METODE INTELIGENTE DE REZOLVARE A PROBLEMELOR REALE

Laura Dioşan

Tema 4

Text mining

- Task-uri
  - Regăsirea informaţiei
  - Clasificarea automată a textelor

Text mining

- Task-uri
  - **Regăsirea informaţiei**
  - 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)*

- Definiţie
  - Regăsirea într-o colecţie de obiecte a unei submulţimi de obiecte care servesc unui anumit scop
    - Ex.
      - Pagini web pt pregătirea unei excursii
      - Materiale educaţionale pentru învăţarea unui concept

Regăsirea informației – tipologie

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

Regăsirea informației – proces

Paşi în procesul de regăsire

- Indexarea şi reprezentarea obiectelor din baza de cunoştinţe
- Formularea interogării
- Potrivirea interogării cu obiectele
- Selectarea rezultatelor

Regăsirea informației – proces

- Indexarea obiectelor din baza de cunoştinţe
  - fixarea unei anumite reprezentări a obiectelor
  - poate fi
    - manuală
    - automată
      - extragerea unor atribute (brute)
        - texte – separarea în cuvinte, eliminarea cuvintelor vide, etc
        - imagini – distribuţia culorilor şi a formelor
        - muzică – frecvenţa notelor

- Formularea interogării
  - Fixarea unei anumite reprezentări a interogării
  - Interogarea → un profil (şablon) 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 funcţii de similaritate sau de tip rang
  - Tipologie
    - potrivire perfectă (exactă)
    - potrivire parţială
- Selectarea rezultatelor
  - ordonarea lor
  - gruparea lor

Regăsirea informației - evaluare

Măsuri de performanţă

- Precizia
  - proporţia obiectelor regăsite care sunt relevante
  - nr. obiecte relevante regăsite / nr. obiecte regăsite
- Rapelul
  - proporţia obiectelor relevante care sunt regăsite
  - nr. obiecte relevante regăsite / nr. obiecte relevante
- Acurateţea
  - proporţia obiectelor corect regăsite
- Scorul F1
  - media armonică a preciziei şi rapelului

Text mining

- Task-uri
  - Regăsirea informaţiei
  - **Clasificarea automată a textelor**

Clasificarea automată a textelor

- Definire
- Direcţii în automatizare
  - Abordarea bazată pe învăţare
  - Abordarea bazată pe cunoştinţe

Clasificarea automată a textelor
Definire

- Categorizarea textelor
  - Atribuirea unor categorii (predefinite) documentelor
  - Documentele
    - rapoarte tehnice, pagini web, mesaje, cărţi
  - Categoriile
    - subiecte (artă, economie),
    - pertinenţe (mesaje spam, pagini web pt adulţi)

- Exemple de probleme

	Cuvinte	Documente
Învăţare supervizată	Etichetarea părţilor de vorbire	Clasificarea textelor, Filtrarea, Detectarea subiectelor
Învăţare nesupervizată	Indexarea semantică, construcţia automată a tezaurelor, extragerea cuvintelor cheie	Clusterizarea documentelor, Detectarea subiectelor

Clasificarea automată a textelor
Direcţii în automatizare

- Abordarea bazată pe învăţare
  - Experţii etichetează o parte din exemple
  - Algoritmul etichetează noi exemple
  - Învăţarea poate fi:
    - supervizată
    - nesupervizată
- Abordarea bazată pe cunoştinţe
  - Cunoştinţele despre clasificare sunt
    - obţinute de la experţi
    - codificate sub formă de reguli

Clasificarea automată a textelor – Învățare – definire

- Definirea problemei
  - Se dă un set de documente D, |D|=N+n şi 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 funcţie necunoscută de clasificare

    $\Phi: DxC \rightarrow \{true, false\}$

    definită astfel:
    - $\Phi(d,c)=$ *true*, dacă $d \in c$

      false, altfel

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

Clasificarea automată a textelor – Învățare – 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
      - Obţinerea unor concepte/termeni reprezentative(i) → atribute
      - Calcularea unor ponderi pt aceste atribute
    - Reducerea dimensiunii (a numărului de concepte/atribute/termeni reprezentative(i) pentru document)
      - Selecţia atributelor
      - Extragerea atributelor
  - Învăţarea 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
      - Obţinerea unor concepte/termeni reprezentative(i) → atribute
      - Calcularea unor ponderi pt aceste atribute
    - Reducerea dimensiunii (a numărului de concepte/atribute/termeni reprezentative(i) pentru document)
      - Selecţia atributelor
      - Extragerea atributelor
  - Învăţarea 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)
  - Obţinerea unor concepte/termeni reprezentative(i) → atribute şi calcularea unor ponderi pt aceste atribute
  - 4 paşi:
    - 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 (atribute)
    - 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)
      - Transforamrea tuturor literelor în litere mici
      - Înlăturarea semnelor de punctuaţie

Iniţial	Liniarizat
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 conţinutului documentului
    - diferenţierea documentului de alte documente dintr-o colecţie dată
  - Înlăturarea celor mai frecvenţi termeni (*stopwords*) – adverbe, prepoziţii
    - găsiţi într-o listă predefinită
    - a căror frecvenţă 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 acţiona asupra mai multor termeni
    - Acţionează în funcţie de context
    - Ex. “better” → “good”
  - Reducerea termenilor la rădăcină (*stemming*)
    - Acţionează 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
      - frecvenţa termenilor (*term frequency – TF*)
    - o colecţie de documente
      - frecvenţa inversă în document (*inverse document frequency – IDF*)
    - o combinaţie între *TF* şi *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 puţin important în descrierea semanticii acelui document
  - Frecvenţele pot fi
    - Binare → prezenţa sau absenţa termenului
    - Reale ([0,1]) → importanţa termenului
  - Fiind dat un set D de documente şi un set T de termeni, ponderea p_{ij} a termenului t_i în documentul d_j (i=1,2,...,|T|, j=1,2,...,|D|) poate fi:
    - binară: p_{ij} = 1, dacă t_i apare în d_j
      0, altfel
    - *TF*: $p_{ij}=tf_{ij}$ (nr. de apariţii 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
      - Obţinerea unor concepte/termeni reprezentative(i) → atribute
      - Calcularea unor ponderi pt aceste atribute
    - **Reducerea dimensiunii (a numărului de concepte / atribute / termeni reprezentative(i) pentru document)**
      - **Selecţia atributelor**
      - Extragerea atributelor
  - Învăţarea unui model de clasificare
- Clasificarea noilor documente(de test)

Clasificarea automată a textelor – Învățare – proces

- Reducerea dimensiunii
  - Are drept scop
    - Creşterea eficacităţii
    - Reducerea timpului de învăţare a modelului de clasificare
    - Evitarea învăţării pe derost a modelului de clasificare
  - Poate consta în
    - Selecţia atributelor (*feature selection*)
      - o submulţime a atributelor iniţiale (originale)
    - Extragerea atributelor
      - o mulţime de noi atribute determinate pe baza celor originale → proiecţia unui vector R-dimensional într-unul r-dimensional ($r < R$)
      - noile atribute (mai puţine) reprezintă o transformare a atributelor originale

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecţia atributelor

- Dându-se o mulţime de atribute $X_k=(x_{k1}, x_{K2},\ldots,x_{km})$ pentru un document $d_k \epsilon D$, să se găsească o submulţime $X_K^p=(x_{K,j1}, x_{K,i2},\ldots,x_{K,ip})$, cu $p < m$ care să optimizeze o funcţie obiectiv $J(X_K^m)$
  - Fc. obiectiv → eroarea de clasificare

- Selecţia implică
  - O strategie de căutare pentru selecţia submulţimilor candidat
    - căutare exhaustivă → toate submulţimile posibile →nefezabil
    - căutare strategică
      - prin ordonarea atributelor
        - pe baza unei metrici
        - şi alegerea celor care depăşesc un anumit prag
      - prin selectarea unei anumite submulţimi de atribute
        - se alege o submulţime optimală
  - O funcţie obiectiv pentru evaluarea acestor submulţimi candidat
    - măsură a calităţii unei submulţimi de atribute
    - ajută selecţia unei noi submulţimi candidat

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecţia atributelor → Prin ordonarea atributelor

- Pp. că avem n date $(\boldsymbol{x}_k, y_k)$, $k=1,2,\ldots,n$
  - $\boldsymbol{x}_k \in \boldsymbol{R}^m \rightarrow \boldsymbol{x}_k = (x_{k1}, x_{k2}, \ldots, x_{km})$
  - $y_k \in \boldsymbol{R}$

- Se calculează o funcţie scor pentru fiecare pereche $S(i)=(x_{ki},y_k)$
  - cu cât scorul este mai mare, cu atât variabila este mai importantă
- şi se ordonează atributele în funcţie de acest scor

- Notaţie

  $X_i \in \boldsymbol{R}^n \rightarrow X_i=(x_{1i}, x_{2i}, \ldots, x_{ni})$

  $Y \in \boldsymbol{R}^n \rightarrow Y=(y_1, y_2, \ldots, y_n)$

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecţia atributelor → Prin ordonarea atributelor

- Scoruri posibile
  - Coeficientul de corelaţie al lui Pearson
    - $R(i)=cov(X_i, Y)/(var(X_i)var(Y))^{1/2}$
    - $R(i)\approx\sum_{k=1,\ldots,n}(x_{k,i}- X_i^a)(y_k-Y^a)/(\sum_{k=1,\ldots,n}(x_{k,i}- X_i^a)^2\sum_{k=1,\ldots,n}(y_k-Y^a)^2)^{1/2}$
    - $R^2(i)$ → relaţie de dependenţă liniară între X_i şi Y
  - Eroarea de clasificare
    - Mai mulţi clasificatori cu o singură variabilă
      - (x_{ki}, y_k), $k=1,2,\ldots,n$
      - Se stabileşte eroarea de clasificare pt fiecare $i=1,2,\ldots,n$
      - Se ordonează variabilele în funcţie de eroare
      - Cu cât eroarea este mai mică cu atât variabila este mai importantă
  - Informaţia teoretică
    - Informaţia mutuală între densitatea variabilei X_i şi densitatatea variabilei Y
    - $I(i)=\int_x\int_y p(x_i,y)log(p(x_i,y)/(p(x_i)p(y)))dxdy$
    - $p(x)$ – probabilitatea densităţii lui x → greu de estimat

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecţia atributelor
→ Prin ordonarea atributelor

- Critici
  - poate determina submulţimi de atribute redundante
  - nu ţine cont de corelarea atributelor
  - un atribut nefolositor în izolaţie poate fi util în combinaţie cu alte atribute

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecţia atributelor → Prin alegerea unei submulţimi de atribute

- Căutarea
  - Căutare exhaustivă – toate submulţimile posibile → nefezabilă
  - Căutare strategică – alegerea doar a unor submulţimi
- Funcţia obiectiv – tipuri
  - *Wrapper*
  - *Filter*
  - *Embedded*

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecţia atributelor → Prin alegerea unei submulţimi de atribute

- Funcţia obiectiv – tipuri
  - *Wrapper*
    - Funcţia obiectiv este un clasificator care evaluează fiecare submulţime prin puterea ei predictivă
    - Alegerea atributelor este **dependentă** de performanţa clasificatorului (algoritmului de învăţare)
    - Algoritmul de învăţare = cutie neagră pentru evaluarea submulţimii de atribute în funcţie de puterea de învăţare (clasificare) a acesteia
  - *Filter*
    - Funcţia obiectiv evaluează fiecare submulţime doar pe baza conţinutului ei
    - Alegerea atributelor este **independentă** de performanţa clasificatorului
    - Selecţia atributelor este un pas anterior învăţării
  - *Embedded*
    - Alegerea atributelor are loc **în timpul** învăţării

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecţia atributelor → Prin alegerea unei submulţimi de atribute → *Wrapper*

- Ideea de bază
  - *Wrapper* → a înveli, a împacheta
  - Funcţia obiectiv este un clasificator care evaluează fiecare submulţime prin puterea ei predictivă
  - Alegerea atributelor este **dependentă** de performanţa clasificatorului (algoritmului de învăţare)
  - Algoritmul de învăţare = cutie neagră pentru evaluarea submulţimii de atribute în funcţie de puterea de învăţare (clasificare) a acesteia

- Algoritm
  - Se alege o metodă de clasificare (învăţare)
  - Se caută configuraţia optimă (submuţime de atribute şi parametri ai clasificatorului)
    - Se alege o submulţime de atribute
    - Se repetă
      - Învăţarea şi optimizarea clasificatorului
      - cuantificarea performanţei clasificatorului
      - alegerea unei noi submulţimi de atribute
    - până când se obţine cea mai bună performanţă în învăţare

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecţia atributelor → Prin alegerea unei submulţimi de atribute → *Wrapper*

- Cum se alege o submulţime?
  - *best-first*
  - *branch-and-bound*
  - *simulated annealing*
  - algoritmi genetici
  - *greedy*
    - *Forward selection*
      - Variabilele sunt încorporate progresiv în submuţimi tot mai mari
    - *Backward selection*
      - Variabilele sunt eliminate progresiv din submulţime
- Cum se stabileşte performanţa algoritmului de învăţare?
  - Validare
  - Validare-încrucişată
- Care algoritm de învăţare să se folosească?
  - Arbori de decizie
  - Reţele neuronale
  - Maşini cu suport vectorial
  - Algoritmi evolutivi, etc

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecţia atributelor → Prin alegerea unei submulţimi de atribute → Filter

- Ideea de bază
  - Funcţia obiectiv evaluează fiecare submulţime doar pe baza conţinutului ei
  - Alegerea atributelor este **independentă** de performanţa clasificatorului
  - Selecţia atributelor este un pas anterior învăţarii
- Evaluare
  - Distanţa sau măsura separabilităţii claselor
    - Ex. distanţa (Euclideană, Hamming, etc) între clase
  - Corelaţia şi măsuri de informaţie teoretică
    - Submulţimile bune conţin atribute
      - puternic corelate cu ieşirea
      - ne-corelate între ele
    - Măsuri liniare
      - Coeficientul de corelaţie
    - Măsuri neliniare
      - Informaţia mutuală

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Selecţia atributelor
→ Prin alegerea unei submulţimi de atribute

- 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
      - Obţinerea unor concepte/termeni reprezentative(i) → atribute
      - Calcularea unor ponderi pt aceste atribute
    - **Reducerea dimensiunii (a numărului de concepte / atribute / termeni reprezentative(i) pentru document)**
      - Selecţia atributelor
      - **Extragerea atributelor**
  - Învăţarea unui model de clasificare
- Clasificarea noilor documente(de test)

Clasificarea automată a textelor – Învățare – proces

Reducerea dimensiunii → Extragerea atributelor

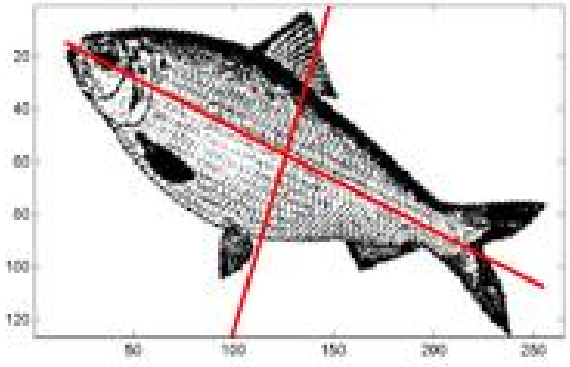
- Definire
  - Determinarea unei noi mulţimi de atribute determinate pe baza celor originale → proiecţia unui vector R-dimensional într-unul r-dimensional ($r < R$)
  - Noile atribute (mai puţine) reprezintă o transformare a atributelor originale

- Dându-se o mulţime de atribute $X_k=(x_{k1}, x_{k2},...,x_{km})$, 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) informaţia atributelor iniţiale
  - Transformarea optimă – cea care nu determină creşterea probabilităţii de eroare
  - Transformarea poate fi
    - Liniară $y = Wx$, $W \in M_{m,p}$
    - Ne-liniară – greu de determinat
  - Transformarea este ghidată de o funcţie obiectiv care trebuie optimizată (min/max)

- Metode de extragere a atributelor în funcţie de criteriul măsurat de funcţia obiectiv:
  - Reprezentare a semnalului → transformarea are drept scop reprezentarea datelor cu o acurateţe cât mai bună într-un spaţiu mai redus
    - Analiza componentelor principale
  - Clasificare → transformarea are drept scop evidenţierea discriminării între clase într-un spaţiu 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 varianţă → 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 spaţiu dimensional mai mic
  - Alte denumiri
    - Transformarea Karhunen-Loève (teoria comunicaţiilor)

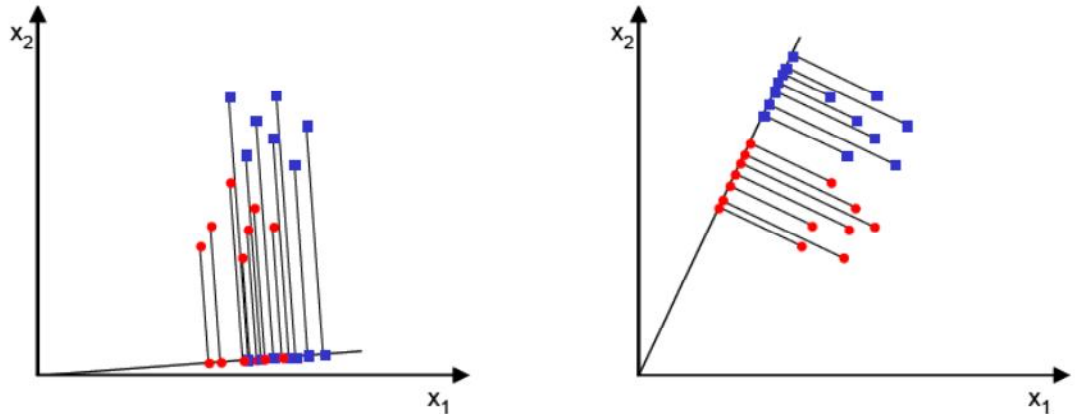


Clasificarea automată a textelor – Învățare – proces

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

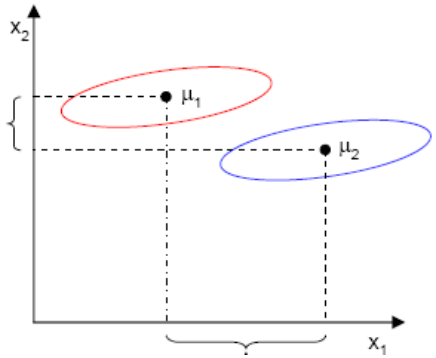
Clasificarea automată a textelor – Învățare – proces

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

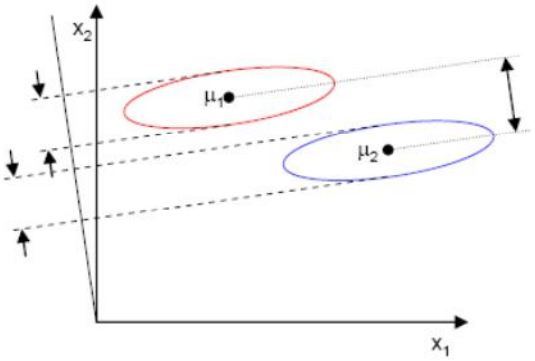


Clasificarea automată a textelor – Învățare – proces

- Metode de reducere a dimensiunii → Extragerea atributelor → Analiza discriminantului liniar
  - Găsirea celei mai bune proiecţii necesită definirea unei măsuri de separare între proiecţiile datelor
    - Distanţa între proiecţiile mediilor corespunzătoare datelor din fiecare clasă
      - Nu este foarte bine pentru că nu se ţine cont de dispersia datelor în interiorul claselor



    - Fisher → maximizarea raportului dintre diferenţa mediilor şi împrăştierea în interiorul claselor
      - → o proiecţie astfel încât:
        - exemplele din aceeaşi clasă sunt proiectate foarte aproape unele de altele
        - proiecţiile mediilor fiecărei clase sunt cât mai depărtate unele de altele



Clasificarea automată a textelor – Învățare – proces

- Metode de reducere a dimensiunii → Extragerea atributelor → Analiza discriminantului liniar
  - Algoritm
    - Pp. că:
      - există 2 clase,
      - μ_i – media instanţelor din clasa i, $i=1,2$
      - n – nr. total de instanţe
      - n_i – nr. de instanţe din clasa i
    - Se calculează
      - împrăştierea intra-clasă (*scatter within class*) S_w
        - $S_w=\sum_{i=1,2}\sum_{x\epsilon clasai}(x-\mu_i)(x-\mu_i)^T$
      - împrăştierea între clase (*scatter between classes*) S_b
        - $S_b=\sum_{i=1,2}n_i(\mu_i-\mu)(\mu_i-\mu)^T$, unde $\mu=1/n\sum_{x\epsilon clasai}n_i\mu_i$
    - Se maximizează
      - raportul dintre
        - pătratul diferenţei mediilor (claselor) şi
        - împrăştierea intra-clasă
    - Soluţie
      - $w=S^{-1}_w(\mu_1-\mu_2)$

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
      - Obţinerea unor concepte/termeni reprezentative(i) → atribute
      - Calcularea unor ponderi pt aceste atribute
    - Reducerea dimensiunii (a numărului de concepte/atribute/termeni reprezentative(i) pentru document)
      - Selecţia atributelor
      - Extragerea atributelor
  - **Învăţarea 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

- Învăţarea unui model de clasificare
  - Alegerea unui algoritm de învăţare
    - Arbori de decizie
    - Reţele neuronale artificiale
    - Maşini cu suport vectorial
    - Algoritmi evolutivi
    - Reţele Bayesiene
  - Fixarea/optimizarea parametrilor algoritmului
    - Cum se aleg parametrii?
  - Construirea modelului de clasificare şi salvarea lui

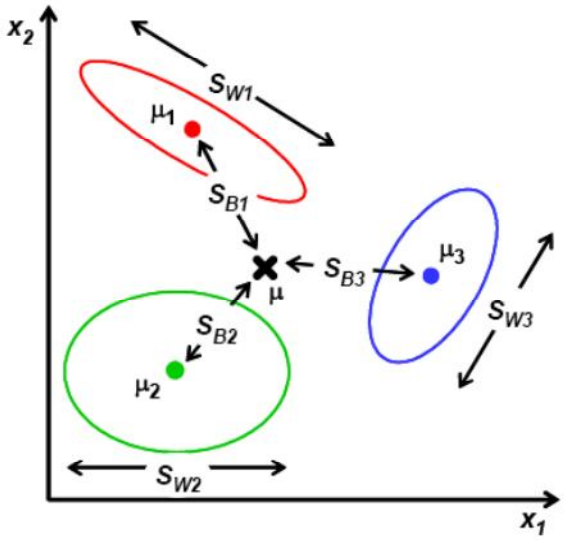
Clasificarea automată a textelor – Învățare – proces

- Analiza documentelor de antrenament
  - Indexarea documentelor
    - Construirea unei reprezentări a documentelor → transformarea documentelor într-o formă interpretabilă de către clasificator
      - Obţinerea unor concepte/termeni reprezentative(i) → atribute
      - Calcularea unor ponderi pt aceste atribute
    - Reducerea dimensiunii (a numărului de concepte/atribute/termeni reprezentative(i) pentru document)
      - Selecţia atributelor
      - Extragerea atributelor
  - Învăţarea unui model de clasificare

- **Clasificarea noilor documente(de test)**
  - Indexarea documentelor
  - Utilizarea modelului de clasificare pentru stabilirea categoriilor fiecărui document de test

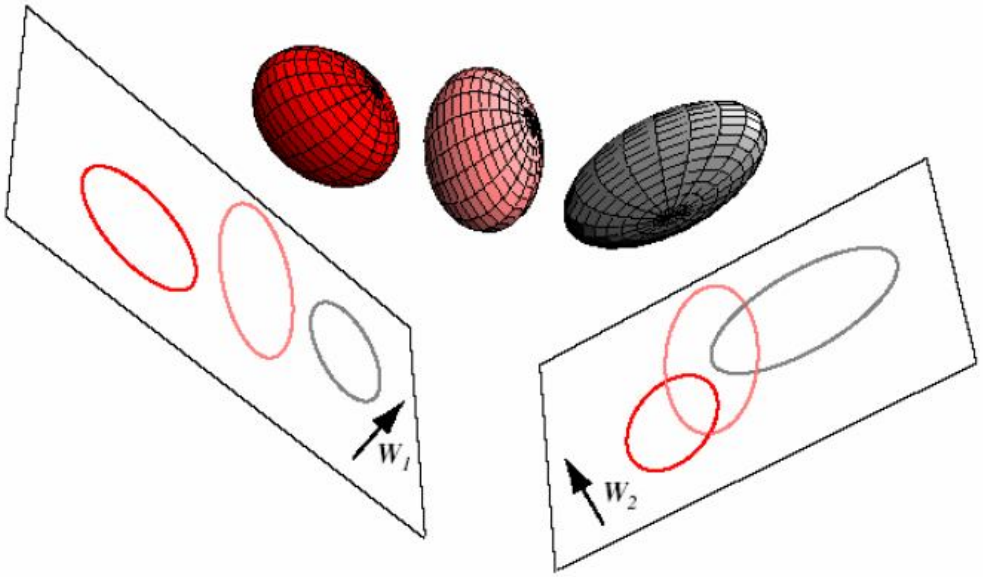
Clasificarea automată a textelor – Învățare – proces

- Metode de reducere a dimensiunii → Extragerea atributelor → Analiza discriminantului liniar
  - Algoritm
    - Pp că:
      - există *k* clase,
      - μ_i – media instanţelor din clasa *i*, *i=1,2,...,k*
      - n – nr total de instanţe
      - n_i – nr de instanţe din clasa *i*, *i=1,2,...,k*
    - *Se caută k-1 vectori de proiecţie*
    - Se calculează
      - Împrăştierea intra-clasă (scatter within class) S_w
        - $S_w=\sum_{i=1,2,...,k}\sum_{x\in clasai}(x-\mu_i)(x-\mu_i)^T$
      - Împrăştierea între clase (scatter between classes) S_b
        - $S_b=\sum_{i=1,2,...,k} n_i(\mu_i-\mu)(\mu_i-\mu)^T$, unde $\mu=1/n\sum_{x\in clasai} n_i\mu_i$
    - Se maximizează
      - Raportul dintre
        - Pătratul diferenței mediilor
        - clasă



sm/lec

h/class



Clasificarea automată a textelor – Învățare – proces

- Metode de reducere a dimensiunii
  - Extragerea atributelor
    - Analiza componentelor principale
    - Analiza componentelor independente
    - Scalare multidimensională
    - Hărţi topografice

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