Metode Clustering: K-means

Pengantar Dalam artikel sebelumnya, telah dibahas perbedaan clustering dan classification. Dalam artikel ini, kita akan membahas salah satu metode clustering yang paling dasar, yaitu K-means Clustering. Apa itu K-means? Sebelum kita melangkah lebih jauh, mungkin ada baiknya mengetahui latar belakang mengapa disebut K-means. K di sini dimaksudkan sebagai konstanta jumlah cluster yang diinginkan. Jadi, berhubung kita sudah mengasumsikan jumlah cluster yang akan dihasilkan algoritma ini, maka K didefiniskan diawal (contoh: K = 5 cluster). Means dalam hal ini berarti nilai suatu rata-rata dari suatu grup data yang dalam hal ini didefinisikan sebagai cluster. Jika kita menggabungkan kedua hal tersebut, maka dapat diartikan bahwa algoritma ini menggunakan K nilai rata-rata yang setiap nilai rata-ratanya dihitung dari suatu cluster. Kalau ada 5 cluster, maka akan ada 5 rata-rata yang dipakai oleh algoritma ini. Model Matematik Seperti yang kita tau bahwa metode K-means ini menggunakan nilai rata-rata yang diambil dari setiap cluster. Maka berikut adalah cara bagaimana K-means menghitung rata-rata dari setiap cluster Ck adalah nilai rata-rata dari cluster K (contoh: C1 adalah nilai rata-rata dari cluster yang pertama). adalah semua anggota dari cluster K. Pertanyaaan berikutnya adalah, bagaimana cara memilih anggota dari suatu cluster? Cara memilihnya mudah. Andaikan ada suatu data, kita ingin mengetahui ke dalam anggota cluster manakah data tersebut paling cocok dimasukkan. Caranya adalah dengan menghitung selisih antara data dan setiap nilai rata-rata cluster. Cluster yang nilai rata-ratanya yang memiliki selisih terkecil dengan data tersebut merupakan cluster dimana data tersebut dikategorisasikan. Secara matematis dapat didefinisikan sebagai berikut. X adalah data yang sedang kita tentukan ke cluster mana harus dimasukkan. Ck adalah nilai rata-rata dari cluster k. K adalah jumlah cluster. Jadi, Cluster T merupakan cluster yang paling cocok untuk data X, karena cluster T memiliki selisih terkecil. Bagaimana cara menghitung selisih? Kita bisa menggunakan berbagai macam metode seperti Eucledian distance, Mahalanobis distance, Manhattan distance, Normalised Cosines distance. Metode yang paling populer adalah dengan menggunakan Eucledian distance. Kita sudah mengetahui bagaimana algoritma menghitung mean dari masing-masing cluster, dan bagaimana algoritma mengelompokan data ke dalam cluster-cluster yang ada. Pertanyaan berikutnya adalah, ketika pertama kali algoritma dijalankan, kita hanya punyai adalah jumlah cluster yang akan dihasilkan (K). Nah! untuk menghitung nilai rata-rata dari setiap cluster diperlukan anggota, dan untuk menentukan anggota, kita memakai informasi nilai rata-rata. Ini mirip dengan masalah ayam dan telur, yang mana yang lebih dahulu. Untuk mengatasinya, kita akan menentukan terlebih dahulu nilai rata-rata dari setiap cluster. Bagaimana caranya? Ada banyak cara, untuk artikel singkat ini, kita akan menentukannya secara acak (random). Tentunya dalam menentukan angka acak kita tidak sembarangan sebab ini akan membuat algoritma tidak berjalan dengan baik. Cara yang paling mudah adalah memilih data yang ada sebagai nilai rata-rata dari suatu cluster. Sebagai contoh, misalkan ada 3 data yaitu (1,0), (1,2), (1,4). K kita tentukan K = 2. Jadi ada dua nilai rata-rata yang perlu kita tentukan diawal. Secara acak kita memilih C1 = (1,0) dan C2 = (1,4). Setelah kita menentukan nilai rata-rata awal dari setiap cluster, selanjutnya algoritma akan meng-update keanggotaan dari setiap cluster. Setelah itu algoritma akan menghitung kembali nilai rata-rata dari setiap cluster berdasarkan anggotanya yang baru saja di-update. Pertanyaan berikutnya, kapan berhenti? Algoritma akan berhenti ketika tidak ada perubahan keanggotaan dari setiap cluster. Algoritma Baik, sekarang kita akan melihat algoritma K-means clustering ini.

- Tentukan K
- Tentukan Ck untuk semua cluster dengan memilih secara acak (random) dari data yang ada
- Update keanggotaan dari setiap cluster
- Update Ck berdasarkan anggota yang baru saja di-update
- Lakukan langkah 3 dan 4, sampai tidak ada perubahan keanggotaan dari setiap cluster. Mudah bukan?

Mungkin ada beberapa dari anda bertanya mengapa algoritma ini benar? Dalam konteks pertanyaan ini, yang ditanyakan adalah apakah algoritma ini selalu konvergen atau dalam arti, apakah algoritma ini dapat selalu berhenti, atau akan berjalan terus. Jawaban nya adalah hampir dapat dipastikan algoritma ini dapat konvergen atau dapat dapat berhenti, dengan beberapa catatan. Untuk para pembaca yang tertarik untuk mengetahui lebih jauh sifat konvergensi dari K-means, dapat meneliti lebih jauh dengan mencari literature yang berhubungan dengan konvergensi K-means. Apakah hasil dari setiap eksekusi algoritma pada data yang sama adalah selalu sama? Jawabannya tidak. Karena ada unsur random disini, maka ada kemungkinan algoritma akan menghasilkan hasil yang berbeda-beda. Untuk memastikannya, biasanya algoritma dieksekusi berulang kali, dan hasilnya dianalisa. Penutup Kita sudah melihat bagaimana algoritma K-means dan model matematiknya. Metode K-means adalah metode yang paling dasar dan paling populer dalam clustering. Namun demikian, banyak sekali kekurangan dalam metode ini. Metode ini biasanya dapat berjalan dengan baik ketika data yang sedang diproses mempunyai model Gaussian. Kalau kita tidak tau model dari data yang kita punya, tidak usah takut karena biasanya dalam clustering kita bisa mencoba-coba terus menerus baik itu mengubah nilai K, maupun mengeksekusi algoritma berulang kali sampai kita puas dengan hasilnya.