Michael Jonathan Halim - 13521124 - GAIB Easy - Unsupervised Learning

1. Jelaskan bagaimana cara kerja dari algoritma yang anda implementasikan!

   Untuk algoritma K Means, cara kerjanya adalah langkah pertama yang dilakukan yaitu meletakkan K centroids pada K random titik/data. Proses tersebut hanya dilakukan satu kali saja pada awal algoritma. Setelah meletakkan K centroids, kita akan membuat clusters berdasarkan centroid yang sudah diletakkan dengan menghitung jarak antara suatu titik data terhadap seluruh centroid dan meng-assign titik data tersebut pada centroid terdekat. Setelah semua titik data sudah diassign dan sudah terbentuk K cluster, maka langkah selanjutnya adalah membuat centroid baru dengan menghitung rata-rata dari seluruh data yang terdapat pada suatu cluster dan meletakkan centroid tersebut pada titik rata-rata data dari cluster tersebut. Proses ini dilakukan untuk menggantikan seluruh centroids yang lama hingga terbentuk K centroids baru. Setelah itu, dihitung sum of squared errors dengan menjumlahkan error dari setiap titik data terhadap centroid. Seluruh proses di atas dilakukan terus menerus (kecuali langkah pertama) hingga iterasi mencapai jumlah iterasi maksimum atau selisih dari sum of squared error yang dihasilkan pada suatu iterasi terhadap iterasi sebelumnya sudah lebih kecil dari nilai toleransi yang ditentukan. Untuk prediksi suatu label terhadap data baru, akan dicari centroid terdekat terhadap data baru tersebut dan meng-assign data tersebut ke cluster terdekat

   Untuk algoritma K Medoids, algoritmanya mirip dengan K Means, bedanya pada penempatan titik pusat cluster tersebut. Untuk K Medoids, tidak digunakan rata-rata data suatu cluster sebagai pusat cluster, namun salah satu titik dari cluster tersebut sebagai pusat cluster. Untuk menentukan medoids baru setelah assign secara random, kita perlu menghitung total dissimilarity dari suatu data terhadap seluruh data lainnya pada cluster tersebut kemudian kita ambil data dengan total dissimilarity terkecil sebagai medoid terbaru dari cluster tersebut. Lalu, kita juga tetap perlu menghitung sum of squared dissimilarity dengan menjumlahkan seluruh dissimilarity setiap data terhadap medoid clusternya. Dan kita hentikan iterasi clustering saat jumlah iterasi mencapai maksimum atau nilai dari selisih antara SSD iterasi saat ini terhadap sebelumnya sudah lebih kecil dari nilai toleransi yang ditentukan. Untuk prediksi, kita cari medoid terdekat dan kita assign data tersebut ke cluster terdekat.

Untuk algoritma DBScan, cara kerjanya adalah dengan mengiterasi seluruh titik pada data, pertama yang dilakukan adalah mencari seluruh tetangga yang memiliki jarak lebih kecil dari epsilon. Apabila jumlah tetangga yang dimiliki lebih kecil dari minimum points yang ditentukan, maka kita anggap titik tersebut sebagai noise (bukan bagian dari suatu cluster). Apabila lebih besar, maka kita akan buat cluster baru dari titik tersebut. Pembuatan clusternya dilakukan dengan berbagai tahap. Pertama, kita tentukan label baru untuk cluster tersebut. Lalu, kita iterasi seluruh tetangga dari titik tersebut. Apabila ada tetangga yang merupakan noise, maka kita assign menjadi bagian dari cluster tersebut. Apabila tetangga tersebut belum tergolong dalam suatu cluster, kita assign menjadi bagian dari cluster tersebut dan kita cari seluruh tetangga dari data baru tersebut. Apabila jumlah tetangga dari data baru tersebut lebih dari minimum points, maka akan kita tambahkan seluruh tetangga tersebut pada iterasi pembentukan cluster saat ini. Jika iterasi algoritma DBScan mencapai titik data yang sudah diassign ke suatu cluster, maka kita skip saja dan lanjut cari data yang belum di-assign ke suatu cluster. Lakukan seluruh proses di atas hingga semua data sudah ditentukan sebagai bagian dari suatu cluster atau noise. Untuk prediksi, kita cari terlebih dahulu seluruh tetangga dari data prediksi tersebut. Jika lebih kecil dari minimum points, maka prediksinya adalah noise. Jika lebih besar, kita cari cluster tetangga yang paling banyak sebagai hasil prediksi.

2. Bandingkan ketiga algoritma tersebut, kemudian tuliskan kelebihan dan kelemahannya!
    a. K Means:
        i. Kelebihan:
            1. Mudah untuk diimplementasikan
            2. Konvergensi lebih cepat
            3. Dapat menangani dataset besar dengan efisien
        ii. Kekurangan:
            1. Pembuatan cluster sangat sensitif terhadap penempatan random centroid pada awal algoritma
            2. Memerlukan penentuan jumlah cluster sebagai masukkan
            3. Tidak menghasilkan hasil clustering yang optimum global
    b. K Medoids:

i. Kelebihan:
    1. Lebih bagus untuk clustering terhadap data dengan outlier
    2. Memilih data paling pusat sebagai titik pusat cluster
ii. Kekurangan:
    1. Komputasi lebih mahal dibandingkan K Means
    2. Tidak cocok untuk dataset yang sangat besar

c. DBScan:
i. Kelebihan:
    1. Deteksi jumlah cluster otomatis, tidak perlu menentukan pada awal
    2. Tahan terhadap noise dan outlier
ii. Kekurangan:
    1. Sulit untuk menemukan hyperparameter epsilon dan minimum points yang cocok terhadap data
    2. Komputasi lebih mahal untuk data yang dimensinya tinggi

3. Jelaskan penerapan dari algoritma unsupervised di dunia nyata!

Algoritma clustering dapat digunakan seperti anomaly detection, pengelompokkan pelanggan, sistem rekomendasi, dan segmentasi gambar. Anomaly detection sering kali menggunakan data yang tidak berlabel sehingga algoritma machine learning yang digunakan biasanya adalah clustering untuk menentukan data yang anomaly (outlier). Data-data seperti perilaku pelanggan dapat digunakan untuk mengelompokkan pelanggan berdasarkan perilaku atau preferensi pembeliannya. Sistem rekomendasi juga bisa menggunakan clustering. Contohnya seperti aplikasi spotify yang dapat merekomendasikan lagu kepada pelanggannya sesuai dengan data lagu yang sering digunakan seperti tempo, loudness, energy, dan lainnya. Segmentasi gambar juga sering menggunakan algoritma clustering untuk membagi gambar menjadi beberapa bagian berdasarkan kesamaan fitur seperti warna atau tekstur.