Bab 9: Unsupervised Learning Techniques

Tujuan Bab

Mempelajari teknik-teknik pembelajaran tanpa pengawasan (*unsupervised learning*), yang digunakan ketika data tidak memiliki label. Fokus utama adalah pada clustering, anomaly detection, dan visualisasi data.

Konsep Utama

1. Unsupervised Learning

Berbeda dari supervised learning, unsupervised learning bekerja tanpa label target. Model mencoba menemukan struktur tersembunyi dalam data.

Contoh kasus:

- Segmentasi pelanggan
- Deteksi outlier
- Visualisasi data kompleks

2. Clustering

Clustering membagi dataset ke dalam kelompok (*cluster*) berdasarkan kemiripan.

a. K-Means

Algoritma clustering paling terkenal. Bekerja dengan:

- Menginisialisasi pusat cluster secara acak
- Mengelompokkan data ke pusat terdekat
- Memperbarui pusat berdasarkan rata-rata anggota
- Mengulang proses hingga konvergen

```
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=3)

kmeans.fit(X)
```

Evaluasi dapat dilakukan dengan inertia atau silhouette score.

b. K-Means++ Initialization

Untuk hasil yang lebih baik, digunakan inisialisasi k-means++ agar pusat awal tersebar baik.

c. Mini-Batch K-Means

Versi K-Means yang lebih cepat untuk dataset besar, menggunakan mini-batch daripada seluruh data.

3. Hierarchical Clustering

Pendekatan lain: membangun pohon (*dendrogram*) dari penggabungan bertahap antar data terdekat.

Contoh: Agglomerative Clustering
Mulai dari setiap data sebagai cluster, lalu menggabungkan yang paling dekat sampai jumlah cluster tercapai.

```
from sklearn.cluster import AgglomerativeClustering

agg_clust = AgglomerativeClustering(n_clusters=3)
```

4. DBSCAN

Clustering berbasis kepadatan. Tidak perlu menentukan jumlah cluster. Cocok untuk mendeteksi *outlier* dan bentuk cluster kompleks.

```
from sklearn.cluster import DBSCAN

dbscan = DBSCAN(eps=0.2, min_samples=5)
```

5. Anomaly Detection

Digunakan untuk mendeteksi data yang menyimpang dari pola umum.

a. One-Class SVM

Melatih model hanya pada data "normal", lalu mendeteksi data luar sebagai anomali.

```
from sklearn.svm import OneClassSVM

ocsvm = OneClassSVM(kernel="rbf", nu=0.05, gamma=0.1)
```

b. Isolation Forest

Mendeteksi anomali dengan membangun banyak pohon yang memisahkan data secara acak. Anomali cenderung lebih cepat terisolasi.

```
from sklearn.ensemble import IsolationForest

iso_forest = IsolationForest(contamination=0.1)
```

6. Dimensionality Reduction for Visualization

Salah satu penerapan utama reduksi dimensi adalah untuk memvisualisasikan data high-dimensional.

a. t-SNE (t-distributed Stochastic Neighbor Embedding)

Mampu mengungkapkan struktur lokal dan klaster dalam data. Sangat baik untuk visualisasi, tetapi tidak cocok untuk pipeline supervised learning karena transformasi tidak generalizable.

```
from sklearn.manifold import TSNE

tsne = TSNE(n_components=2)

X_reduced = tsne.fit_transform(X)
```

Proyek / Notebook Praktik

Isi Praktik:

- Clustering dengan K-Means dan evaluasinya menggunakan *silhouette score*
- Visualisasi hasil clustering
- Penggunaan DBSCAN untuk mendeteksi outlier
- Membangun model One-Class SVM dan Isolation Forest untuk anomali

- Visualisasi hasil dengan PCA dan t-SNE

Inti Pelajaran

Konsep	Penjelasan
Clustering	Teknik untuk mengelompokkan data berdasarkan kemiripan
K-Means	Clustering populer dengan pendekatan centroid dan jarak Euclidean
DBSCAN	Clustering berbasis kepadatan, tahan terhadap noise dan bentuk cluster kompleks
Anomaly Detection	Teknik untuk mengenali data yang menyimpang dari distribusi normal
t-SNE	Digunakan untuk memvisualisasikan data kompleks secara 2D atau 3D

Kelebihan dan Kekurangan

Teknik	Kelebihan	Kekurangan
K-Means	Cepat dan sederhana	Perlu menentukan jumlah cluster, tidak cocok untuk bentuk cluster kompleks
DBSCAN	Tidak perlu jumlah cluster, bisa bentuk bebas	Sensitif terhadap parameter
Hierarchical	Tidak perlu jumlah cluster, hasil berupa pohon	Tidak skalabel untuk dataset besar
One-Class SVM	Baik untuk data normal saja	Tidak cocok untuk high-dimensional data
Isolation Forest	Skalabel, cocok untuk anomali	Hasil bisa bervariasi antar run
t-SNE	Visualisasi sangat baik	Tidak bisa digunakan untuk prediksi baru, lambat untuk dataset besar

Bab ini memperkuat pemahaman tentang struktur data dan pentingnya memahami pola yang tersembunyi bahkan tanpa supervisi. Topik ini juga menjadi dasar penting untuk memahami model-model yang mengandalkan representasi data, termasuk dalam NLP dan computer vision.