JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Algorithms for Data Science

Unsupervised Learning: Evaluation Metrics

Evaluation Metrics in Unsupervised Learning

Provide insights into algorithm quality based on properties like compactness, separation, and anomaly detection effectiveness.

Internal Metrics

Evaluate clustering quality based on intrinsic data properties.

Detection Metrics

Quantify model performance in identifying anomalies.

Internal Metrics

Measuring clustering quality based on intrinsic properties of the data.

- **Silhouette Score:**

$$\text{Silhouette} = \frac{b - a}{\max(a, b)}$$

  - Where:
    - a: Mean intra-cluster distance.
    - b: Mean nearest-cluster distance.
    - Higher absolute values indicate better clustering.

Internal Metrics

Measuring clustering quality based on intrinsic properties of the data.

- **Davies-Bouldin Index:**

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{j \neq i}\frac{s_i + s_j}{d_{i,j}}$$

  - Where:
    - s_i: Average dispersion of cluster i.
    - $d_{i,j}$,: Distance between cluster centroids i and j.
    - Lower values indicate better clustering.

Internal Metrics

Measuring clustering quality based on intrinsic properties of the data.

- **Inertia:**

$$WCSS = \sum_{i=1}^{N} \min_{j} \left\| x_i - c_j \right\|^2$$

  - Where:
    - x_i: The i-th observation.
    - c_j: The centroid of the j-th cluster.
    - Lower values indicate better clustering but biased by choice of k.

Anomaly Detection Metrics

Quantify performance in identifying anomalies within data.

- **Precision:**

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Minimizing FPs

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Minimizing FNs

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Balances
Precision & Recall

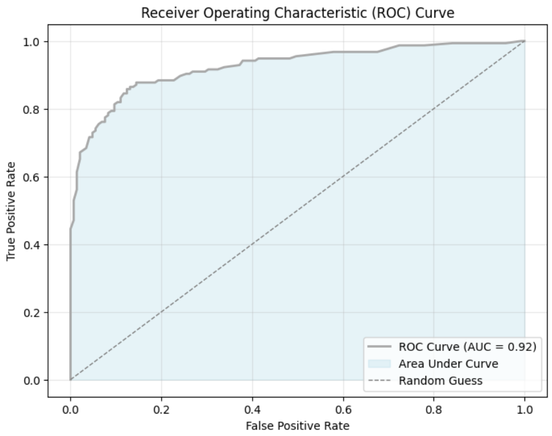
Area Under the ROC Curve (AUC-ROC)

ROC Curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various decision thresholds.

AUC quantifies the overall performance of the model by measuring the area under the ROC curve.

Interpretation:

- $AUC = 1.0$ -> Perfect Model
- $AUC = 0.5$ -> Random Guessing



Key Takeaways

1. **Unsupervised Learning:**
   - ✓ Focused on uncovering hidden patterns in unlabeled data with applications in clustering, anomaly detection, and dimensionality reduction.
2. **Clustering Methods:**
   - ✓ Algorithms like K-Means and DBSCAN group data based on similarity and provide a foundation for anomaly detection and exploratory analysis.
3. **Anomaly Detection:**
   - ✓ Techniques like One-Class SVM identify outliers by learning boundaries that separate normal data from anomalies.
4. **Evaluation Metrics:**
   - ✓ Internal metrics assess clustering quality, while metrics like AUC-ROC and F1-Score evaluate anomaly detection models.

References

[1] Nasir Ahmed, T Natarajan, and K R Rao. "Discrete cosine transform". In: IEEE Transactions on Computers 23.1 (1974), pp. 90–93.

[2] David Arthur and Sergei Vassilvitskii. "k-means++: the advantages of careful seeding". In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (2007), pp. 1027–1035.

[3] Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006. url: https : / / www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf.

[4] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). 1996, pp. 226–231.

[5] Hugging Face. Hugging Face Contribution Guide. Accessed: 2024-10-17. 2024. url: https://huggingface.co/docs/transformers/main/en/contributing.

[6] Hugging Face. Hugging Face Developer Guide. Accessed: 2024-10-17. 2024. url: https://huggingface.co/docs/transformers/main/en/developers.

[7] Gene H Golub and Charles F Van Loan. Matrix Computations. Johns Hopkins University Press, 2013.

[8] Rafael C Gonzalez and Richard E Woods. Digital Image Processing. Prentice Hall, 2008.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn. "Data clustering: A review". In: ACM Computing Surveys 31 (21999), pp. 264–323.

[10] Anil K Jain. "Data clustering: 50 years beyond K-Means". In: Pattern Recognition Letters 31.8 (2010), pp. 651–666.

[11] Ian T Jolliffe. Principal Component Analysis. Springer, 2002.

[12] Stuart P Lloyd. "Least squares quantization in PCM". In: IEEE Transactions on Information Theory 28.2 (1982), pp. 129–137.

[13] J MacQueen. "Some methods for classification and analysis of multivariate observations". In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. University of California Press. 1967, pp. 281–297.

[14] James MacQueen. "Some methods for classification and analysis of multivariate observations". In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. University of California Press. 1967, pp. 281–297.

[15] K R Rao. Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press, 1990.

[16] Bernhard Schölkopf et al. "Estimating the Support of a High-Dimensional Distribution". In: Neural Computation 13.7 (2001), pp. 1443–1471.

[17] Erich Schubert et al. "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN". In: ACM Transactions on Database Systems (TODS) 42.3 (2017), pp. 1–21.

JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING