Personal Notebook: Unsupervised Machine Learning

Comprehensive Study Guide - IBM Machine Learning Course

Generated by AI Research Agent

October 29, 2025

Contents

1 Module 1: Introduction to Unsupervised Learning and K-Means Clustering

1.1 Introduction to Unsupervised Learning

Unsupervised Learning la hinh thuc hoc may ma du lieu khong co nhan, muc tieu la tim cau truc an trong du lieu. Cac loai bao gom:

- Clustering (Phan cum)
- Dimensionality Reduction (Giam chieu du lieu)
- Association Rule Learning
- Anomaly Detection

1.2 K-Means Clustering

Muc tieu: Phan cum du lieu thanh k cum, toi thieu hoa tong binh phuong khoang cach den centroid.

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Cong thuc centroid:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Khoang cach Euclidean:

$$d(x, \mu) = \sqrt{\sum_{j=1}^{n} (x_j - \mu_j)^2}$$

1.2.1 Algorithm Steps

1. Chon so cum k
2. Khoi tao k centroid
3. Gan moi diem du lieu cho centroid gan nhat
4. Cap nhat lai centroid
5. Lap lai den khi hoi tu

1.3 Initialization and Elbow Method

K-Means++ chon centroid dau tien ngau nhien, tiep theo voi xac suat ty le binh phuong khoang cach. Elbow method chon k tai diem uon do thi WCSS.

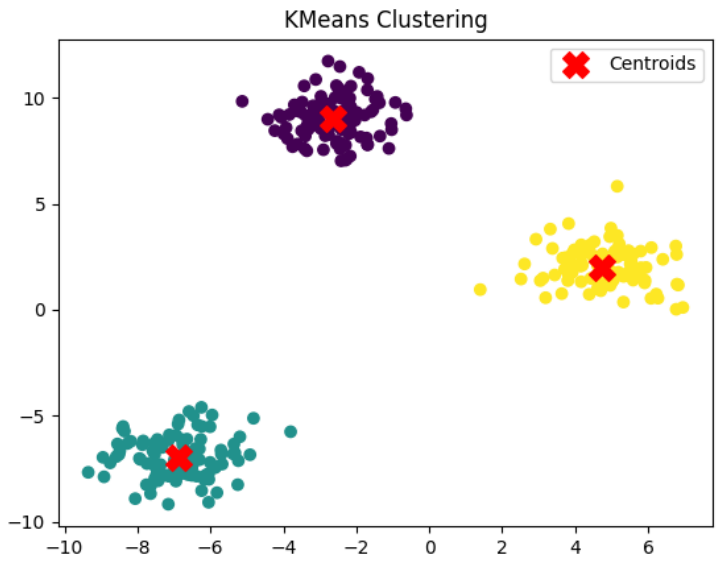

Figure 1: Minh hoa KMeans Clustering

1.4 Silhouette Score

Silhouette score $s(i)$ danh gia chat luong clustering:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$a(i)$ la khoang cach trung binh trong cum, $b(i)$ la khoang cach trung binh toi cum gan nhat.

1.5 Gaussian Mixture Models

Mo hinh hoa du lieu bang hon hop cac Gaussian:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

Thuat toan EM lap giua E-step (tinh responsibility) va M-step (cap nhat tham so).

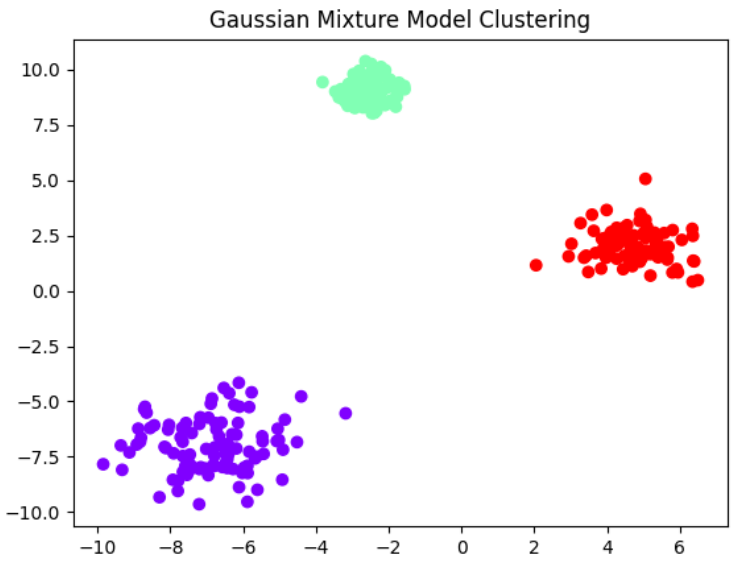

Figure 2: Minh hoa Gaussian Mixture Model

2 Module 2: Distance Metrics and Curse of Dimensionality

2.1 Distance Metrics

Euclidean Distance:

$$d_E(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Manhattan Distance:

$$d_M(p,q) = \sum_{i=1}^{n}|p_i - q_i|$$

Cosine Similarity & Distance:

$$\text{cosine similarity} = \frac{A \cdot B}{||A||\ ||B||}, \quad \text{cosine distance} = 1 - \text{cosine similarity}$$

Jaccard Similarity & Distance:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \quad d_J = 1 - J(A,B)$$

2.2 Curse of Dimensionality

Khong gian da chieu lam khoang cach dong deu va du lieu thua thot, gay kho khan cho thuat toan.

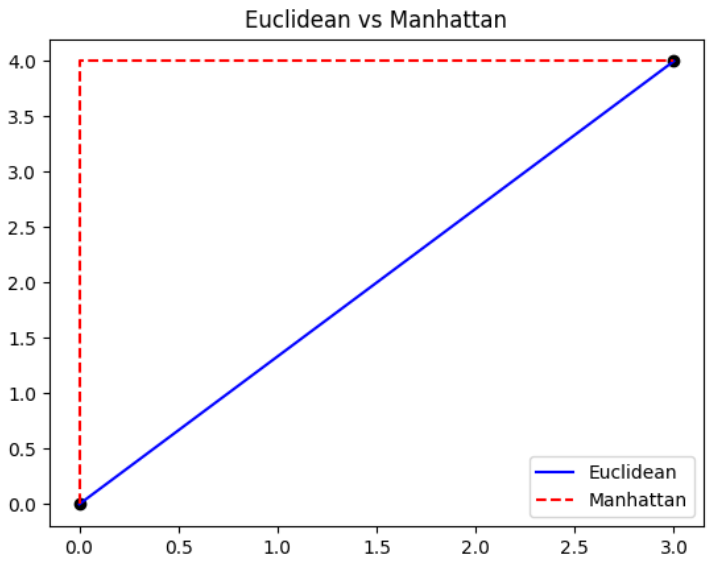

Figure 3: So sanh Euclidean va Manhattan distance

distance_metrics_comparison.png

Figure 4: Cac loai distance metrics pho bien

curse_dimensionality.png

Figure 5: Curse of Dimensionality

3 Module 3: Advanced Clustering Algorithms

3.1 Hierarchical Agglomerative Clustering

Cac phuong phap linkage: Single, Complete, Average, Ward.

3.2 DBSCAN

Chia diem thanh diem loi, bien va nhieu dua tren mat do lan can.

3.3 Mean Shift

Tim mode cua phan phoi mat do kernel.

hierarchical_dendrogram.png

Figure 6: Hierarchical Dendrogram

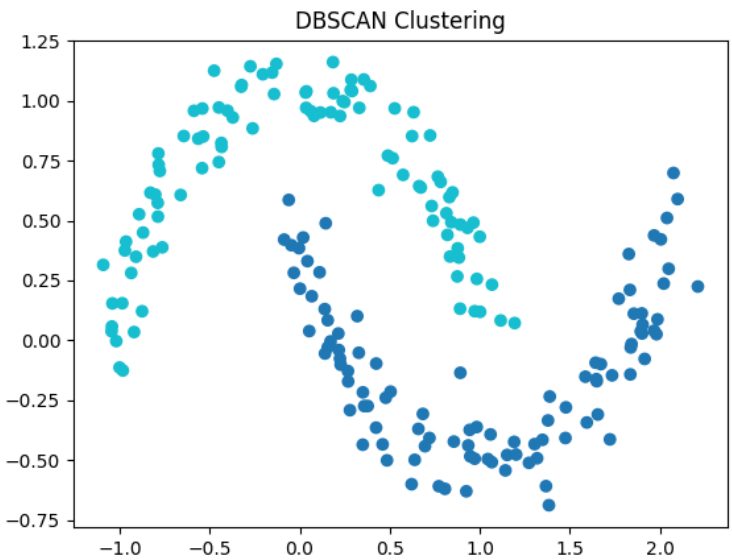

Figure 7: Minh hoa DBSCAN

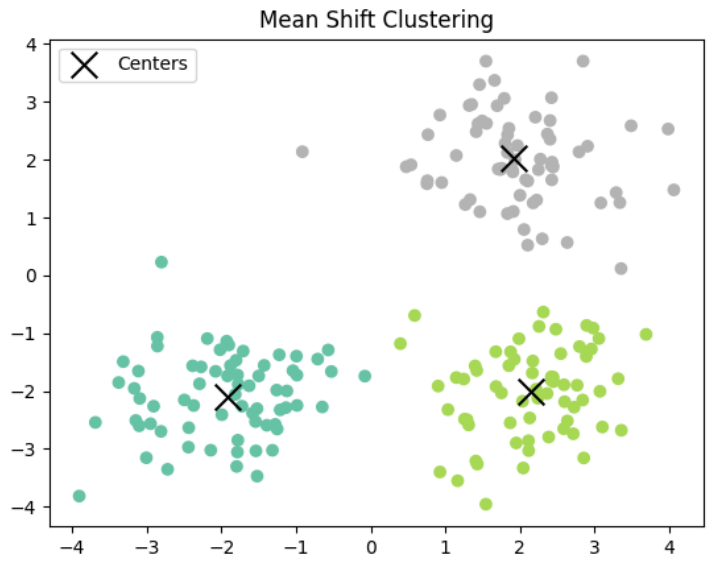

Figure 8: Mean Shift Clustering

3.4 So sanh cac thuat toan clustering

Aspect	K-Means	HAC	DBSCAN	Mean Shift	GMM
Number of Clusters	Need k	Dendrogram	Auto	Auto	Need k
Cluster Shape	Spherical	Depends	Arbitrary	Arbitrary	Elliptical
Scalability	Good	Poor	Medium	Poor	Medium
Noise Handling	Poor	Depends	Good	Good	Moderate
Parameters	k, init	Linkage	ϵ, MinPts	Bandwidth	k, covar
Probabilistic	No	No	No	No	Yes

4 Module 4: Principal Component Analysis (PCA)

PCA tim cac thanh phan chinh orthogonal voi phuong sai lon nhat.

$$C = \frac{1}{n-1} X^T X, \quad Cv = \lambda v, \quad Z = XW$$

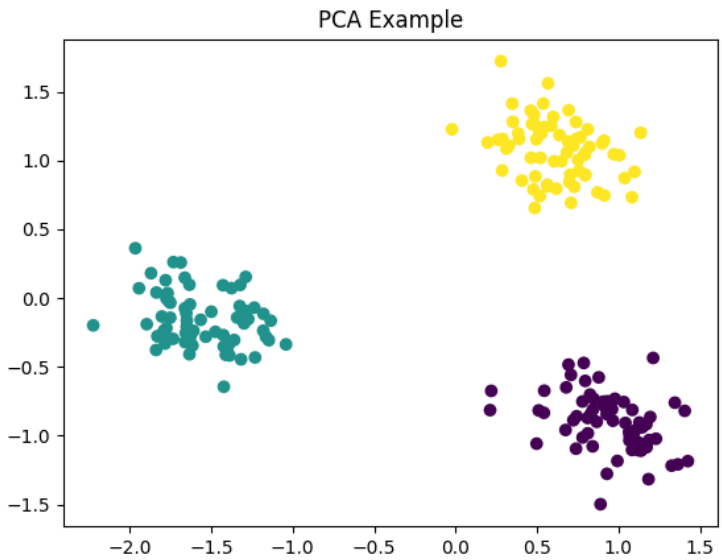

Figure 9: PCA Visualization

5 Module 5: Advanced Dimensionality Reduction Techniques

5.1 Kernel PCA

Chieu du lieu phi tuyen vao space cao chieu bang kernel, lam PCA.

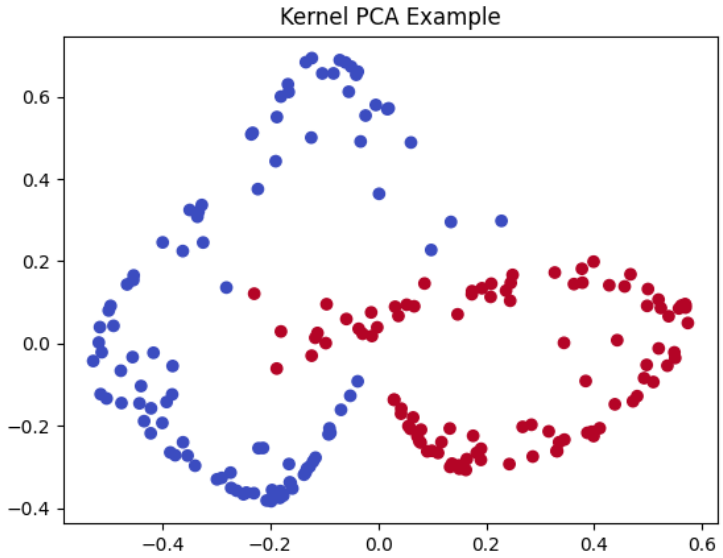

Figure 10: Kernel PCA Visualization

5.2 Multidimensional Scaling (MDS)

Giam chieu dua tren ma tran khoang cach.

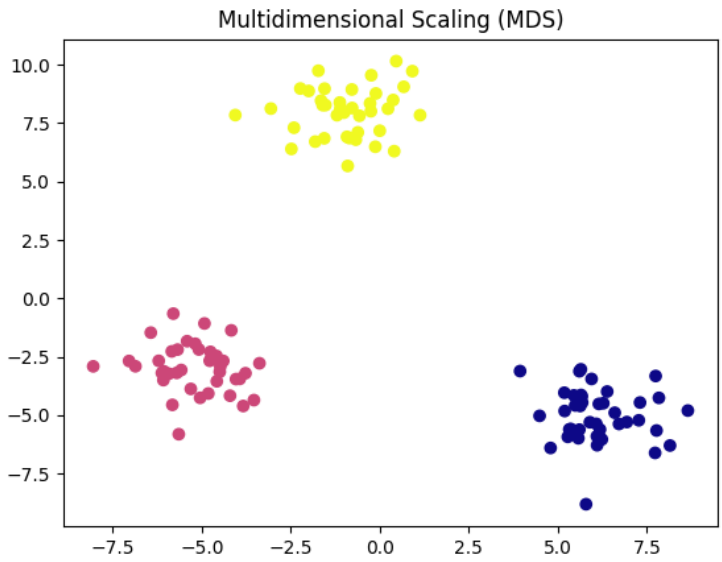

Figure 11: MDS Visualization

6 Module 6: Non-Negative Matrix Factorization (NMF)

Phan tich ma tran khong am V thanh W, H cung khong am:

$$V \approx WH$$

Quy tac cap nhat multiplicative:

$$W \leftarrow W \odot \frac{VH^T}{WHH^T}, \quad H \leftarrow H \odot \frac{W^TV}{W^TWH}$$

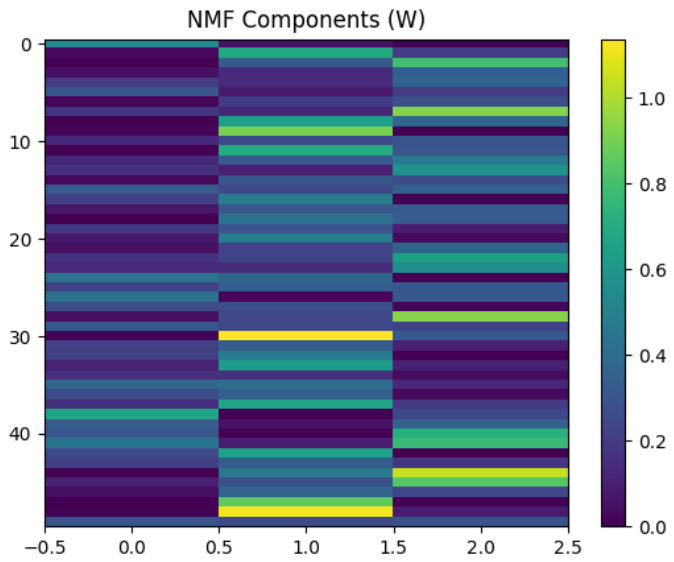

Figure 12: NMF Topic Modeling

7 Module 7: Summary and Best Practices

Tom tat cac ky thuat hoc khong giam sat, cac metric, thuat toan clustering, giam chieu:

- Clustering: K-Means, HAC, DBSCAN, Mean Shift, GMM
- Do khoang cach: Euclidean, Manhattan, Cosine, Jaccard
- Giam chieu: PCA, Kernel PCA, MDS, NMF
- Cac diem quan trong: Curse of dimensionality, chon so cum, danh gia clustering

References to images

- K-Means clustering visualization `kmeans_clustering.png`
- Distance metrics comparison `euclidean_vs_manhattan.png`, `distance_metrics_comparison.pn`
- PCA visualizations `pca_visualization.png`
- Curse of dimensionality `curse_dimensionality.png`
- Cluster comparison plots
- Mean Shift clustering `mean_shift_clustering.png`
- MDS visualization `mds_voting_patterns.png`
- Kernel PCA `kernel_pca.png`
- GMM clustering `gmm_clustering.png`

- Silhouette score
- Covariance matrix