Nama : Satria Nur Hidayatullah
NIM : 1103130035
Kelas : MachineLearning-Gab-01

Section 1 : Theory

1. Machine learning adalah sebuah metode dalam pembuatan suatu task dalam komputer, yang bertujuan membuat mesin menirukan manusia dalam hal pembelajaran. Sehingga pada tujuan akhirnya terbentuk suatu program yang dapat memahami permasalahan yang diberikan serta memberikan solusi yang diinginkan tanpa harus mendapat perintah dari manusia.

2. Contohnya adalah suatu program yang dapat mengenali gambar asli atau gambar hasil editan yang biasa disebut dengan *image spoofing*

3. (a) untuk perhitungan cosine hanya digunakan range [0 - 1] sedangkan cosine correlation memiliki nilai rentang [-1 - 1

   (b) Tidak tentu, cosine maksimal dengan nilai 1 menunjukkan bahwa 2 objek tersebut adalah objek yang "similar" apabila digambarkan dengan vektor arah, maka 2 objek yang memiliki nilai cosine sama dengan 1 adalah 2 objek yang memiliki arah yang sama. Arah yang sama artinya adalah atribut yang dimiliki dibedakan dengan nilai yang konstan.

   (c) Berdasarkan fungsi:

$$CosSim(x,y) = \frac{\sum_i^n x_i y_i}{\sqrt{\sum_i^n x_i^2}\sqrt{\sum_i^n y_i^2}} \tag{1}$$

$$Corr(x,y) = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2}\sqrt{\sum_i^n (y_i - \bar{y})^2}} \tag{2}$$

   sehingga diperoleh fungsi:

$$Corr(x,y) = CosSim(x - \bar{x}, y - \bar{y}) \tag{3}$$

   sehingga dapat dibuktikan dari fungsi (3) bahwa correlation didapatkan dari hasil cosine similarity yang telah di kurangi oleh rata-rata. Sehingga correlation adalah fungsi yang digunakan untuk mencari cosine similarity dengan mengambil nilai asli dari vector itu sendiri, karena pengurangan terhadap rata-rata digunakan untuk mencari nilai asli dari suatu data. Dan apabila nilai rata rata dari x dan y = 0, maka $corr(x,y) = cos(x,y)$

   (d) misal x dan y adalah 2 vektor yang memiliki panjan $L_2 = 1$. Untuk kasus vektor seperti ini, nilai variansi dari ke-2 vektor adalah n kali jumlah akar dari nilai atribut dan nilai korelasi nya adalah hasil dari

dot product kedua vektor dibagi dengan n.

$$\begin{aligned} d(x,y) &= \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \\ &= \sqrt{\sum_{k=1}^{n} x_k^2 - 2x_k y_k + y_k^2} \\ &= \sqrt{1 - 2\cos(x,y) + 1} \\ &= \sqrt{2(1 - \cos(x,y))} \end{aligned}$$

(e) misal x dan y adalah 2 vektor yang memiliki nilai mean 0 dan standar deviasi 1. Untuk kasus vektor seperti ini, nilai variansi dari ke-2 vektor adalah n kali jumlah akar dari nilai atribut dan nilai korelasi nya adalah hasil dari dot product kedua vektor dibagi dengan n.

$$\begin{aligned} d(x,y) &= \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \\ &= \sqrt{\sum_{k=1}^{n} x_k^2 - 2x_k y_k + y_k^2} \\ &= \sqrt{n - 2ncorr(x,y) + n} \\ &= \sqrt{2n(1 - corr(x,y))} \end{aligned}$$

4. (a) Yang pertama dengan cara mengitung jarak dengan suatu centroid, dengan menggunakan nilai euclidian distance, seperti saat melakukan clustering. Yang kedua adalah dengan menggunakan kedekatan antar 2 data. Salah satunya dengan menggunakan nilai minimum similarity, atau maksimum similarity.

   (b) Salah satu caranya adalah menghitung jarak antara centroid dengan 2 titik nilai.

   (c) Dengan menghitung jarak dari kedua nilai tersebut dengan 1 titik centorid. Atau dengan cara mengambil langsung nilai kedekatan minimum atau maksimum.

5. (a) Hasil log pada perhitungan invers menjadi n jumlah dokumen, sehingga memperoleh nilai maksimal dari frekuensi dokumen invers.

   (b) Hasil log pada perhitungan invers menjadi 0, sehingga memiliki nilai minimum frekuensi dokumen invers.

   (c) Hasil invers menunjukkan bahwa kata yang muncul di semua dokumen tidak bisa membedakan ciri-ciri antar dokumen, sedangkan kata yang hanya di sedikit dokumen, bisa membedakan antar dokumen.

(d) Salah satu penggunaannya adalah ketika digunakan pada perhitungan suatu film yang disukai oleh beberapa orang dengan kesenangan genre tersendiri. Bisa digunakan untuk film x yang sangat disukai oleh orang-orang dengan kesenangan genre x ketika dia hanya disukai oleh sekelompok orang.

Reference: Hampir semua jawaban saya dapatkan rata-rata dari mencoba memahami buku Introduction to Data Mining karya Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Namun ada beberapa yang kata-katanya hampir mirip karena tidak bisa memahami. Dan buku ini saya dapatkan dengan cara googling soal secara langsung pada saat menyerah.hehe