point into clusters based on their distance simalinty.

Dimensionality Reduction :- Reduction high-~~demensity~~ dimensional data while retaining important information for analysis or visulization Some of the technique are:-

1) PCA (Principal componet Analysis): Project data into principal component to reduce dimension
2 LDA (linear discriminant Analysis): Maximize class separability while reducing dimensionality

Advantages of supervised learning
1) Simplicity & clarity
2 High accuracy
3 versality
4 Generalization
5 wide application

Disadvantage of supervised learning
1 Required labeled data
2 Bias from data
3 overfitting risk
4 limited adaptability
5 Not scable for some problem

Advantages of unsupervised learning

a. No need for labeled data
b. Discovers hidden patterns
c. Handels complex and large datasets
d. Useful for anomly detection

Challenges of unsupervised learning

a. Noisy data
b. Overfitting Risk
c. limited Guidance
d. Cluster Interpretability