IDS-Unit-5

18 June 2025 23:01

1. What is Principal Component Analysis (PCA)? How is it used for dimensionality reduction?

Principal Component Analysis (PCA) is a powerful dimensionality reduction technique used in data science and machine learning to simplify complex datasets while preserving as much variability (information) as possible.

From <https://copilot.cloud.microsoft/?fromcode=bingchat&redirectid=823852B9497E4A72B8F26930484FC937>

What Is PCA?

PCA transforms the original features into a new set of uncorrelated variables called principal components. These components are ordered so that the first few retain most of the variation present in the original dataset.

Key Concepts:

- Principal Components: Linear combinations of original features.
- Variance Maximization: PCA finds directions (components) that maximize variance.
- Orthogonality: Components are orthogonal (uncorrelated).

How PCA Is Used for Dimensionality Reduction

Step-by-Step Process:

1. Standardize the Data
   Ensure all features have mean = 0 and standard deviation = 1.
2. Compute the Covariance Matrix
   Measures how features vary together.
3. Calculate Eigenvalues and Eigenvectors
   Eigenvectors determine the direction of components; eigenvalues indicate their magnitude (importance).
4. Select Top k Components
   Choose the top components that explain most of the variance (e.g., 95%).
5. Transform the Data
   Project the original data onto the selected components.

Benefits of PCA:

- Reduces computational cost and complexity.
- Helps visualize high-dimensional data.
- Removes multicollinearity.
- Improves model performance by eliminating noise.

Example Use Case:

In image processing, PCA can reduce the number of pixels (features) while retaining the essential structure of the image, making it faster to process and classify.

==========

2. Discuss different techniques for evaluating clustering performance. What are the internal and external evaluation metrics?

Evaluating clustering performance is essential to determine how well the algorithm has grouped the data. Unlike supervised learning, clustering lacks ground truth labels, so evaluation techniques are divided into internal and external metrics.

Internal Evaluation Metrics

These assess the quality of clustering based on the data itself, without external labels.

1. Silhouette Score

- Measures how similar an object is to its own cluster vs. other clusters.
- Range: -1 to 1 (higher is better).
- Good for evaluating cluster cohesion and separation.

2. Davies-Bouldin Index

- Measures average similarity between each cluster and its most similar one.
- Lower values indicate better clustering.

3. Dunn Index

- Ratio of the minimum inter-cluster distance to the maximum intra-cluster distance.
- Higher values indicate better separation.

4. Within-Cluster Sum of Squares (WCSS)

- Measures compactness of clusters.
- Used in the Elbow Method to determine optimal number of clusters.

◇ External Evaluation Metrics

These compare clustering results to ground truth labels (if available).

1. Adjusted Rand Index (ARI)

- Measures similarity between predicted clusters and true labels.
- Adjusted for chance; range: -1 to 1.

2. Normalized Mutual Information (NMI)

- Measures mutual dependence between clusters and true labels.
- Range: 0 to 1 (higher is better).

3. Fowlkes-Mallows Index

- Geometric mean of precision and recall between predicted and true clusters.

4. Purity

- Measures the extent to which clusters contain a single class.
- Simple but doesn't penalize for too many clusters.

☑ Summary Table

Metric Type	Metric Name	Ideal Value	Description
Internal	Silhouette Score	Close to 1	Measures cohesion and separation
Internal	Davies-Bouldin Index	Close to 0	Measures cluster similarity
Internal	Dunn Index	High	Measures cluster separation
External	Adjusted Rand Index	Close to 1	Compares clustering with true labels
External	Normalized Mutual Info	Close to 1	Measures shared information
External	Purity	High	Measures class homogeneity in clusters

==========

3. Discuss the evaluation metrics used to assess the performance of Regression models.

Evaluating the performance of regression models is essential to understand how well the model predicts continuous outcomes. Here are the most commonly used regression evaluation metrics, explained in detail:

From <https://copilot.cloud.microsoft/?fromcode=bingchat&redirectid=823852B9497E4A72B8F26930484FC937>

1. Mean Absolute Error (MAE)

- **Definition:** Average of the absolute differences between predicted and actual values.
- **Formula:**
  $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- **Interpretation:** Lower MAE indicates better performance. It treats all errors equally.

Screen clipping taken: 18-06-2025 23:08

2. Mean Squared Error (MSE)

- **Definition:** Average of the squared differences between predicted and actual values.
- **Formula:**
  $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Interpretation:** Penalizes larger errors more than MAE. Sensitive to outliers.

Screen clipping taken: 18-06-2025 23:08

3. Root Mean Squared Error (RMSE)

- **Definition:** Square root of MSE.
- **Formula:**
  $RMSE = \sqrt{MSE}$
- **Interpretation:** Same units as the target variable. Easier to interpret than MSE.

Screen clipping taken: 18-06-2025 23:09

4. R-squared (Coefficient of Determination)

- **Definition:** Proportion of variance in the dependent variable explained by the model.
- **Formula:**
  $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$
- **Interpretation:**
  - $R^2 = 1$ → Perfect fit
  - $R^2 = 0$ → Model explains none of the variance
  - Can be negative if the model performs worse than a horizontal line

Screen clipping taken: 18-06-2025 23:09

5. Adjusted R-squared

- Definition: Modified version of R^2 that adjusts for the number of predictors.
- Interpretation: Useful when comparing models with different numbers of features.

From <https://copilot.cloud.microsoft/?fromcode=bingchat&redirectid=823852B9497E4A72B8F26930484FC937>

🔹 6. Mean Absolute Percentage Error (MAPE)

- **Definition:** Average of absolute percentage errors.
- **Formula:**

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **Interpretation:** Expresses error as a percentage. Not suitable when actual values are close to zero.

Screen clipping taken: 18-06-2025 23:10

==========

Describe the DBSCAN clustering algorithm. How is it different from K-Means?

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based clustering algorithm that groups together points that are closely packed and marks points in low-density regions as outliers.

How DBSCAN Works

4. Parameters:

- ε (epsilon): Radius to search for neighboring points.
- MinPts: Minimum number of points required to form a dense region.

5. Steps:

- Pick a point.
- If it has MinPts neighbors within ε, it becomes a core point.
- All points within ε of a core point are part of the same cluster.
- Points not reachable from any core point are labeled as noise.

Advantages of DBSCAN

- Can find arbitrarily shaped clusters.
- Automatically detects outliers.
- No need to specify the number of clusters in advance.

❌ Limitations

- Sensitive to the choice of ε and MinPts.
- Struggles with varying density clusters.

DBSCAN vs. K-Means

Feature	DBSCAN	K-Means
Cluster Shape	Arbitrary shapes	Spherical clusters
Number of Clusters	Not required	Must be specified
Outlier Detection	Yes (labels noise)	No (forces all points into clusters)
Scalability	Slower on large datasets	Faster and scalable
Sensitivity	Sensitive to ε and MinPts	Sensitive to initial centroids

Example Use Case:

- DBSCAN: Ideal for spatial data like GPS coordinates, where clusters may not be circular.
- K-Means: Works well for customer segmentation when clusters are well-separated.

==========

5. Explain different techniques to evaluate Clustering Performance

Evaluating clustering performance is essential to determine how well the algorithm has grouped the data. Since clustering is typically unsupervised, evaluation techniques fall into two main categories:

◇ 1. Internal Evaluation Metrics

These assess clustering quality based on the data itself, without using external labels.

☑ Common Internal Metrics:

Metric	Description	Ideal Value
Silhouette Score	Measures how similar a point is to its own cluster vs. other clusters.	Close to 1
Davies-Bouldin Index	Measures average similarity between each cluster and its most similar one.	Close to 0
Dunn Index	Ratio of minimum inter-cluster distance to maximum intra-cluster distance.	Higher is better
Calinski-Harabasz Index	Ratio of between-cluster dispersion to within-cluster dispersion.	Higher is better

◇ 2. External Evaluation Metrics

These compare clustering results to known ground truth labels (if available).

☑ Common External Metrics:

Metric	Description	Ideal Value
Adjusted Rand Index (ARI)	Measures similarity between predicted clusters and true labels, adjusted for chance.	Close to 1
Normalized Mutual Information (NMI)	Measures mutual dependence between clusters and true labels.	Close to 1
Fowlkes-Mallows Index	Geometric mean of precision and recall.	Close to 1
Purity	Measures the extent to which clusters contain a single class.	Close to 1

◇ 3. Visualization Techniques

These help visually assess clustering quality:

- t-SNE or PCA plots: Reduce dimensions to 2D/3D for visual inspection.
- Dendrograms: For hierarchical clustering.
- Cluster heatmaps: Show relationships between clusters and features.

From <https://copilot.cloud.microsoft/?fromcode=bingchat&redirectid=823852B9497E4A72B8F26930484FC937>