KOM341
Temu Kembali Informasi

KULIAH #8
- Text Classification

Ad Hoc Retrieval

- User mencari informasi dengan memberikan satu atau lebih query terhadap koleksi terkini.
- Contoh: mencari *multicore computer chips* terbaru.
  - Query : multicore AND computer AND chip
  - Akan dieksekusi setiap ada penambahan dokumen baru → standing query
  - Mungkin tidak menemukan artikel baru lain yang relevan, misalnya multicore processors.
  - Gunakan Boolean: (multicore OR multi-core) AND (chip OR processor OR microprocessor)

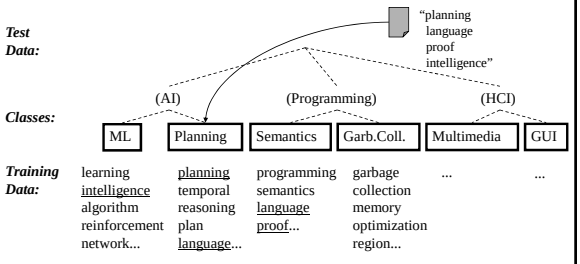
Classification

- Lebih mudah kalau dokumen dikelompokkan menjadi misalnya dua kelas, yaitu dokumen tentang multicore computer chips dan dokumen BUKAN tentang multicore computer chips.
- Kelas biasanya merujuk ke topik dokumen.
- Prosesnya sering disebut sebagai text classification, text categorization, topic classification, topic spotting.

Categorization / Classification

- Given:
  - Deskripsi dokumen $d \in X$, dimana X adalah kumpulan dokumen.
  - Himpunan kelas atau kategori: $C = \{c_1, c_2, ..., c_n\}$
- Tujuan:
  - Menentukan kategori dari d: $c(d) \in C$, *dimana* $c(d)$ adalah fungsi kategorisasi (*classifier*).

Document Classification



Learning Method

- Kita mempelajari fungsi klasifikasi yang memetakan dokumen ke kategori tertentu:

$$\gamma : X \rightarrow C$$

- Disebut juga *supervised learning*, karena supervisor (orang yang menentukan kategori dokumen) berperan langsung di dalam proses pembelajaran.

Metode

- Manual
  - Digunakan oleh Yahoo!, Looksmart, about.com, ODP, Medline
  - Sangat akurat karena dilakukan oleh ahli.
  - Konsisten pada saat ukurannya kecil/sedikit.
  - Sulit dan mahal

Metode

- Automatic document classification
  - Hand-coded rule-based systems
    - Digunakan oleh CS dept's spam filter, Reuters, CIA, Verity, …
    - Masukkan ke kategori jika dokumen mengandung kombinasi kata tertentu.
    - Akurasi tinggi jika rule dibuat dengan sangat baik oleh ahli dan kompleks.

Metode

- Automatic document classification
  - Supervised learning
    - Beberapa menggunakan machine learning (Autonomy, MSN, Verity, Enkata, Yahoo!, …)
      - k-Nearest Neighbors (simple, powerful)
      - Naive Bayes (simple, common method)
      - Support-vector machines (new, more powerful)
      - dsb
      - Membutuhkan hand-classified training data
      - Data dapat dibangun oleh amatir
- Banyak sistem komersial menggunakan metode campuran

Metode Bayes

- Berbasis teori peluang
- Utamanya teorema Bayes
- Untuk kejadian ***a*** dan ***b***, Bayes Rules:

$$p(a,b) = p(a \cap b) = p(a \mid b)\,p(b) = p(b \mid a)\,p(a)$$

$$p(\bar{a} \mid b)\,p(b) = p(b \mid \bar{a})\,p(\bar{a})$$

$$p(a \mid b) = \frac{p(b \mid a)\,p(a)}{p(b)} = \frac{p(b \mid a)\,p(a)}{\sum_{x=a,\bar{a}} p(b \mid x)\,p(x)}$$

Prior

Posterior

Naïve Bayes Model

- Supervised learning method
- Multinomial Naïve Bayes Model
- Peluang dokumen ***d*** dalam kelas ***c*** :

$$P(c \mid d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k \mid c)$$

dimana $P(t_k|c)$ adalah peluang term t_k muncul pada dokumen kelas c, P(c) peluang dokumen ada pada kelas c.

Pendugaan Parameter

Pendugaan parameter

$$\hat{P}(c) = \frac{N_c}{N}, \quad \hat{P}(t \mid c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

dimana N_c adalah banyaknya dokumen dalam kelas c, N adalah total dokumen, T_{ct} adalah banyaknya t dalam dokumen training dari kelas c.

Laplace smoothing

- Atau Add-One Smoothing.
- Untuk menghilangkan dugaan parameter yang bernilai nol.

$$\hat{P}(t \mid c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{\left(\sum_{t' \in V} T_{ct'} \right) + B'}$$

dimana B = |V| = banyaknya term dalam vocabulary.

Contoh

	docID	words in document	in c = China?
TRAINING SET	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
TEST SET	5	Chinese Chinese Chinese Tokyo Japan	?

P(c) = ¾ dan P(¬c) = ¼
P(Chinese|c) = (5+1)/(8+6) = 6/14 = 3/7
P(Tokyo|c) = P(Japan|c) = (0+1)/(8+6) = 1/14
P(Chinese| ¬c) = (1+1)/(3+6) = 2/9
P(Tokyo| ¬ c) = P(Japan| ¬ c) = (1+1)/(3+6) = 2/9

$$\hat{P}(c \mid d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003$$

$$\hat{P}(\bar{c} \mid d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001$$

Bernoulli Model

- Kejadian Bernoulli
- Multivariate Bernoulli Model
- $\hat{P}(t \mid c)$: rasio dokumen dari kelas c yang mengandung term t. Dalam multinomial didefinisikan sebagai rasio token dalam dokumen kelas c yang mengandung term t.

Contoh

	docID	words in document	in c = China?
TRAINING SET	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
TEST SET	5	Chinese Chinese Chinese Tokyo Japan	?

P(c) = ¾ dan P(¬c) = ¼
P(Chinese|c) = (3+1)/(3+2) = 4/5
P(Tokyo|c) = P(Japan|c) = (0+1)/(3+2) = 1/5
P(Beijing|c) = P(Shanghai|c) = P(Macao|c) = (1+1)/(3+2) = 2/5
P(Chinese| ¬c) = (1+1)/(1+2) = 2/3
P(Tokyo| ¬ c) = P(Japan| ¬ c) = (1+1)/(1+2) = 2/3
P(Beijing| ¬ c) = P(Shanghai|¬c) = P(Macao|¬c) = (0+1)/(1+2) = 1/3

Contoh

$$\hat{P}(c \mid d_5) \propto \hat{P}(c).\hat{P}(Chinese \mid c).\hat{P}(Japan \mid c).\hat{P}(Tokyo \mid c)$$
$$.(1-\hat{P}(Beijing \mid c)).(1-\hat{P}(Shanghai \mid c)).(1-\hat{P}(Macao \mid c))$$
$$= 3/4 \cdot 4/5 \cdot 1/5 \cdot (1-2/5).(1-2/5).(1-2/5) \approx 0.005$$

$$\hat{P}(\bar{c} \mid d_5) \propto 1/4 \cdot 2/3 \cdot 2/3 \cdot 2/3.(1-1/3).(1-1/3).(1-1/3) \approx 0.022$$

Jadi, dokumen d_5 diklasifikasikan ke ¬c (bukan China)

Maximum a Posteriori

- Tujuan klasifikasi: mendapatkan kelas terbaik untuk suatu dokumen.
- Kelas terbaik : sangat mirip atau maximum a posteriori (MAP) kelas c_{map} :

$$c_{map} = \arg\max_{c \in C} \hat{P}(c \mid d) = \arg\max_{c \in C} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k \mid c)$$

diduga dari training set

Maximum a Posteriori

$$c_{map} = \arg\max_{c \in C} P(c \mid d) = \arg\max_{c \in C} \frac{P(d \mid c)P(c)}{P(d)}$$
$$= \arg\max_{c \in C} P(d \mid c)P(c)$$

- Multinomial P(d|c)=P(<t_1, …, t_k, …, t_{nd}>|c)
- Bernoulli P(d|c)=P(<e_1, …, e_k, …, e_M>|c)

Asumsi Saling Bebas

- Kejadian A dan B saling bebas
  P(A∩B) = P(A,B) = P(A).P(B)
- Maka:

Multinomial $P(d \mid c) = P(\langle t_1,...,t_{n_d} \rangle \mid c) = \prod_{1 \le k \le n_d} P(X_k = t_k \mid c)$

Bernoulli $P(d \mid c) = P(\langle e_1,...,e_M \rangle \mid c) = \prod_{1 \le i \le M} P(U_i = e_i \mid c)$

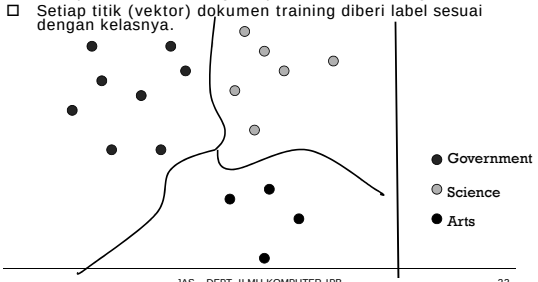
Multinomial vs Bernoulli

	multinomial model	Bernoulli model
event model	generation of token	generation of document
random variable(s)	$X = t$ iff t occurs at given pos	$U_t = 1$ iff t occurs in doc
document representation	$d = \langle t_1, \ldots, t_k, \ldots, t_{n_d} \rangle, t_k \in V$	$d = \langle e_1, \ldots, e_i, \ldots, e_M \rangle, e_i \in \{0, 1\}$
parameter estimation	$\hat{P}(X = t \mid c)$	$\hat{P}(U_i = e \mid c)$
decision rule: maximize	$\hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(X = t_k \mid c)$	$\hat{P}(c) \prod_{t_i \in V} \hat{P}(U_i = e_i \mid c)$
multiple occurrences	taken into account	ignored
length of docs	can handle longer docs	works best for short docs
# features	can handle more	works best with fewer
estimate for term the	$\hat{P}(X = \text{the} \mid c) \approx 0.05$	$\hat{P}(U_{\text{the}} = 1 \mid c) \approx 1.0$

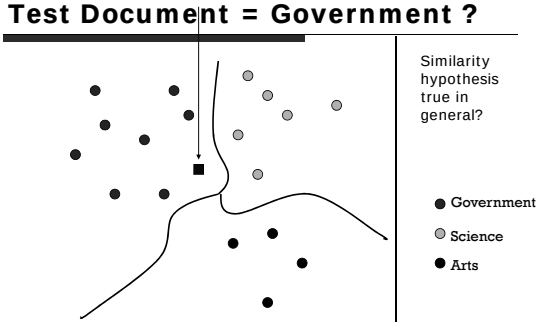
Vector Space Classification

Klasifikasi Menggunakan Ruang Vektor

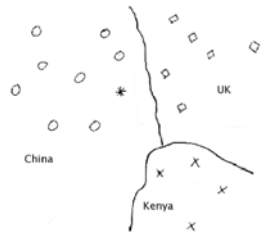
- Setiap dokumen training direpresentasikan sebagai vektor.
- Setiap titik (vektor) dokumen training diberi label sesuai dengan kelasnya.



Test Document = Government ?



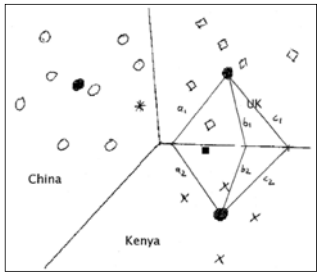
Rocchio Classification



Centroid dari kelas c:

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d)$$

Rocchio Classification



Batas antara dua kelas adalah titik yang memiliki jarak sama ke kedua centroid-nya → $|a_1| = |a_2|$, $|b_1| = |b_2|$, $|c_1| = |c_2|$

Rocchio Classification

Dokumen d dikelompokkan ke dalam kelas c

- Menggunakan jarak

$$\arg\min_c |\vec{\mu}_c - \vec{v}(d)|$$

- Menggunakan ukuran kesamaan Cosine

$$\arg\max_c \cos(\vec{\mu}(c), \vec{v}(d))$$

Contoh

Dari contoh sebelumnya, diperoleh:

	term weights					
vector	Chinese	Japan	Tokyo	Macao	Beijing	Shanghai
$\vec{d}_1$	0	0	0	0	1.0	0
$\vec{d}_2$	0	0	0	0	0	1.0
$\vec{d}_3$	0	0	0	1.0	0	0
$\vec{d}_4$	0	0.71	0.71	0	0	0
$\vec{d}_5$	0	0.71	0.71	0	0	0
$\vec{\mu}_c$	0	0	0	0.33	0.33	0.33
$\vec{\mu}_{\bar{c}}$	0	0.71	0.71	0	0	0

Jarak d_5 terhadap centroid:

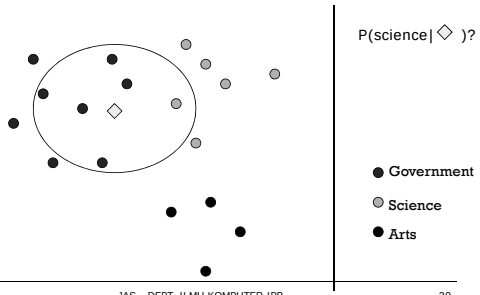
- $|\mu_c - d_5| \approx 1.15$ dan $|\mu_{\neg c} - d_5| \approx 0.00$
- maka Rocchio mengklasifikasikan d_5 ke kelas $\neg c$ (bukan China).

k Nearest Neighbor Classification

- Mengklasifikasikan dokumen ***d*** ke dalam kelas ***c***
- Tentukan k-neighborhood ***N*** atau kNN sebagai ***k*** terdekat dari ***d***
- Hitung banyaknya dokumen ***i*** dalam ***N*** pada kelas ***c***
- Duga nilai P(c|d) = i/k
- Pilih

$$c_{map} = \arg\max_{c \in C} P(c \mid d)$$

Contoh: k=6 (6NN)



Ukuran Kemiripan

- Metode kNN tergantung pada ukuran kemiripan (bisa juga jarak) yang digunakan.
- Paling sederhana adalah jarak Euclidean.
- Untuk teks, yang paling efektif adalah ukuran kemiripan cosine dengan bobot vektor tf.idf.
- Skor dokumen di suatu kelas:

$$\text{score}(c,d) = \sum_{d' \in S_k} I_c(d') \cos(\vec{v}(d'), \vec{v}(d))$$

dimana $I_c(d')=1$ jjk d' ada dalam kelas c, dan sebaliknya = 0.

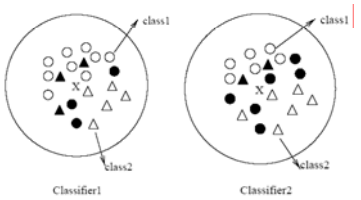
Contoh : 1NN

	term weights					
vector	Chinese	Japan	Tokyo	Macao	Beijing	Shanghai
$\vec{d}_1$	0	0	0	0	1.0	0
$\vec{d}_2$	0	0	0	0	0	1.0
$\vec{d}_3$	0	0	0	1.0	0	0
$\vec{d}_4$	0	0.71	0.71	0	0	0
$\vec{d}_5$	0	0.71	0.71	0	0	0

Dengan menggunakan jarak Euclidean, maka:
$|d_1-d_5| = |d_2-d_5| = |d_3-d_5| = 1.4171$
$|d_4-d_5| = 0.0000$
Maka d_5 lebih dekat ke kelas d_4.

Kombinasi Metode Klasifikasi

Beberapa peneliti menunjukkan bahwa kombinasi beberapa classifier yang berbeda dapat meningkatkan akurasi.



Classifier 1:
X → class1
Classifier 2:
X → class2

Jadi, X dimasukkan kemana?

Kombinasi Metode Klasifikasi

- Simple voting
  Untuk tiap dokumen test, kita klasifikasikan ke kelas c_i jika mayoritas classifier memasukkan dokumen test ke kelas c_i.
- Dynamic classifier selection (DCS)
  Pendekatan kNN dengan ukuran kesamaan Cosine, dilakukan iterasi.
- Adaptive classifier combination (ACC)
  Kombinasi NB dan kNN