Data Mining

Munawar, PhD

14. Review

Unsupervised Learning

- Proses suatu sistem mempelajari unlabeled data berdasarkan fitur –fitur dari data tersebut,
- Tujuan akhir dari *unsupervised learning* adalah mengelompok data – data ke dalam suatu group yang berupa cluster terdiri dari data yang memiliki kemiripan yang sama untuk satu cluster dan apa bila ditemukan perbedaan dapat di kelompokkan ke dalam cluster yang lain atau dapat diangap sebagai *outlier.*

Unsupervised Learning …

- Algoritma data mining mencari pola dari semua variable (atribut).
- Dataset tidak memiliki label/*class*
- Contoh algoritma *unsupervised learning* adalah algoritma clustering,maupun SOM
- Algoritma *clustering* : K-Means, K-Medoids, Hierarcichal Clustering

Decision Tree

- Salah satu algoritma klasifikasi yang sangat powerful
- Waktu komputasi lebih singkat dibandingkan yang lain
- Rule-rule yang sederhana dan mudah untuk dimengerti

Algoritma Decision Tree

1. Siapkan data *training* (data latih)
2. Pilih atribut sebagai akar

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

3. Buat cabang untuk tiap –tiap nilai
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yg sama

Naïve Bayes

- Naive Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*.
- Naive Bayes didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network.
- Naive Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar

Formula Naïve Bayes

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$$

- X adalah data sample yang belum diketahui kelasnya
- H adalah dugaan bahwa X adalah anggota C
- Klasifikasi ditentukan oleh P(H|X) , *(posteriori probability)*, probabilitas bahwa dugaan terhadap data *sample* X
- P(H) adalah *prior probability*
- Probabilitas dari sample data yang diamati
- P(X|H) (likelyhood), probabilitas dari sample X dengan memperhatikan dugaan.

Klasifikasi Naïve Bayes

- Misal D adalah *record data training* dan setiap *record* terdapat label kelasnya dan masing-masing *record* dinyatakan n atribut (n field) X= (x1,x2, ….., xn).
- Misal terdapat *m* kelas C1, C2, C3,… Cm
- Klasifikasi diperoleh maksimum posterior yaitu, maksimum P(Ci|X)
- Ini dapat diperoleh dari teorema Bayes

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Karena P(X) adalah konstan untuk semua kelas, hanya

Perlu dimaksimumkan

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

Algoritma Naïve Bayes

1. Baca Data Training
2. Hitung jumlah class
3. Hitung jumlah kasus yang sama dengan class yang sama
4. Kalikan semua nilai hasil sesuai dengan data X yang dicari class-nya

Feature Selection

Feature Selection

=

Variabel Selection

=

Variabel Subset Selection

=

Attribute Selection

Proses pemilihan subset dari fitur yang relevan (variabel, prediktor) untuk digunakan dalam konstruksi model.

Teknik pemilihan fitur digunakan untuk empat alasan:

1. Penyederhanaan model agar lebih mudah ditafsirkan oleh peneliti / pengguna
2. Waktu pelatihan yang lebih pendek
3. Untuk menghindari kutukan dimensi
4. Generalisasi yang disempurnakan dengan mengurangi overfitting (secara formal, pengurangan varians)

Feature Selection

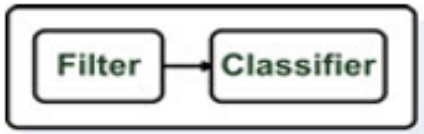
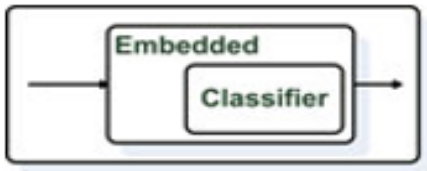
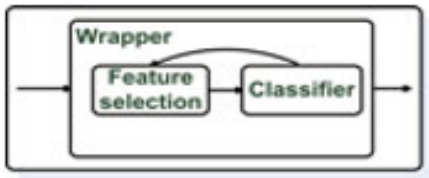
Tujuan *feature selection* adalah:

- Mengurangi jumlah fitur yang terlibat dalam menentukan suatu nilai kelas target.
- Mengurangi fitur *irelevan*
- Mengurangi data yang berlebihan
- Mengurangi data yang menyebabkan salah pengertian terhadap kelas target yang membuat efek segera bagi aplikasi

Aplikasi data mining bisa dipercepat

Mempertinggi kinerja mining, seperti akurasi peramalan

Teknik Feature Selection

Method	Advantages	Disadvantages
Filter 	Independence of the classifier Lower computational cost than wrappers Fast Good generalization ability	No interaction with the classifier
Embedded 	Interaction with the classifier Lower computational cost than wrappers Captures feature dependencies	Classifier-dependent selection
Wrapper 	Interaction with the classifier Captures feature dependencies	Computationally expensive Risk of overfitting Classifier-dependent selection

Complex Data Type

Berkembangnya data komplek

- **Spatial data:** Data geographis, data kesehatan dan data gambar satellite
- **Multimedia data:** images, audio, dan video
- **Time-series data:** Data perbangkan dan stock exchange data

FOCUS

- **Text data:** Word descriptions for objects
- **World-Wide-Web:** teks dan data multimedia yang sangat tidak terstruktur

Text Mining

- Text mining merujuk pada data mining yang menggunakan dokumen teks sebagai data
- Hampir semua tugas Text Mining menggunakan metode Information Retrieval (IR) untuk pra-proses dokumen teks.
- Metode ini sedikit berbeda daripada metode pra-proses data yang digunakan dalam tabel relasional
- Web search juga berakar pada IR

Text Mining

Menemukan informasi yang berguna dari kumpulan teks besar dimana informasi sebelumnya tidak diketahui

- Pola
- Trends
- Associations (Hubungan yang menarik yang tersembunyi dalam dataset besar).

Definisi Text Mining

Text mining di pahami sebagai proses secara otomatis untuk mengekstrak informasi yang bermakna, berguna, dimana sebelumnya tidak di ketahui dan pada akhirnya dapat di pahami dari penyimpanan dokumen tekstual.

Text Mining

= Data Mining (yang diterapkan dalam bentuk data teks)

+ basic linguistic

Text Mining

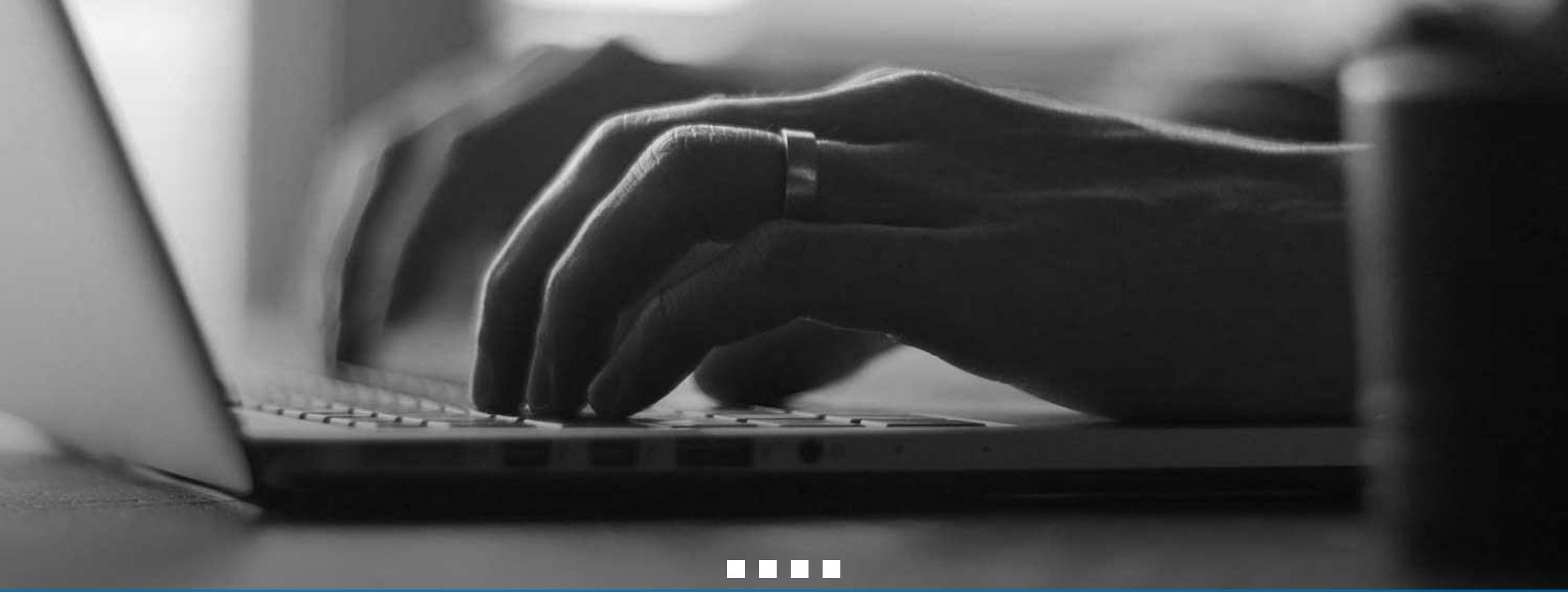
Bagaimana *Text Mining* bekerja?

1. Langkah dasar dalam *text mining* melibatkan konversi teks menjadi data semi terstruktur

2. Setelah mengubah teks yang tidak terstruktur menjadi data semi terstruktur, langkah selanjutnya adalah menerapkan teknik analisis untuk proses *classification, clustering, prediction*

3. Menemukan pola yang lebih baik dari hasil perubahan teks yang tidak terstruktur menjadi data semi terstruktur

4. Melakukan pelatihan model untuk mendeteksi pola pada teks baru dan tidak terlihat



THANK YOU

Munawar, PhD – moenawar@gmail.com – www.moenawar.web.id