Assignment 1 C

Date
Page No. 1

Bagging and boosting

-> Bagging and Boosting are ensemble learning techniques used in machine learning. Ensemble learning means combining multiple models to improve overall performance. Both methods help in increasing accuracy and reducing errors, but they work in different ways.

1) Bagging (Bootstrap Aggregating)
Bagging is an ensemble technique in which multiple models are trained independently using different random samples of the dataset and their results are combined.

Working of Bagging

i)-> From the orijnal dataset, multiple random samples are created using bootstrap sampling (sampling with replacement)

ii)-> A seperate model is trained on each sample

iii)-> All models work in parallel.

iv)-> The final output is obtained by:
- voting (for classification)
- Averaging (for regression)

Advantages of Bagging

→ Reduces overfitting
→ Improves model stability
→ Works well with high-variance models like decision trees

Example: Random forest is a popular algorithm based on bagging.

Disadvantages of Bagging

→ Requires more computation due to multiple models

2) Boosting

Boosting is an ensemble technique where models are trained sequentially, and each new model focuses on correcting the errors of the previous model.

Working of Boosting

i) Initially all data points are given equal importance.
ii) A model is trained and errors are identified.
iii) More weights is given to misclassified data.
iv) The next model focuses more on these difficult data points.
v) Final prediction is made by weighted voting of all models.

Advantages of Boosting

- Reduces bias
- Improves prediction accuracy
- Works well with weak learners

Examples: AdaBoost, Gradient Boosting, XGBoost etc.

Disadvantages of Boosting:

- Can overfit if data is noisy
- Slower than bagging due to sequential training.

Conclusion:

Both Bagging and Boosting are important ensemble techniques. Bagging is mainly used to reduce variance and avoid overfitting, while Boosting focuses on improving accuracy by learning from previous mistakes. Choosing the right method depends on the dataset and problem type.

2) K-means problem:

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

Let's suppose k=3 and initial centroids are A(2,10), ~~B(2,5)~~ B(5,8) and C(1,2)

Points	X	Y	Dist. to A	Dist. to B	Dist. to C	Cluster	New clust.
A_1	2	10	0	$\sqrt{13}=3.61$	$\sqrt{65}=8.06$		A
A_2	2	5	5	$\sqrt{18}=4.24$	$\sqrt{10}=3.16$		C
A_3	8	4	$\sqrt{72}=8.49$	5	$\sqrt{53}=7.28$		B
B_1	5	8	$\sqrt{13}=3.61$	0	$\sqrt{52}=7.21$		B
B_2	7	5	$\sqrt{50}=7.07$	$\sqrt{13}=3.61$	$\sqrt{45}=6.70$		B
B_3	6	4	$\sqrt{52}=7.21$	$\sqrt{20}=4.47$	$\sqrt{29}=5.39$		B
C_1	1	2	$\sqrt{65}=8.06$	$\sqrt{52}=7.21$	0		C
C_2	4	9	$\sqrt{5}=2.24$	$\sqrt{2}=1.41$	$\sqrt{58}=7.62$		B

Cluster assignment

A → A_1

B → A_3, B_1, B_2, B_3, C_2

C → A_2, C_1

New centroids

A → (2, 10)

B → $\left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right)$

$= (6, 6)$

C → $\left(\frac{2+1}{2}, \frac{5+2}{3}\right)$

$= (1.5, 3.5)$

Point	X	Y	Dist. to A	Dist to B	Dist. to C	Cluster	New cluster
A_1	2	10	0	5.61	6.52	A	A
A_2	2	5	5	4.12	1.58	C	C
A_3	8	4	8.48	2.83	6.67	B	B
B_1	5	8	3.61	2.24	5.52	B	B
B_2	7	5	7.07	1.41	5.87	B	B
B_3	6	4	7.21	2	4.74	B	B
C_1	1	2	8.06	6.40	1.58	C	C
C_2	4	9	2.24	3.61	6.20	B	A

Cluster assignment

A → A_1, C_2

B → A_3, B_1, B_2, B_3

C → A_2, C_1

New centroids

A → $\left(\frac{2+4}{2}, \frac{10+9}{2}\right) = (3, 9.5)$

B → $\left(\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}\right) = (6.5, 5.25)$

C → $\left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

Date
Page No. 6

Points	X	Y	Dist. to A	Dist. to B	Dist to C	Cluster	New clust.
A_1	2	10	1.50	6.53	6.52	A	A
A_2	2	5	4.50	4.50	1.58	C	C
A_3	8	4	7.43	1.225	6.52	B	B
B_1	5	8	2.50	3.13	5.70	B	A
B_2	7	5	6.02	0.56	5.70	B	B
B_3	6	4	6.26	1.35	4.54	B	B
C_1	1	2	7.76	6.39	1.58	C	C
C_2	4	9	1.11	3.51	6.04	A	A

Cluster assignment:

A → A_1, C_2, B_1

B → A_3, B_1, B_2, B_3

C → A_2, C_1

Cluster assignment did not change from the previous step so, the final clusters are

Clusters	Centroid	Points
A	(3, 9.5)	A_1, C[A_1(2,10), C_2(4,9)]
B	(6.5, 5.25)	A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4)
C	(1.5, 3.5)	A_2(2,5), C_2(4,9)

New centroids

$$A = \left(\frac{2+5+4}{3}, \frac{10+8+9}{3}\right) = (3.6, 9)$$

$$B = \left(\frac{8+7+6}{3}, \frac{4+5+4}{3}\right) = (7, 4.3)$$

$$C = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$$

Points	X	Y	Dist. to A	Dist to B	Dist to C	cluster	New cluste
A1	2	10	1.95	7.55	6.52	A	A
A2	2	5	4.34	3.05	1.58	C	C
A3	8	4	6.61	1.05	6.52	B	B
B1	5	8	1.66	4.00	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	5.56	6.04	A	A

Since new cluster did change from previous cluster.
So, final clusters are

Clusters	centroid	Points
A	(3.6, 9)	A1(2,10), B1(5,8), C2(4,9)
B	(7, 4.3)	A3(8,4), B2(7,5), B3(6,4)
C	(1.5, 3.5)	A2(2,5), C1(1,2)

3) Supervised learning models

=> Classification model

Advantages:

- High prediction accuracy when sufficient labeled data is available.
- Clear performance evaluation using metrics (accuracy, precision, recall, F1-score)
- Effective for tasks like spam detection, image classification, disease prediction etc..
- Many supervised algorithms like decision trees provide insights into feature importance, making it easier to understand model decisions and iterate on improvements

Disadvantages

- Requires large amount of labeled data, which is costly and time consuming.
- Overfitting can occur (especially Decision Trees)
- performance drops if data is noisy or imbalanced.

Date
Page No. 9

⇒ Regression models

Advantages
- Predicts continuous values effectively
- Easy to interpret (especially linear regression)
- Useful for forecasting tasks (house prices, stock prices)

Disadvantages
- Sensitive to outliers (linear regression)
- Assumes relationship between variables (may not hold true)
- Complex models may overfit

Unsupervised learning models

1) Clustering models

Advantages
- No labeled data is required
- Useful for discovering hidden patterns
- Simple and computationally efficient
- Widely used for customer segmentation and anomaly detection

Disadvantages
- Need to predifine number of clusters (k)
- Sensitive to initial centroid selection
- Struggles with non-spherical or uneven cluster sizes

2) Dimensionality Reduction

Advantages
- Reduces complexity of high-dimensional data
- Improves visualization and performance
- Removes redundant features

Disadvantages
- Possible loss of important information
- Reduced interpretability