Nama : Tassya Ramadhanti
NIM : 1103204016

Judul : Analisis Komponen Utama (PCA)

PCA adalah metode untuk mengompresi banyak data menjadi sesuatu yang menangkap esensi dari data asli.

Pengenalan Dimensi

- Dimensi 1 (1-D) = garis angka, contoh (Anggaplah satu set data RNA-seq untuk satu sel). Ada dua output (Distribusi seragam dan Distribusi tidak seragam) (jika kita memiliki satu data)
- 2-D = grafik normal, dalam grafik dua dimensi kita memiliki dua sumbu. Kita mungkin melihat dua ekspresi di dua sel yang berkorelasi dan tidak berkorelasi. Jika ekspresi berkorelasi, ini berarti gen yang ditranskripsi tinggi di sel satu juga ditranskripsi tinggi di sel dua dan gen yang ditranskripsi rendah di sel satu juga ditranskripsi rendah di sel dua dan jika ekspresi tidak berkorelasi, ini berarti gen yang ditranskripsi tinggi di sel satu tidak memberi tahu kita apa-apa tentang apakah gen tersebut ditranskripsi tinggi atau rendah di sel dua. (jika kita memiliki data dari dua sel)
- 3-D = grafik mewah yang memiliki kedalaman = memiliki tiga sumbu terpisah (jika kita memiliki data dari tiga sel)

PCA mengambil dataset dengan banyak dimensi dan meratakannya menjadi 2 atau 3 dimensi sehingga kita dapat melihatnya. PC1 adalah arah variasi terbesar dalam ekspresi gen, PC 2 adalah arah variasi paling besar ke-2 dalam ekspresi gen.

- Jika kita memiliki 2 sel = PC1 menangkap arah di mana sebagian besar variasi berada, PC2 menangkap arah dengan variasi paling besar ke-2.
- Jika kita memiliki 3 sel = PC1 mencakup arah variasi paling besar, PC2 mencakup arah variasi paling besar ke-2, PC3 mencakup arah variasi paling besar ke-3
- Jika kita memiliki 4 sel = PC1 mencakup arah variasi paling besar, PC2 mencakup arah variasi paling besar ke-2, PC3 mencakup arah variasi paling besar ke-3, PC4 = mencakup arah variasi paling besar ke-4

BAGAIMANA KITA MENEPATKAN SEL? Panjang dan arah PC1 sebagian besar ditentukan oleh gen yang dilingkari. Gen dengan sedikit pengaruh pada PC1 mendapatkan nilai mendekati nol, dan gen dengan lebih banyak pengaruh mendapatkan angka yang lebih jauh dari nol.

Untuk mengidentifikasi gen kunci, kita ingin mengetahui gen mana yang memiliki pengaruh besar dalam menempatkan sel dermal di kiri sel saraf di kanan, kita dapat melihat skor pengaruh di PC1 dan jika kita ingin menemukan gen mana yang membantu membedakan sel darah dari sel saraf dan sel dermal, kita dapat melihat skor pengaruh di PC2.

Judul : K-nearest neighbors (KNN)

- K-nearest neighbors (KNN) adalah metode pembelajaran mesin untuk mengklasifikasikan data. KNN bekerja dengan mencari data yang paling mirip dengan data baru, dan kemudian mengategorikan data baru tersebut ke dalam kategori yang sama dengan data yang paling mirip.
  Berikut langkah-langkahnya:

Langkah 1: Mulailah dengan kumpulan data dengan kategori yang sudah diketahui
Langkah 2: Tambahkan sel baru, dengan kategori yang tidak diketahui, ke plot PCA
Langkah 3: Kita mengklasifikasikan sel baru dengan melihat sel yang telah diberi label yang paling dekat

- Terminologi Pembelajaran Mesin/Penambangan Data = Data yang digunakan untuk pengelompokan awal (data di mana kita sudah tahu kategorinya sebelumnya) disebut "data pelatihan". Data pelatihan adalah data yang digunakan untuk menghitung jarak antara data baru dan data yang sudah diketahui. Data pelatihan harus mencakup data dari semua kategori yang akan digunakan untuk mengklasifikasikan data baru.
- Nilai K adalah jumlah data yang digunakan untuk menentukan kategori data baru. Nilai K yang lebih besar akan menghasilkan klasifikasi yang lebih halus, tetapi juga lebih rentan terhadap noise. Nilai K yang lebih kecil akan menghasilkan klasifikasi yang lebih kasar, tetapi juga lebih tahan terhadap noise.

Judul: Pohon Keputusan dan Klasifikasi

Pohon keputusan adalah algoritma pembelajaran mesin yang dapat digunakan untuk tugas klasifikasi dan regresi. Pohon keputusan bekerja dengan cara membagi data secara rekursif menjadi subset yang lebih kecil dan lebih kecil, hingga setiap subset hanya berisi titik data dari kelas atau kategori yang sama.

Perbedaan antara pohon keputusan dan pohon regresi

- Tujuan: Pohon keputusan digunakan untuk tugas klasifikasi, sedangkan pohon regresi digunakan untuk tugas regresi.
- Keluaran: Pohon keputusan menghasilkan kelas, sedangkan pohon regresi menghasilkan nilai numerik.
- Terminologi: Pohon keputusan menggunakan terminologi seperti "simpul daun" dan "pengotor jin", sedangkan pohon regresi menggunakan terminologi seperti "simpul daun" dan "error kuadrat rata-rata".

Terminologi yang umum digunakan dalam pohon keputusan sebagai berikut:

- Simpul akar: Simpul paling atas dari pohon.
- Simpul internal: Simpul yang memiliki lebih dari satu anak.
- Simpul daun: Simpul yang tidak memiliki anak.
- Fitur: Atribut data yang digunakan untuk membuat keputusan.
- Ambang batas: Nilai yang digunakan untuk membagi data menjadi dua subset.
- Pengotor jin: Ukuran ketidakmurnian dari suatu simpul.