Chapter 9: Unsupervised Learning Techniques

Pendahuluan

Hingga saat ini, sebagian besar topik yang dibahas adalah *supervised learning*, di mana data pelatihan memiliki label. Chapter ini berfokus pada **Unsupervised Learning** (Pembelajaran Tak Terarah), di mana data pelatihan tidak memiliki label. Tugas utama dalam *unsupervised learning* meliputi:

- **Clustering**: Mengelompokkan data yang mirip ke dalam grup atau *cluster*.
- **Anomaly Detection**: Mendeteksi sampel yang tidak biasa atau *outlier*.
- **Density Estimation**: Mengestimasi distribusi probabilitas dari data.

1. Clustering

Clustering adalah tugas mengelompokkan sekumpulan objek sedemikian rupa sehingga objek dalam grup yang sama (disebut *cluster*) lebih mirip[1] satu sama lain daripada dengan objek di grup lain.

K-Means

K-Means adalah salah satu algoritma clustering yang paling sederhana dan paling banyak digunakan. Algoritma ini bertujuan untuk mempartisi *n* pengamatan ke dalam *k* cluster di mana setiap pengamatan termasuk dalam cluster dengan *mean* (rata-rata) terdekat, yang berfungsi sebagai prototipe dari cluster tersebut. *Mean* ini disebut **centroid**.

Proses Algoritma K-Means:

1. **Inisialisasi**: Pilih *k* centroid secara acak dari dataset.
2. **Penugasan (Assignment)**: Setiap instance data ditugaskan ke centroid terdekatnya.
3. **Pembaruan (Update)**: Posisi setiap centroid diperbarui dengan menghitung rata-rata dari semua instance yang ditugaskan ke centroid tersebut.
4. Langkah 2 dan 3 diulangi hingga centroid tidak lagi berubah secara signifikan.

Tantangan dalam K-Means:

- **Inisialisasi Centroid**: Hasil akhir K-Means sangat bergantung pada penempatan centroid awal. Solusi yang umum adalah **K-Means++**, sebuah strategi inisialisasi cerdas yang menempatkan centroid awal agar saling berjauhan, yang terbukti meningkatkan kualitas hasil akhir.
- **Menemukan Jumlah Cluster (k) yang Optimal**:
    - **Metode Siku (Elbow Method)**: Dengan memplot metrik **inertia** (jumlah kuadrat jarak antara setiap instance dan centroid terdekatnya) sebagai fungsi dari *k*. Grafik yang dihasilkan biasanya terlihat seperti siku. Titik "siku" pada kurva dianggap sebagai indikasi jumlah cluster yang optimal.
    - **Silhouette Score**: Metrik ini mengukur seberapa mirip sebuah objek dengan clusternya sendiri dibandingkan dengan cluster lain. Skornya berkisar dari -1

hingga +1. Skor +1 menunjukkan bahwa instance berada jauh dari cluster tetangga. Skor 0 menunjukkan bahwa instance berada sangat dekat dengan batas keputusan antar cluster. Skor -1 menunjukkan bahwa instance mungkin telah ditempatkan di cluster yang salah. Rata-rata *silhouette score* dari semua instance dapat digunakan untuk mengevaluasi kualitas partisi secara keseluruhan.

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

DBSCAN adalah algoritma clustering berbasis kepadatan. Algoritma ini mendefinisikan cluster sebagai area padat yang dipisahkan oleh area dengan kepadatan rendah.

Konsep Utama DBSCAN:

- **epsilon (ε)**: Jarak yang mendefinisikan "lingkungan" di sekitar setiap instance.
- **min_samples**: Jumlah minimum instance yang harus ada dalam lingkungan ε agar sebuah instance dianggap sebagai *core instance*.

Cara Kerja DBSCAN:

1. Untuk setiap instance, algoritma menghitung jumlah instance lain dalam lingkungan ε-nya.
2. Jika sebuah instance memiliki setidaknya min_samples tetangga, ia ditandai sebagai **core instance**.
3. Semua instance yang berada dalam lingkungan sebuah *core instance* dianggap bagian dari cluster yang sama.
4. Setiap instance yang bukan *core instance* tetapi berada dalam lingkungan *core instance* disebut **border instance**.
5. Instance yang bukan *core* maupun *border* dianggap sebagai **noise** atau *outlier*.

Keunggulan utama DBSCAN adalah kemampuannya untuk menemukan cluster dengan bentuk yang tidak beraturan dan ketahanannya terhadap *outlier*. Ia juga tidak memerlukan penentuan jumlah cluster di awal.

2. Gaussian Mixture Models (GMM)

GMM adalah model probabilistik yang mengasumsikan bahwa semua data dihasilkan dari campuran sejumlah distribusi Gaussian dengan parameter yang tidak diketahui. Setiap cluster pada dasarnya sesuai dengan satu distribusi Gaussian.

Berbeda dengan K-Means yang melakukan *hard clustering* (setiap instance hanya milik satu cluster), GMM melakukan **soft clustering**: ia menghitung probabilitas sebuah instance untuk menjadi anggota dari setiap cluster.

Algoritma Expectation–Maximization (EM)

GMM dilatih menggunakan algoritma **Expectation–Maximization (EM)**, sebuah pendekatan iteratif untuk menemukan estimasi kemungkinan maksimum dari parameter dalam model statistik.

1. **Expectation Step (E-step)**: Langkah ini mengestimasi probabilitas setiap instance untuk dimiliki oleh setiap cluster (penugasan lunak).
2. **Maximization Step (M-step)**: Langkah ini memperbarui parameter dari setiap distribusi Gaussian (rata-rata, kovarians, dan bobot) berdasarkan probabilitas yang dihitung pada E-step.
3. Langkah-langkah ini diulangi hingga parameter model konvergen.

Deteksi Anomali dengan GMM

GMM juga dapat digunakan untuk deteksi anomali. Instance yang terletak di daerah dengan kepadatan probabilitas rendah (jauh dari pusat semua komponen Gaussian) dapat dianggap sebagai anomali.

Seleksi Model

Untuk memilih jumlah cluster *k* yang optimal dan tipe kovarians, kita tidak bisa menggunakan inertia atau silhouette score. Sebaliknya, kita menggunakan kriteria informasi teoretis yang mencoba menyeimbangkan antara seberapa baik model cocok dengan data dan seberapa kompleks model tersebut.

- **Bayesian Information Criterion (BIC)**: $BIC = \log(m)p - 2\log(\hat{L})$
- **Akaike Information Criterion (AIC)**: $AIC = 2p - 2\log(\hat{L})$

Di mana:

- m adalah jumlah instance.
- p adalah jumlah parameter yang dipelajari oleh model.
- $\hat{L}$ adalah nilai maksimum dari fungsi kemungkinan (*likelihood function*) model.

Baik BIC maupun AIC "menghukum" model yang memiliki lebih banyak parameter. Tujuannya adalah untuk memilih model yang paling sederhana namun tetap cocok dengan data. Model dengan nilai BIC atau AIC **terendah** dianggap yang terbaik.