UTS I

Penambangan Data dalam Sains

Mohammad Rizka Fadhli - 20921004

31 March 2022

SOAL 1 a

Andaikan anda bekerja dalam sebuah lembaga riset yang fokusnya dalam bidang Sains Fisika atau Kimia. Jelaskan bagaimana data mining (Klasifikasi, Clustering dan Asosiasi) dapat membantu anda dalam riset di bidang Fisika atau Kimia? Pilih bidang riset sesuai keahlian anda!

Jawab

Pada bidang saya di *marketing riset*, aplikasi data mining sangat banyak sekali. Sebagai contoh:

1. Klasifikasi dapat digunakan untuk memprediksi apakah seorang pelanggan akan membeli produk kita atau tidak berdasarkan data prediktor yang ada.
2. *Clustering* dapat digunakan untuk mengelompokkan pelanggan berdasarkan beberapa atribut yang ada.
3. Asosiasi dapat digunakan untuk membuat analisa *consumer basket*, yakni mengelompokkan barang-barang apa yang dibeli secara bersamaan oleh pelanggan. Bisa juga dijadikan acuan untuk membuat produk *bundling* atau kumpulan produk yang dijual bersamaan oleh *retailer*.

SOAL 1 b

Jelaskan tentang berbagai jenis atribut dalam Data Mining dan metode normalisasi berbagai jenis atribut tersebut?

Jawab

Beberapa jenis atribut antara lain:

1. Tipe nominal atau kategorik, yakni berupa representasi dari pengamatan. Misalkan *gender*, warna, nama hari, dll.
2. *Binary*, yakni berupa pilihan biner [0,1].
3. Numerik, yakni berupa angka. Bisa dalam diskrit (bilangan bulat) atau bilangan interval / ratio atau bilangan *real*.
4. *Ordinal*, yakni data berupa rentang yang memiliki urutan tertentu.
5. *String* atau *character*, yakni berupa teks saja.

Beberapa metode normalisasi antara lain:

Data Numerik

Min-max normalization

$$v' = \frac{v - min}{max - min}$$

Z-score normalization

$$v' = \frac{v - \mu}{\sigma}$$

Data Ordinal

$$z = \frac{v - 1}{M - 1}$$

dimana v diurutkan dan dijadikan angka dan M adalah elemen dari v yang terbesar.

SOAL 1 c

Diberikan data:

id	A	B	C	class
1	0	2	1	Y
2	1	2	2	Y
3	1	1	0	Y
4	2	0	2	Y
5	1	1	1	Y
6	2	2	2	N
7	0	1	0	N
8	0	0	1	N
9	2	1	0	N
10	1	0	0	N

Hitung *dissimilarity* untuk:

- Objek 1 dan 6
- Objek 3 dan 8

Jawab

Untuk memudahkan, kita akan ubah atribut `class` menjadi *binary* sebagai berikut:

id	A	B	C	class_Y	class_N
1	0	2	1	1	0
2	1	2	2	1	0
3	1	1	0	1	0
4	2	0	2	1	0
5	1	1	1	1	0
6	2	2	2	0	1
7	0	1	0	0	1
8	0	0	1	0	1
9	2	1	0	0	1
10	1	0	0	0	1

Euclidean Distance didefinisikan sebagai:

$$d(X, Y) = \sqrt{\sum (x_i - y_i)^2}$$

Minkowski Distance didefinisikan sebagai:

$$d(X, Y) = (\sum |x_i - y_i|^p)^{\frac{1}{p}}$$

Untuk soal ini, saya akan gunakan $p = 1$.

Jarak objek 1 dan 6 adalah:

Objek 1 = 0,2,1,1,0

Objek 6 = 2,2,2,0,1

Euclidean Distance

$$d(id_1, id_6) = \sqrt{(0-2)^2 + (2-2)^2 + (1-2)^2 + (1-0)^2 + (0-1)^2} = \sqrt{7} \approx 2.646$$

Minskowski Distance saat $p = 1$

$$d(id_1, id_6) = |0-2| + |2-2| + |1-2| + |1-0| + |0-1| = 5$$

Jarak objek 3 dan 8 adalah:

Objek 3 = 1,1,0,1,0

Objek 8 = 0,0,1,0,1

Euclidean Distance

$$d(id_3, id_8) = \sqrt{(1-0)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2} = \sqrt{5} \approx 2.236$$

Minskowski Distance saat $p = 1$

$$d(id_3, id_8) = |1-0| + |1-0| + |0-1| + |1-0| + |0-1| = 5$$

SOAL 2 a

Mengapa kita perlu melakukan permrosesan awal data?

Jawab

Data yang pertama kali kita terima belum tentu siap untuk dianalisa. Beberapa analisa tidak memperbolehkan data kosong dan data numerik yang terlalu lebar *range*-nya. Oleh karena itu perlu ada tahapan *pre-processing* terlebih dahulu. Hal yang biasa dilakukan antara lain:

1. Melihat konsistensi format penulisan data.
2. Melakukan normalisasi untuk data numerik.
3. Melihat adanya data yang kosong atau bolong.
4. Melihat keberadaan nilai pencilan pada data numerik.

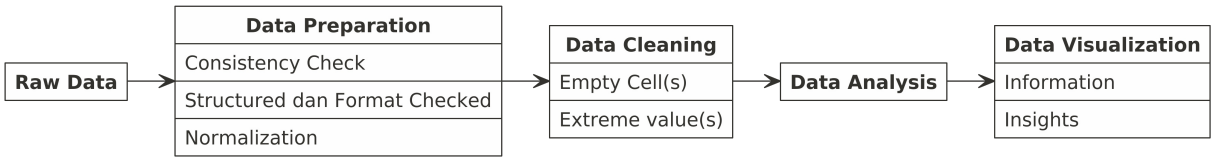

Figure 1: Tahapan Data Processing

SOAL 2 b

Diberikan data sebagai berikut:

id	body_temp	birth	four_leg	hibernate	class
1	warm	Y	Y	Y	Y
2	warm	Y	Y	N	Y
3	warm	Y	N	Y	N
4	warm	Y	N	N	N
5	cold	N	Y	Y	N
6	cold	N	Y	N	N
7	cold	N	N	Y	N
8	cold	N	N	N	N
9	warm	N	N	N	N
10	cold	Y	N	N	N

Di bawah ini diberikan 10 buah data dengan masing-masing 6 atribut. Buat algoritma reduksi dimensi menggunakan PCA jika kita ingin mereduksi dimensi data di atas. Bagaimana kita memilih atribut yang berisi informasi 90% dari data di atas?

Jawab

Untuk membuat PCA-nya, data yang ada perlu kita ubah menjadi *binary* terlebih dahulu. Dari bentuk berikut ini:

body_temp	birth	four_leg	hibernate
warm	Y	Y	Y
warm	Y	Y	N
warm	Y	N	Y
warm	Y	N	N
cold	N	Y	Y
cold	N	Y	N
cold	N	N	Y
cold	N	N	N
warm	N	N	N
cold	Y	N	N

menjadi berikut ini:

body_tempcold	body_tempwarm	birthN	birthY	four_legN	four_legY	hibernateN	hibernateY
0	1	0	1	0	1	0	1

body_tempcold	body_tempwarm	birthN	birthY	four_legN	four_legY	hibernateN	hibernateY
0	1	0	1	0	1	1	0
0	1	0	1	1	0	0	1
0	1	0	1	1	0	1	0
1	0	1	0	0	1	0	1
1	0	1	0	0	1	1	0
1	0	1	0	1	0	0	1
1	0	1	0	1	0	1	0
0	1	1	0	1	0	1	0
1	0	0	1	1	0	1	0

Untuk menghitung PCA, berikut algoritmanya:

```
STEP 1
  Menghitung covariance matrix dari data
    misal C
STEP 2
  Mencari eigen values dan eigen vectors dari C
STEP 3
  Eigen value merepresentasikan varians yang bisa di-"explained" oleh PC
  Penjumlahan k eigen values pertama adalah varians explained dari k-dimensi
```

Berikut adalah matriks kovariansi dari data yang ada:

```
##                body_tempcold body_tempwarm     birthN     birthY   four_legN
## body_tempcold     0.2777778    -0.2777778  0.1666667 -0.1666667  0.00000000
## body_tempwarm    -0.2777778     0.2777778 -0.1666667  0.1666667  0.00000000
## birthN            0.1666667    -0.1666667  0.2777778 -0.2777778  0.00000000
## birthY           -0.1666667     0.1666667 -0.2777778  0.2777778  0.00000000
## four_legN         0.0000000     0.0000000  0.0000000  0.0000000  0.26666667
## four_legY         0.0000000     0.0000000  0.0000000  0.0000000 -0.26666667
## hibernateN        0.0000000     0.0000000  0.0000000  0.0000000  0.04444444
## hibernateY        0.0000000     0.0000000  0.0000000  0.0000000 -0.04444444
##                 four_legY  hibernateN  hibernateY
## body_tempcold  0.00000000  0.00000000  0.00000000
## body_tempwarm  0.00000000  0.00000000  0.00000000
## birthN         0.00000000  0.00000000  0.00000000
## birthY         0.00000000  0.00000000  0.00000000
## four_legN     -0.26666667  0.04444444 -0.04444444
## four_legY      0.26666667 -0.04444444  0.04444444
## hibernateN    -0.04444444  0.26666667 -0.26666667
## hibernateY     0.04444444 -0.26666667  0.26666667
```

Berikut adalah nilai eigen:

```
## [1] 0.889 0.622 0.444 0.222 0.000 0.000 0.000 0.000
```

dan vektor eigennya:

```
##      [,1] [,2] [,3] [,4]         [,5]          [,6]          [,7]          [,8]
## [1,]  0.5  0.0  0.0  0.5 0.000000e+00  0.000000e+00 -7.071068e-01  0.000000e+00
## [2,] -0.5  0.0  0.0 -0.5 0.000000e+00 -4.891900e-17 -7.071068e-01  0.000000e+00
## [3,]  0.5  0.0  0.0 -0.5 0.000000e+00  7.071068e-01  5.551115e-17  0.000000e+00
## [4,] -0.5  0.0  0.0  0.5 0.000000e+00  7.071068e-01 -5.551115e-17  0.000000e+00
## [5,]  0.0  0.5  0.5  0.0 0.000000e+00  0.000000e+00  0.000000e+00 -7.071068e-01
## [6,]  0.0 -0.5 -0.5  0.0 1.369272e-17  0.000000e+00  0.000000e+00 -7.071068e-01
## [7,]  0.0  0.5 -0.5  0.0 7.071068e-01  0.000000e+00  0.000000e+00  2.359224e-16
## [8,]  0.0 -0.5  0.5  0.0 7.071068e-01  0.000000e+00  0.000000e+00 -2.498002e-16
```

Untuk memilih atribut yang berisi 90% informasi di atas, kita cukup mengambil **2** dimensi karena penjumlahan dua vektor eigen pertama sudah $> 90\%$.

SOAL 3

Diberikan data sebagai berikut:

id	A	B	C	class
1	0	2	1	Y
2	1	2	2	Y
3	1	1	0	Y
4	2	0	2	Y
5	1	1	1	Y
6	2	2	2	N
7	0	1	0	N
8	0	0	1	N
9	2	1	0	N
10	1	0	0	N

Buat klasifikasi dengan *Naive Bayes*!

Jawab

Naive Bayes Classifier berlandaskan peluang bersyarat. Berikut adalah langkah perhitungannya:

STEP 1

Menghitung peluang class `Y` dan `N`.

$P(class = Y) = \frac{5}{10} = 0.5$

$P(class = N) = \frac{5}{10} = 0.5$

STEP II

Menghitung peluang bersyarat untuk atribut A.

Peluang_A
P(A = 0| N) = 0.4
P(A = 1| N) = 0.2
P(A = 2| N) = 0.4
P(A = 0| Y) = 0.2
P(A = 1| Y) = 0.6
P(A = 2| Y) = 0.2

STEP III

Menghitung peluang bersyarat untuk atribut B.

Peluang_B
P(B = 0| N) = 0.4
P(B = 1| N) = 0.4
P(B = 2| N) = 0.2
P(B = 0| Y) = 0.2
P(B = 1| Y) = 0.4
P(B = 2| Y) = 0.4

STEP IV

Menghitung peluang bersyarat untuk atribut C.

Peluang_C
P(C = 0| N) = 0.6
P(C = 1| N) = 0.2
P(C = 2| N) = 0.2
P(C = 0| Y) = 0.2
P(C = 1| Y) = 0.4
P(C = 2| Y) = 0.4

STEP V

Berikut adalah *summary* peluang yang ada:

$P(class = Y) = \frac{5}{10} = 0.5$

$P(class = N) = \frac{5}{10} = 0.5$

dan

Table 10: Tabel Peluang NB

Peluang_A	Peluang_B	Peluang_C
P(A = 0| N) = 0.4	P(B = 0| N) = 0.4	P(C = 0| N) = 0.6
P(A = 1| N) = 0.2	P(B = 1| N) = 0.4	P(C = 1| N) = 0.2
P(A = 2| N) = 0.4	P(B = 2| N) = 0.2	P(C = 2| N) = 0.2
P(A = 0| Y) = 0.2	P(B = 0| Y) = 0.2	P(C = 0| Y) = 0.2
P(A = 1| Y) = 0.6	P(B = 1| Y) = 0.4	P(C = 1| Y) = 0.4
P(A = 2| Y) = 0.2	P(B = 2| Y) = 0.4	P(C = 2| Y) = 0.4

SOAL 3 b

Buat matriks/tabel klasifikasi/misklasifikasi dan hitung nilai akurasi, presisi, dan *recall*!

Jawab

Sekarang dari data yang ada berikut ini:

id	A	B	C	class
1	0	2	1	Y
2	1	2	2	Y
3	1	1	0	Y
4	2	0	2	Y
5	1	1	1	Y
6	2	2	2	N
7	0	1	0	N
8	0	0	1	N
9	2	1	0	N
10	1	0	0	N

Saya akan lakukan prediksi berdasarkan `Tabel Peluang NB` pada jawaban sebelumnya.

Saya akan berikan gambaran untuk `id = 1`

$X = (A = 0, B = 2, C = 1)$

$P(X|N) = 0.4 \times 0.2 \times 0.2 \times 0.5 = 0.008$

$P(X|Y) = 0.2 \times 0.4 \times 0.4 \times 0.5 = 0.016$

$P(X|Y) > P(X|N)$

Kesimpulan: Maka pada `id = 1` diprediksi nilainya adalah `Y`.

Kita lakukan hal serupa untuk baris yang lain.

id	class	P(X|N)	P(X|Y)	class_predict
1	Y	0.008	0.016	Y
2	Y	0.004	0.048	Y
3	Y	0.024	0.024	N
4	Y	0.016	0.008	N
5	Y	0.008	0.048	Y
6	N	0.008	0.016	Y
7	N	0.048	0.008	N
8	N	0.016	0.008	N
9	N	0.048	0.008	N
10	N	0.024	0.012	N

Sekarang kita buat *confusion matrix* dari hasil di atas:

```
##
##     N Y
##   N 4 1
##   Y 2 3
```

Sumbu Y - Prediksi

Sumbu X - *Actual*

Akurasi = $\frac{4+3}{10} = 0.7$

Presisi = $\frac{3}{3+1} = 0.75$

Recall = $\frac{3}{3+2} = 0.6$