

- a. Apa itu unsupervised learning?
- b. Jelaskan bagaimana cara kerja dari algoritma yang anda implementasikan!
- c. Bandingkan ketiga algoritma tersebut, kemudian tuliskan kelebihan dan kelemahannya!
- d. Jelaskan penerapan dari algoritma unsupervised! (misal di bidang kesehatan atau industri)

Jawab

- a. Unsupervised learning merupakan paradigma dalam machine learning yang tidak membutuhkan label atau “supervisor”. Dengan paradigma ini, model dapat menentukan dan menemukan sendiri pola dan informasi yang ada pada data.

b.

- DBScan

Ada beberapa istilah atau parameter yang digunakan dalam DBScan

1. epsilon, merupakan parameter yang menentukan radius ketetanggaan
2. minimal points, jumlah tetangga minimal yang harus dimiliki sebuah titik data agar menjadi core point
3. core points, sebuah titik data yang memiliki tetangga berjumlah minimal points atau lebih yang berada di dalam radius epsilon
4. directly reachable, sebuah titik q disebut sebagai directly reachable dari titik p jika titik q berada dalam lingkup epsilon dari core point p
5. reachable, titik q disebut reachable dari titik p, jika terdapat jalur dari p_1, \dots, p_n dengan $p_1=p$ dan $p_n=q$, sebagaimana sehingga setiap titik p_{i+1} directly reachable dari p_i
6. outlier atau noise, semua titik yang tidak reachable

Algoritma

1. Cari seluruh titik pada radius epsilon untuk seluruh titik data, dan tentukan core point
2. Menentukan seluruh komponen yang terhubung ke core points dan mengabaikan non-core point
3. Mengklasifikasikan setiap non-core point ke cluster terdekat jika termasuk ke dalam sebuah epsilon core point tertentu

- KMeans

1. Tentukan jumlah cluster, K
2. Pilih K data sebagai centroid awal secara random
3. Untuk seluruh data, hitung jaraknya ke setiap K centroid
4. Klasifikasikan data berdasarkan centroid terdekat
5. Hitung rata-rata dari tiap cluster
6. Nilai yang didapat pada no 5 dijadikan sebagai centroid baru untuk masing-masing cluster

7. Ulangi langkah 3-6 dengan centroid yang baru didapat sebanyak iterasi tertentu atau hingga konvergen

- KMedoids

1. Tentukan jumlah cluster, K
2. Pilih sembarang K data sebagai medoid awal
3. Untuk seluruh data, hitung jaraknya ke setiap K medoid
4. Klasifikasikan data berdasarkan medoid terdekat
5. Hitung cost, yaitu jumlah jarak setiap data ke medoid terdekat
6. Untuk setiap medoid dan data non-medoid
 - Tukar medoid dan non-medoid dan hitung cost
 - Jika cost lebih kecil, maka data non-medoid yang ditukar menjadi medoid baru
 - Jika cost tidak lebih baik, batalkan penukaran

c.

- DBScan

- i. Kelebihan

1. Tidak perlu menentukan banyak cluster yang akan dibuat
2. Tidak dirusak oleh outlier
3. Dapat meminimalisir mengklasifikasikan cluster yang berbeda ke satu cluster yang sama dengan adanya parameter epsilon dan minimal point

- ii. Kekurangan

1. Kualitas dbscan bergantung pada fungsi jarak yang digunakan
2. Tidak berjalan dengan baik untuk data dengan dimensi yang tinggi
3. Kurang cocok untuk data dengan densitas rendah

- KMeans

- i. Kelebihan

1. Implementasi relatif lebih mudah
2. Dijamin akan konvergen
3. Memiliki skalabilitas yang baik untuk dataset yang besar

- ii. Kekurangan

1. Perlu menentukan K secara manual
2. Sensitif terhadap outlier
3. Bermasalah untuk data dengan cluster yang memiliki ukuran dan densitas yang bervariasi

- KMedoids

- i. Kelebihan

1. Relatif lebih efektif terhadap outlier
2. Mudah diimplementasikan
3. Relatif lebih fleksibel

- ii. Kekurangan

1. Perlu menentukan K secara manual
2. Proses pengklasifikasian lebih mahal

3. Skalabilitas yang tidak begitu baik
- d. Unsupervised learning dapat digunakan dalam bidang e-commerce. E-commerce dapat memanfaatkan clustering untuk mengklasifikasikan barang yang dijual, sehingga nantinya pembeli dapat direkomendasikan barang yang berada di cluster yang sama dengan barang yang sedang dicari atau akan dibeli. Dengan begitu, rekomendasi akan lebih efektif sebab lebih tepat sasaran.