Introduction Data Mining

Basic Data Mining

Data Mining adalah suatu proses yang dilakukan untuk menemukan *pattern/insights* dalam sebuah dataset yang berukuran besar yang melibatkan proses machine learning, statistics, dan database. Insights yang di dapatkan dari data mining sendiri akan digunakan untuk system pengambilan keputusan. Data mining sendiri pada saat ini sudah bisa di implementasikan di berbagai bidang industry yang antara lain adalah finance, otomotif, marketing. Human resource, dan lain-lain.

Kasus dari Data Mining Sendiri dibagi menjadi 2 permaslaahan yaitu unsupervised learning dan supervised learning.

Supervised Learning

- Regresi

Regresi adalah kasus dari maching learning yang memiliki label berupa bilangan kontinyu seperti contohnya harga rumah. Regresi bertujuan untuk menemukan suatu fungsi yang memodelkan data dengan meminimalkan *error* atau selisih antara nilai prediksi dengan nilai sebenarnya.

- Klasifikasi

Klasifikasi adalah sebuah teknik untuk mengklasifikasikan atau mengkategorikan beberapa *item* yang belum berlabel ke dalam sebuah set kelas diskrit. Klasifikasi mencoba mempelajari hubungan antara kumpulan variabel fitur dan variabel target. Dalam klasifikasi, variabel targetnya bertipe kategori.

Unsupervised Learning

- Klastering

Clustering dilakukan jika kita ingin menemukan klaster dari sebuah dataset. Klaster adalah sebuah kumpulan data atau objek yang memiliki kemiripan satu sama lain di dalam kumpulan atau kelompok tersebut, dan berbeda dengan objek di kelompok lain. Tidak seperti klasifikasi dimana setiap data latih sudah memiliki label alias sudah ditentukan kelasnya, dalam clustering, data tidak berlabel.

Data dam Attributes

Data merupakan *bahan* atau komponen utama yang sangat diperlukan dalam Data Mining. Data biasa nya tersusun dari objek. Data yang yang dikumpulkan menjadi satuan besar bisa disebut dengan Dataset dan kumpulan dari Dataset disebut juga dengan Database.

Jenis Nilai Atribut

- Nominal

Nominal berarti "yang berkaitan dengan nama-nama." Nilai-nilai atribut nominal adalah simbol atau nama-nama dari suatu benda. Setiap nilai merupakan semacam kategori, kode, atau status dan sebagainya sehingga atribut nominal juga disebut sebagai kategorikal. Nilai-nilai di dalamnya tidak memiliki urutan. Contoh warna rambut, customer id, dll

- Ordinal

Sebuah atribut ordinal adalah atribut dengan nilai-nilai yang memiliki urutan atau peringkat, tapi besaran nilai-nilai yang berurutan tidak diketahui. Contoh nilai(A, B C, D), nilai rating, dll

- Interval

Memiliki nilai yang sudah pasti (tetap) diantara tiap data interval. Contoh : tanggal kalender (diantara dua tanggal tidak ada tanggal lain), temperatur dalam Celcius atau Fahrenheit.

- Rasio

Atribut numerik dengan titik nol absoulut. Artinya kita dapat menghitung perkalian atau perbandingan antara suatu nilai dengan nilai yang lain. Selain itu, nilai-nilai tersebut juga bisa diurutkan, dihitung perbedaan/selisihnya, bisa dihitung mean (rata-rata), median (nilai tengah), dan modus (yang paling sering muncul). Contoh : Contoh : temperatur dalam Kelvin, panjang, waktu, massa dan jumlah. Operasi yang bisa dilakukan adalah semua operasi yang ada.

Similarity and Dissimilarity Data

Kemiripan (*similarity*) adalah ukuran numerik dimana dua objeknya mirip, nilai 0 jika tidak mirip dan nilai 1 jika mirip penuh. Sementara ketidakmiripan (*dissimilarity*) adalah derajat numerik dimana dua objek yang berbeda, jangkauan nilai 0 sampai 1.

Istilah ketidakmiripan juga dapat disebut sebagai ukuran jarak (*distance*) antara dua data. Jika s adalah ukuran kemiripan dan d adalah ukuran ketidakmiripan, serta jika interval/range nilainya adalah [0,1], maka dapat dirumuskan bahwa s+d=1. Sebenarnya ukuran kemiripan dan ketidakmiripan tidak harus selalu dalam interval [0,1], tetapi boleh juga menggunakan interval seperti [0,10], [0,100], bahkan menggunakan nilai negative seperti [-1,1], [-10,10] dan sebagainya. Transformasi nilai s dan d tidak hanya terbatas pada formula s+d=1, karena ada juga yang menggunakan s = 11+d atau s = e-d.

- Jarak Euclidean

Euclidean distance adalah perhitungan jarak dari 2 buah titik dalam Euclidean space. Euclidean ini berkaitan dengan Teorema Phytagoras dan biasanya diterapkan pada 1, 2 dan 3 dimensi. Tapi juga sederhana jika diterapkan pada dimensi yang lebih tinggi. Jarak Euclidean sendiri memiliki rumus seperti di bawah.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- Jarak Manhattan

Merupakan salah satu pengukuran yang paling banyak digunakan meliputi penggantian perbedaan kuadrat dengan menjumlahkan perbedaan absolute dari variabel. Prosedur ini disebut blok absolute atau lebih dikenal dengan *city block distance.*

$$d(x, y) = \left(\sum_i (x_i - y_i)^2 \right)^{\frac{1}{2}}$$

Basic Python for Data Mining

Python merupakan Bahasa pemrograman tingkat tinggi yang dikenal juga sebagai pemrograman interpreter. Python dapat dijalankan di command prompt, script sampai dengan Notebook.

Loop pada Python

Loop atau perulangan adalah pernyataan pada bahasa pemrograman akan dieksekusi secara berurutan. Pernyataan pertama dalam sebuah fungsi dijalankan pertama, diikuti oleh yang kedua, dan seterusnya. Tetapi akan ada situasi dimana Anda harus menulis banyak kode, dimana kode tersebut sangat banyak. Jika dilakukan secara manual maka Anda hanya akan membuang-buang tenaga dengan menulis beratus-ratus bahkan beribu-ribu kode. Untuk itu Anda perlu menggunakan pengulangan di dalam bahasa pemrograman Python.

While Loop

Pengulangan While Loop di dalam bahasa pemrograman Python dieksesusi statement berkali-kali selama kondisi bernilai benar atau True. Dibawah ini adalah contoh penggunaan pengulangan While Loop.

```
count = 0
while (count < 9):
    print ("The count is: ", count)
    count = count + 1

print ("Good bye!")
```

For Loop

Pengulangan for pada Python memiliki kemampuan untuk mengulangi item dari urutan apapun, seperti list atau string. Dibawah ini adalah contoh penggunaan pengulangan For Loop.

```
angka = [1,2,3,4,5]
for x in angka:
    print(x)

#Contoh pengulangan for
buah = ["nanas", "apel", "jeruk"]
for makanan in buah:
    print ("Saya suka makan", makanan)
```