

METODE INTELIGENTE DE REZOLVARE A PROBLEMELOR REALE



Laura Dioşan
Tema 4

Text mining

□ Task-uri

- Regăsirea informației
- Clasificarea automată a textelor

Text mining

□ Task-uri

- **Regăsirea informației**
- Clasificarea automată a textelor

Regăsirea informației

- Definire
- Tipologie
- Proces
- Evaluare

Regăsirea informației – definire

□ Alte denumiri

- *Information retrieval (IR),*
- *Information storage and retrieval (ISR)*
- *Information organization and retrieval (IOR)*

□ Definiție

- Regăsirea într-o colecție de obiecte a unei submulțimi de obiecte care servesc unui anumit scop
 - Ex.
 - Pagini web pt pregătirea unei excursii
 - Materiale educaționale pentru învățarea unui concept

Regăsirea informației – tipologie

- În funcție de tipul de informație
 - Regăsirea textelor → text mining
 - Regăsirea imaginilor
 - Regăsirea muzicii
 - Regăsirea vorbirii
 - Regăsirea încrucișată a limbajului
 - Întrebarea într-o limbă, răspunsul în altă(e) limbă(i)

Regăsirea informației – proces

Pași în procesul de regăsire

- ❑ Indexarea și reprezentarea obiectelor din baza de cunoștințe
- ❑ Formularea interogării
- ❑ Potrivirea interogării cu obiectele
- ❑ Selectarea rezultatelor

Regăsirea informației – proces

- Indexarea obiectelor din baza de cunoștințe
 - fixarea unei anumite reprezentări a obiectelor
 - poate fi
 - manuală
 - automată
 - extragerea unor attribute (brute)
 - texte – separarea în cuvinte, eliminarea cuvintelor vide, etc
 - imagini – distribuția culorilor și a formelor
 - muzică – frecvența notelor
- Formularea interogării
 - Fixarea unei anumite reprezentări a interogării
 - Interogarea → un profil (șablon) pe care îl vor respecta anumite obiecte (documente)
 - texte → anumite cuvinte care trebuie să apară în text
 - imagini → anumite culori sau forme care trebuie să apară în imagini
 - muzică → anumite (succesiuni de) note care trebuie să apară în melodii

Regăsirea informației – proces

- ❑ Potrivirea interogării cu obiectele
 - Cu ajutorul unei funcții de similaritate sau de tip rang
 - Tipologie
 - ❑ potrivire perfectă (exactă)
 - ❑ potrivire parțială
- ❑ Selectarea rezultatelor
 - ordonarea lor
 - gruparea lor

Regăsirea informației - evaluare

Măsuri de performanță

□ Precizia

- proporția obiectelor regăsite care sunt relevante
- $\text{nr. obiecte relevante regăsite} / \text{nr. obiecte regăsite}$

□ Rapelul

- proporția obiectelor relevante care sunt regăsite
- $\text{nr. obiecte relevante regăsite} / \text{nr. obiecte relevante}$

□ Acuratețea

- proporția obiectelor corect regăsite

□ Scorul F1

- media armonică a preciziei și rapelului

Text mining

□ Task-uri

- Regăsirea informației
- **Clasificarea automată a textelor**

Clasificarea automată a textelor

□ Definire

□ Direcții în automatizare

- Abordarea bazată pe învățare
- Abordarea bazată pe cunoștințe

Clasificarea automată a textelor

Definire

□ Categorizarea textelor

- Atribuirea unor categorii (predefinite) documentelor
- Documentele
 - rapoarte tehnice, pagini web, mesaje, cărți
- Categoriile
 - subiecte (artă, economie),
 - pertinente (mesaje spam, pagini web pt adulți)

□ Exemple de probleme

	Cuvinte	Documente
Învățare supervizată	Etichetarea părților de vorbire	Clasificarea textelor, Filtrarea, Detectarea subiectelor
Învățare nesupervizată	Indexarea semantică, construcția automată a tezaurelor, extragerea cuvintelor cheie	Clusterizarea documentelor, Detectarea subiectelor

Clasificarea automată a textelor

Dirrecții în automatizare

- Abordarea bazată pe învățare
 - Experții etichetează o parte din exemple
 - Algoritmul etichetează noi exemple
 - Învățarea poate fi:
 - supervizată
 - nesupervizată
- Abordarea bazată pe cunoștințe
 - Cunoștințele despre clasificare sunt
 - obținute de la experți
 - codificate sub formă de reguli

Clasificarea automată a textelor – Învățare – definire

□ Definirea problemei

- Se dă un set de documente D , $|D|=N+n$ și un set de categorii C , $|C|=k$, sub forma

- date de antrenament – (d_i, c_i) , unde
 - $i = 1, N$ (N = nr datelor de antrenament)
 - $d_i \in D, c_i \in C$
- date de test
 - $(d_i), i = 1, n$ (n = nr datelor de test)

- Se cere să se aproximeze o funcție necunoscută de clasificare

$$\Phi: D \times C \rightarrow \{\text{true}, \text{false}\}$$

definită astfel:

- $\Phi(d, c) = \text{true}$, dacă $d \in c$
 false , altfel

pentru orice pereche de documente și categorii (d, c) .

Clasificarea automată a textelor – Învăţare – definire

□ Tipuri de categorii

- În funcţie de modul de organizare
 - Categorii ierarhice
 - Directoarele de e-mail, MESH
 - Categorii liniare
 - Secţiunile unui ziar, Reuters
- În funcţie de apartenenţa documentelor la categorii
 - Categorii suprapuse
 - Reuters, MESH
 - Categorii disjuncte
 - Directoarele de e-mail, secţiunile unui ziar

Clasificarea automată a textelor – Învățare – proces

- Analiza documentelor de antrenament
 - Indexarea documentelor
 - Construirea unei reprezentări a documentelor → transformarea documentelor într-o formă interpretabilă de către clasificator
 - Obținerea unor concepte/termeni reprezentative(i) → attribute
 - Calcularea unor ponderi pt aceste attribute
 - Reducerea dimensiunii (a numărului de concepte/attribute/termeni reprezentative(i) pentru document)
 - Selecția atributelor
 - Extragerea atributelor
 - Învățarea unui model de clasificare
- Clasificarea noilor documente(de test)
 - Indexarea documentelor
 - Utilizarea modelului de clasificare pentru stabilirea categoriilor fiecărui document de test

Clasificarea automată a textelor – Învățare – proces

- Analiza documentelor de antrenament
 - **Indexarea documentelor**
 - Construirea unei reprezentări a documentelor → transformarea documentelor într-o formă interpretabilă de către clasificator
 - Obținerea unor concepte/termeni reprezentative(i) → attribute
 - Calcularea unor ponderi pt aceste attribute
 - Reducerea dimensiunii (a numărului de concepte/attribute/termeni reprezentative(i) pentru document)
 - Selecția atributelor
 - Extragerea atributelor
 - Învățarea unui model de clasificare
- Clasificarea noilor documente(de test)
 - Indexarea documentelor
 - Utilizarea modelului de clasificare pentru stabilirea categoriilor fiecărui document de test

Clasificarea automată a textelor – Învățare – proces

□ Indexarea documentelor

- Construirea unei reprezentări a documentelor → transformarea documentelor într-o formă
 - interpretabilă de către clasificator
 - indexată (organizată, ordonată)
- Obținerea unor concepte/termeni reprezentative(i) → attribute și calcularea unor ponderi pt aceste attribute
- 4 pași:
 - Linearizarea documentelor
 - Filtrarea
 - Aducerea la formă canonică
 - Ponderarea

} Reducerea dimensiunii
vocabularului

Clasificarea automată a textelor – Învățare – proces

- Linearizarea documentelor (segmentare)
 - Procesul de reducere a documentelor la un vector de termeni (attribute)
 - modelul sac de cuvinte (*bag of words*)
 - o matrice
 - pe linii documentele
 - pe coloane termenii
 - o celulă → 1/0 – dacă termenul curent apare în documentul curent
 - Identificarea termenilor se face în 2 etape:
 - Înlăturarea formatării
 - Ex. eliminarea etichetelor în cazul documentelor HTML
 - *Tokenization*
 - Parsare (segmentare)
 - Transformarea tuturor literelor în litere mici
 - Înlăturarea semnelor de punctuație

Inițial	Liniazat
Interactive query expansion modifies queries using terms from a user. Automatic query expansion expands queries automatically.	interactive query expansion modifies queries using terms from a user automatic query expansion expands queries automatically

Clasificarea automată a textelor – Învățare – proces

□ Filtrarea

- Alegerea termenilor care să reprezinte documentul astfel încât să permită
 - descrierea conținutului documentului
 - diferențierea documentului de alte documente dintr-o colecție dată
- Înlăturarea celor mai frecvenți termeni (*stopwords*) – adverbe, prepoziții
 - găsiți într-o listă predefinită
 - a căror frecvență în toate documentele este mai mică de un anumit prag (5%)

Segmentat	Filtrat
interactive query expansion modifies queries using terms from a user automatic query expansion expands queries automatically	interactive query expansion modifies queries terms automatic query expansion expands queries automatically

Clasificarea automată a textelor – Învățare – proces

□ Aducerea la formă canonică

■ Lematizarea

- Analiză morfologică a termenilor pentru identificarea tuturor formelor de bază posibile
- Poate acționa asupra mai multor termeni
- Acționează în funcție de context
- Ex. "better" → "good"

■ Reducerea termenilor la rădăcină (*stemming*)

- Acționează asupra unui singur termen
- Ex. "computer", "computing", "compute" → "comput"
- Algoritmul de *stemming*
 - al lui Martin Porter
 - din WordNet

Filtrat	Redus
interactive query expansion modifies queries terms automatic query expansion expands queries automatically	interact queri expan modifi queri term automat queri expan expand queri automat

Clasificarea automată a textelor – Învățare – proces

□ Ponderarea

- Ponderarea termenilor conform unui anumit model
- Ponderi relative la
 - un singur document
 - frecvența termenilor (*term frequency* – *TF*)
 - o colecție de documente
 - frecvența inversă în document (*inverse document frequency* – *IDF*)
 - o combinație între *TF* și *IDF*
 - *TF* → cu cât un termen este mai frecvent într-un document, cu atât el este mai important pentru acel document
 - *IDF* → cu cât un termen apare în mai multe documente, cu atât el este mai puțin important în descrierea semanticii acelui document
- Frecvențele pot fi
 - Binare → prezența sau absența termenului
 - Reale $[0,1]$ → importanța termenului
- Fiind dat un set D de documente și un set T de termeni, ponderea p_{ij} a termenului t_i în documentul d_j ($i=1,2,\dots,|T|$, $j=1,2,\dots,|D|$) poate fi:
 - binară: $p_{ij} = 1$, dacă t_i apare în d_j
0, altfel
 - *TF*: $p_{ij}=tf_{ij}$ (nr. de apariții a termenului t_i în documentul d_j)
 - *TF.IDF*: $p_{ij}=tf_{ij}*\log_2(|D|/df_i)$, unde df_i =nr. de documente în care apare termenul t_i

Clasificarea automată a textelor – Învățare – proces

- Analiza documentelor de antrenament
 - Indexarea documentelor
 - Construirea unei reprezentări a documentelor → transformarea documentelor într-o formă interpretabilă de către clasificator
 - Obținerea unor concepte/termeni reprezentative(i) → attribute
 - Calcularea unor ponderi pt aceste attribute
 - **Reducerea dimensiunii (a numărului de concepte/attribute/termeni reprezentative(i) pentru document)**
 - **Selecția atributelor**
 - Extragerea atributelor
 - Învățarea unui model de clasificare
- Clasificarea noilor documente(de test)

Clasificarea automată a textelor – Învățare – proces

❑ Reducerea dimensiunii

■ Are drept scop

- ❑ Creșterea eficacității
- ❑ Reducerea timpului de învățare a modelului de clasificare
- ❑ Evitarea învățării pe derost a modelului de clasificare

■ Poate consta în

- ❑ Selecția atributelor (*feature selection*)
 - o submulțime a atributelor inițiale (originale)
- ❑ Extragerea atributelor
 - o mulțime de noi attribute determinate pe baza celor originale
→ proiecția unui vector R -dimensional într-unul r -dimensional ($r < R$)
 - noile attribute (mai puține) reprezintă o transformare a atributelor originale

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor

- Dându-se o mulțime de atribute $X_k = (x_{k1}, x_{k2}, \dots, x_{km})$ pentru un document $d_k \in D$, să se găsească o submulțime $X_k^p = (x_{k,i1}, x_{k,i2}, \dots, x_{k,ip})$, cu $p < m$ care să optimizeze o funcție obiectiv $J(X_k^m)$
 - Fc. obiectiv → eroarea de clasificare
- Selecția implică
 - O strategie de căutare pentru selecția submulțimilor candidat
 - căutare exhaustivă → toate submulțimile posibile → nefezabil
 - căutare strategică
 - prin ordonarea atributelor
 - pe baza unei metrici
 - și alegerea celor care depășesc un anumit prag
 - prin selectarea unei anumite submulțimi de atribute
 - se alege o submulțime optimă
 - O funcție obiectiv pentru evaluarea acestor submulțimi candidat
 - măsură a calității unei submulțimi de atribute
 - ajută selecția unei noi submulțimi candidat

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor

□ Metode

■ Nesupervizate

- Clusterizare
- Factorizarea matricilor

■ Supervizate

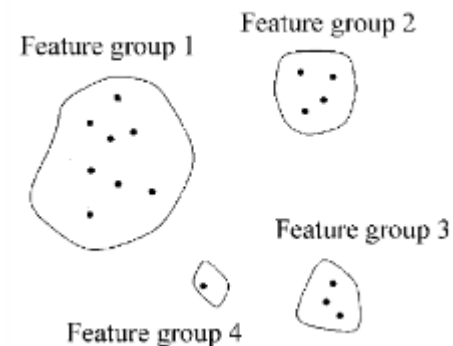
- Ordonarea atributelor
- Selectia unei submultimi de attribute

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor → Metode
Nesupervizate → Clusterizare

■ Se grupează attributele în clusteri

- K-means
- Hierarchical clustering



■ Se înlocuiesc (multe) attribute similare din același cluster cu
centrul clusterului

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor → Metode

Nesupervizate → factorizarea matricilor

- ❑ Analiza componentelor principale
- ❑ Descompunerea in valori singulare
- ❑ Factorizarea matricilor non-negative
- ❑ Isomap-uri

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor → Prin ordonarea atributelor

- Pp. că avem n date $(\mathbf{x}_k, y_k), k=1,2,\dots,n$
 - $\mathbf{x}_k \in \mathbf{R}^m \rightarrow \mathbf{x}_k = (x_{k1}, x_{k2}, \dots, x_{km})$
 - $y_k \in \mathbf{R}$
- Se calculează o funcție scor pentru fiecare pereche $S(i)=(x_{ki}, y_k)$
 - cu cât scorul este mai mare, cu atât variabila este mai importantă
- și se ordonează attributele în funcție de acest scor
- Notăție
 - $\mathbf{X}_i \in \mathbf{R}^n \rightarrow \mathbf{X}_i = (x_{1i}, x_{2i}, \dots, x_{ni})$
 - $\mathbf{Y} \in \mathbf{R}^n \rightarrow \mathbf{Y} = (y_1, y_2, \dots, y_n)$

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor → Prin ordonarea atributelor

□ Scoruri posibile

■ Coeficientul de corelație al lui Pearson

- $R(i) = \text{cov}(X_i, Y) / (\text{var}(X_i) \text{var}(Y))^{1/2}$
- $R(i) \approx \sum_{k=1, \dots, n} (x_{k,i} - X_i^a)(y_k - Y^a) / (\sum_{k=1, \dots, n} (x_{k,i} - X_i^a)^2 \sum_{k=1, \dots, n} (y_k - Y^a)^2)^{1/2}$
- $R^2(i) \rightarrow$ relație de dependență liniară între X_i și Y

■ Eroarea de clasificare

- Mai mulți clasificatori cu o singură variabilă
 - $(x_{k,i}, y_k), k=1, 2, \dots, n$
 - Se stabilește eroarea de clasificare pt fiecare $i=1, 2, \dots, n$
 - Se ordonează variabilele în funcție de eroare
 - Cu cât eroarea este mai mică cu atât variabila este mai importantă

■ Informația teoretică

- Informația mutuală între densitatea variabilei X_i și densitatea variabilei Y
- $I(i) = \int_x \int_y p(x_i, y) \log(p(x_i, y) / (p(x_i)p(y))) dx dy$
- $p(x)$ – probabilitatea densității lui $x \rightarrow$ greu de estimat

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor
→ Prin ordonarea atributelor

❑ Critici

- poate determina submulțimi de attribute redundante
- nu ține cont de corelarea atributelor
- un atribut nefolositor în izolație poate fi util în combinație cu alte attribute

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor → Prin alegerea unei submulțimi de attribute

□ Căutarea

- Căutare exhaustivă – toate submulțimile posibile → nefezabilă
- Căutare strategică – alegerea doar a unor submulțimi

□ Funcția obiectiv – tipuri

- *Wrapper*
- *Filter*
- *Embedded*

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor → Prin alegerea unei submulțimi de attribute

□ Funcția obiectiv – tipuri

■ *Wrapper*

- Funcția obiectiv este un clasificator care evaluează fiecare submulțime prin puterea ei predictivă
- Alegerea atributelor este **dependentă** de performanța clasificatorului (algoritmului de învățare)
- Algoritmul de învățare = cutie neagră pentru evaluarea submulțimii de attribute în funcție de puterea de învățare (clasificare) a acesteia

■ *Filter*

- Funcția obiectiv evaluează fiecare submulțime doar pe baza conținutului ei
- Alegerea atributelor este **independentă** de performanța clasificatorului
- Selecția atributelor este un pas anterior învățării

■ *Embedded*

- Alegerea atributelor are loc **în timpul** învățării

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor → Prin alegerea unei submulțimi de attribute → *Wrapper*

❑ Ideea de bază

- *Wrapper* → a înveli, a împacheta
- Funcția obiectiv este un clasificator care evaluează fiecare submulțime prin puterea ei predictivă
- Alegerea atributelor este **dependentă** de performanța clasificatorului (algoritmului de învățare)
- Algoritmul de învățare = cutie neagră pentru evaluarea submulțimii de attribute în funcție de puterea de învățare (clasificare) a acesteia

❑ Algoritm

- Se alege o metodă de clasificare (învățare)
- Se caută configurația optimă (submulțime de attribute și parametri ai clasificatorului)
 - ❑ Se alege o submulțime de attribute
 - ❑ Se repetă
 - Învățarea și optimizarea clasificatorului
 - cuantificarea performanței clasificatorului
 - alegerea unei noi submulțimi de attribute
 - ❑ până când se obține cea mai bună performanță în învățare

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor → Prin alegerea unei submulțimi de attribute → *Wrapper*

- ❑ Cum se alege o submulțime?
 - *best-first*
 - *branch-and-bound*
 - *simulated annealing*
 - algoritmi genetici
 - *greedy*
 - ❑ *Forward selection*
 - Variabilele sunt încorporate progresiv în submulțimi tot mai mari
 - ❑ *Backward selection*
 - Variabilele sunt eliminate progresiv din submulțime
- ❑ Cum se stabilește performanța algoritmului de învățare?
 - Validare
 - Validare-încrucișată
- ❑ Care algoritm de învățare să se folosească?
 - Arbori de decizie
 - Rețele neuronale
 - Mașini cu suport vectorial
 - Algoritmi evolutivi, etc

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor → Prin alegerea unei submulțimi de attribute → Filter

□ Ideea de bază

- Funcția obiectiv evaluează fiecare submulțime doar pe baza conținutului ei
- Alegerea atributelor este **independentă** de performanța clasificatorului
- Selecția atributelor este un pas anterior învățării

□ Evaluare

- Distanța sau măsura separabilității claselor
 - Ex. distanța (Euclideană, Hamming, etc) între clase
- Corelația și măsuri de informație teoretică
 - Submulțimile bune conțin attribute
 - puternic corelate cu ieșirea
 - ne-corelate între ele
 - Măsuri liniare
 - Coeficientul de corelație
 - Măsuri neliniare
 - Informația mutuală

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecția atributelor
→ Prin alegerea unei submulțimi de
attribute

- <http://jmlr.csail.mit.edu/papers/volume3/guyon03a/guyon03a.pdf>
- <http://jmlr.csail.mit.edu/proceedings/papers/v4/guerif08a/guerif08a.pdf>
- http://courses.cs.tamu.edu/rgutier/cs790_w02/l5.pdf

Clasificarea automată a textelor – Învățare – proces

- Analiza documentelor de antrenament
 - Indexarea documentelor
 - Construirea unei reprezentări a documentelor → transformarea documentelor într-o formă interpretabilă de către clasificator
 - Obținerea unor concepte/termeni reprezentative(i) → attribute
 - Calcularea unor ponderi pt aceste attribute
 - **Reducerea dimensiunii (a numărului de concepte/attribute/termeni reprezentative(i) pentru document)**
 - Selecția atributelor
 - **Extragerea atributelor**
 - Învățarea unui model de clasificare
- Clasificarea noilor documente(de test)

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Extragerea atributelor

□ Definire

- Determinarea unei noi mulțimi de atribute determinate pe baza celor originale → proiecția unui vector R -dimensional într-unul r -dimensional ($r < R$)
- Noile atribute (mai puține) reprezintă o transformare a atributelor originale

□ Dându-se o mulțime de atribute $X_k = (x_{k1}, x_{k2}, \dots, x_{km})$, să se găsească o transformare $z_k = g(x_k): R^m \rightarrow R^p$ cu $p < m$ astfel încât transformarea z_k să păstreze (cea mai parte din) informația atributelor inițiale

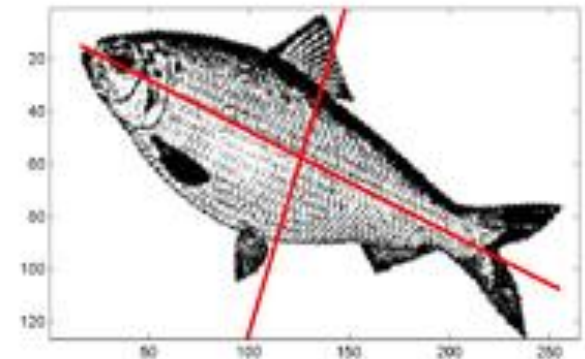
- Transformarea optimă – cea care nu determină creșterea probabilității de eroare
- Transformarea poate fi
 - Liniară $y = Wx$, $W \in M_{m,p}$
 - Ne-liniară – greu de determinat
- Transformarea este ghidată de o funcție obiectiv care trebuie optimizată (min/max)

□ Metode de extragere a atributelor în funcție de criteriul măsurat de funcția obiectiv:

- Reprezentare a semnalului → transformarea are drept scop reprezentarea datelor cu o acuratețe cât mai bună într-un spațiu mai redus
 - Analiza componentelor principale
- Clasificare → transformarea are drept scop evidențierea discriminării între clase într-un spațiu mai mic
 - Analiza discriminantului liniar

Clasificarea automată a textelor – Învățare – proces

- Metode de reducere a dimensiunii → Extragerea atributelor → Analiza componentelor principale
 - Scop
 - Transformarea unui set de variabile posibil corelate într-un set de variabile necorelate între ele (componente principale)
 - Prima componentă principală are cea mai mare varianță → cuantifică cea mai mare variabilitate posibilă a datelor
 - ACP determină axele care explică cel mai bine dispersia datelor (norul de puncte)
 - Descrierea datelor într-un spațiu dimensional mai mic
 - Alte denumiri
 - Transformarea Karhunen-Loève (teoria comunicațiilor)

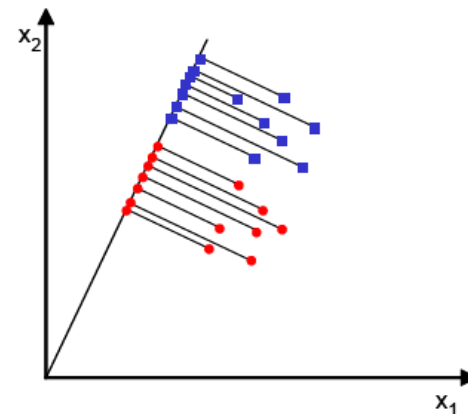
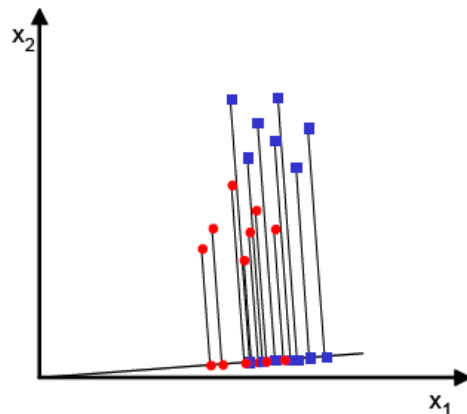


Clasificarea automată a textelor – Învățare – proces

- Metode de reducere a dimensiunii → Extragerea atributelor → Analiza componentelor principale
 - Tipologie
 - ACP liniară – date separabile liniar
 - ACP bazată pe kernele – date neseperabile liniar
 - Algoritm
 - Pp că avem un set de date $x_i, i=1,2,\dots,n$ cu m attribute ($x_i \in R^m \rightarrow x_i = (x_{i1}, x_{i2}, \dots, x_{im})$)
 - Scăderea mediei din fiecare dată (pe fiecare dimensiune) → centrarea datelor
 - $x'_{ij} = x_{ij} - x_j^a$, unde $x_j^a = (x_{1j} + x_{2j} + \dots + x_{nj})/n$
 - Calcularea matricii de covariație C
 - $C = (c_{ij}), i, j = 1, 2, \dots, m, c_{ij} = \text{cov}(x_i, x_j)$, unde $x_i = (x_{i1}, x_{i2}, \dots, x_{ni})$
 - $\text{cov}(X, Y) = \sum_{i=1,2,\dots,n} (X_i - X_a)(Y_i - Y_a)/(n-1)$
 - Determinarea vectorilor proprii \mathbf{v}_p și a valorilor proprii v_p (*eigenvector, eigenvalue*) corespunzătoare matricii de covariație $A \mathbf{v}_p = v_p \mathbf{v}_p$
 - Alegerea componentelor și formarea vectorului de caracteristici (attribute)
 - Se ordonează vectorii proprii descrescător după valorile proprii → attributele în ordinea importanței
 - Formarea vectorului de caracteristici cu acei vectori proprii care se doresc a fi reținuți
 - Derivarea noilor date
 - Se înmulțește vectorul de caracteristici cu vectorul datelor centrate

Clasificarea automată a textelor – Învățare – proces

- Metode de reducere a dimensiunii → Extragerea atributelor → Analiza discriminantului liniar
 - Scop
 - Determinarea unei combinații liniare de attribute care să separe datele (în clase) cât mai bine
 - Modelarea diferențelor între clase
 - Proiectarea datelor pe o linie/plan/hiperplan pentru a se observa o mai bună separabilitate a datelor → care este cea mai bună proiecție?
 - $y = w^T x$



Clasificarea automată a textelor – Învățare – proces

- Metode de reducere a dimensiunii → Extragerea atributelor → Analiza discriminantului liniar

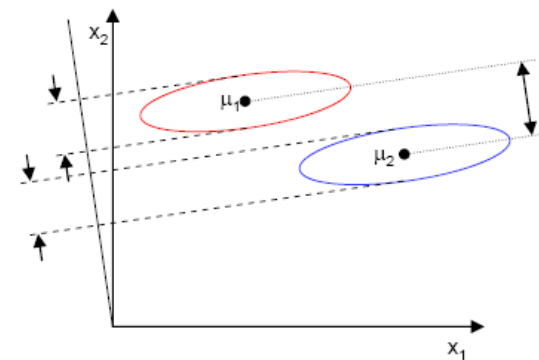
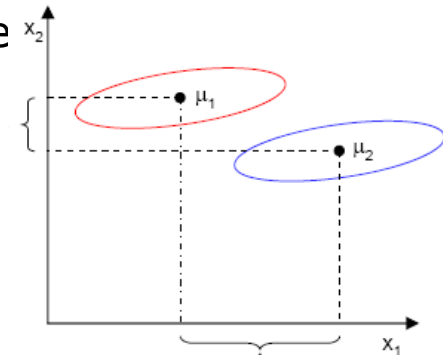
- Găsirea celei mai bune proiecții necesită definirea unei separare între proiecțiile datelor

- Distanța între proiecțiile mediilor corespunzătoare datelor din fiecare clasă

- Nu este foarte bine pentru că nu se ține cont de dispersia datelor în interiorul claselor

- Fisher → maximizarea raportului dintre diferența mediilor și împrăștierea în interiorul claselor

- o proiecție astfel încât:
 - exemplele din aceeași clasă sunt proiectate foarte aproape unele de altele
 - proiecțiile mediilor fiecărei clase sunt cât mai depărtate unele de altele

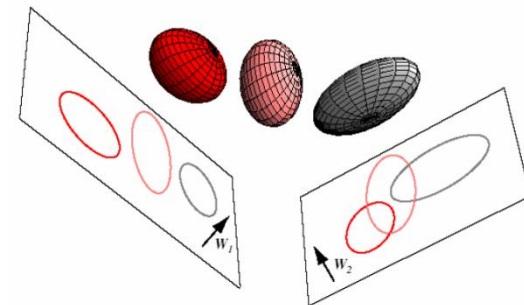
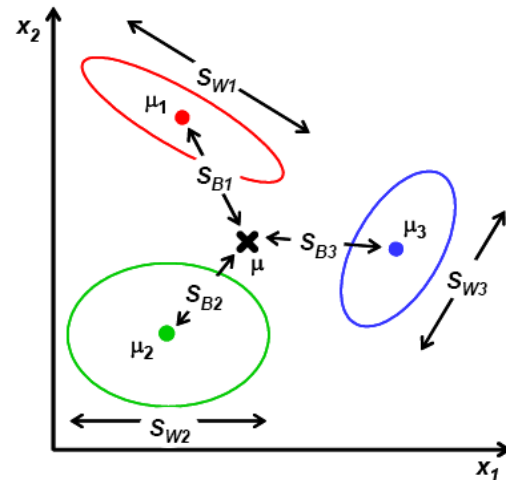


Clasificarea automată a textelor – Învățare – proces

Metode de reducere a dimensiunii → Extragerea a
Analiza discriminantului liniar

■ Algoritm

- Pp că:
 - există k clase,
 - μ_i – media instanțelor din clasa i , $i=1,2,\dots,k$
 - n – nr total de instanțe
 - n_i – nr de instanțe din clasa i , $i=1,2,\dots,k$
- Se caută $k-1$ vectori de proiecție
- Se calculează
 - Împrăștierea intra-clasă (scatter within class) S_w
 - $S_w = \sum_{i=1,2,\dots,k} \sum_{x \in \text{clasa } i} (x - \mu_i)(x - \mu_i)^T$
 - Împrăștierea între clase (scatter between classes) S_b
 - $S_b = \sum_{i=1,2,\dots,k} n_i (\mu_i - \mu)(\mu_i - \mu)^T$, unde $\mu = 1/n \sum_{x \in \text{clasa } i} x$
- Se maximizează
 - Raportul dintre
 - Pătratul diferenței mediilor (claselor) și
 - Împrăștierea intra-clasă
- Soluție
 - $w = S_w^{-1}(\mu_1 - \mu_2)$



- http://research.cs.tamu.edu/prism/lectures/pr/pr_l10.pdf
- <http://www.dtrek.com/lda.htm>
- http://www.music.mcgill.ca/~ich/classes/mumt611_05/classifiers/lda_theory.pdf

Clasificarea automată a textelor – Învățare – proces

- Analiza documentelor de antrenament
 - Indexarea documentelor
 - Construirea unei reprezentări a documentelor → transformarea documentelor într-o formă interpretabilă de către clasificator
 - Obținerea unor concepte/termeni reprezentative(i) → attribute
 - Calcularea unor ponderi pt aceste attribute
 - Reducerea dimensiunii (a numărului de concepte/attribute/termeni reprezentative(i) pentru document)
 - Selecția atributelor
 - Extragerea atributelor
 - **Învățarea unui model de clasificare**
- Clasificarea noilor documente(de test)
 - Indexarea documentelor
 - Utilizarea modelului de clasificare pentru stabilirea categoriilor fiecărui document de test

Clasificarea automată a textelor – Învățare – proces

- ❑ Învățarea unui model de clasificare
 - Alegerea unui algoritm de învățare
 - ❑ Arbori de decizie
 - ❑ Rețele neuronale artificiale
 - ❑ Mașini cu suport vectorial
 - ❑ Algoritmi evolutivi
 - ❑ Rețele Bayesiene
 - Fixarea/optimizarea parametrilor algoritmului
 - ❑ Cum se aleg parametrii?
 - Construirea modelului de clasificare și salvarea lui

Clasificarea automată a textelor – Învățare – proces

- Analiza documentelor de antrenament
 - Indexarea documentelor
 - Construirea unei reprezentări a documentelor → transformarea documentelor într-o formă interpretabilă de către clasificator
 - Obținerea unor concepte/termeni reprezentative(i) → attribute
 - Calcularea unor ponderi pt aceste attribute
 - Reducerea dimensiunii (a numărului de concepte/attribute/termeni reprezentative(i) pentru document)
 - Selecția atributelor
 - Extragerea atributelor
 - Învățarea unui model de clasificare
- **Clasificarea noilor documente(de test)**
 - Indexarea documentelor
 - Utilizarea modelului de clasificare pentru stabilirea categoriilor fiecărui document de test

Clasificarea automată a textelor – Învățare – proces

□ Metode de reducere a dimensiunii

■ Extragerea atributelor

- Analiza componentelor principale
- Analiza componentelor independente
- Scalare multidimensională
- Hărți topografice

- http://134.58.34.50/~marc/DM_course/slides_selection.pdf
- <http://www.esi.uem.es/~jmgomez/tutorials/eacl03/slides.pdf>