LAPI – Laboratorul de Analiza şi Prelucrarea Imaginilor

Facultatea de Electronică, Telecomunicaţii şi Tehnologia Informaţiei

Universitatea POLITEHNICA din Bucureşti

Tehnici de analiză și clasificare automată a informației

Conf. dr. ing. Bogdan IONESCU
http://imag.pub.ro/~bionescu

București, 2015

Organizare disciplină

Conf. Bogdan Ionescu
Titular disciplină
http://imag.pub.ro/~bionescu/

Dr. Ionuț Mironică
Titular laborator
http://imag.pub.ro/~imironica/

Materiale curs:
http://imag.pub.ro/~bionescu/index_files/Page328.htm

Laborator: B135

Punctaj:
- examen (scris) 50%;
- colocviu laborator (practic) 50%.

Plan Curs

M1. Introducere (concept, aplicații)

M2. Prelucrarea și reprezentarea datelor de intrare

M3. Tehnici de clasificare ne-supervizată (“clustering”)

M4. Tehnici de clasificare supervizată (“classification”)

M5. Evaluarea performanței clasificatorilor

Bibliografie

[1] Curs;

[2] B. Ionescu, “*Analiza și Prelucrarea Secvențelor Video: Indexarea Automată după Conținut*”, Editura Tehnică București, 2009;

[3] B. Ionescu, I. Mironică, “*Conceptul de Indexare Automată după Conținut în Contextul Datelor Multimedia*”, Editura MartrixRom, 2013;

[4] I.H. Witten, E. Frank, M.A. Hall, “*Data Mining: Practical Machine Learning Tools and Techniques*”, Morgan Kaufmann Publishers, 2011;

[5] A.K. Jain, M.N. Murty, P.J. Flynn, “*Data Clustering: A Review*”, ACM Computing Surveys, 31(3), 1999.

> M1. Introducere

1.1. [Introducere]

1.2. [Aplicații concrete]

1.3. [Conceptul de învățare]

1.4. [Terminologie]

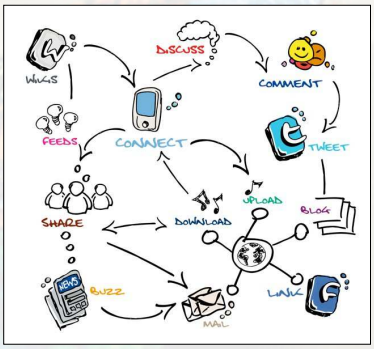
1.5. [Tehnici existente]

1.6. [Utilitare]

Informație

> volum imens de date (**Big Data**) care nu mai poate fi “gestionat” (vizualizat, analizat, înțeles, prelucrat) de către utilizator;

+ complexitatea datelor depășește de multe ori puterea de înțelegere și de procesare umană.



Exemplu date multimedia

Informație (cont.)

Exemplu date multimedia:

- >100 de ore video sunt încărcate în fiecare minut pe platforma YouTube;

- >600 de ani video de pe platforma YouTube sunt vizualizați zilnic pe platforma Facebook;

- >900 de secvențe video de pe platforma YouTube sunt partajate în fiecare minut pe platforma Twitter;

- în 2015 se estimează ca > 1 milion de minute video (674 de zile) vor tranzita Internet-ul în fiecare secundă!

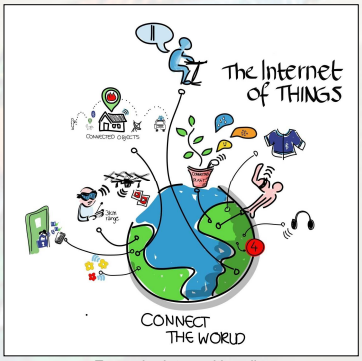
- bazele de date ale lumii se dublează la fiecare 20 de luni.

[date din 2014]

Informație (cont.)

[sursă imagine Wikipedia]

> în mod constant, apar noi "generatori" de informație, ex. dispozitive purtabile, Internet-ul lucrurilor, roboți etc.

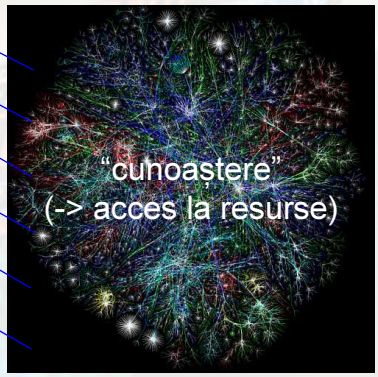


Exemplu date multimedia

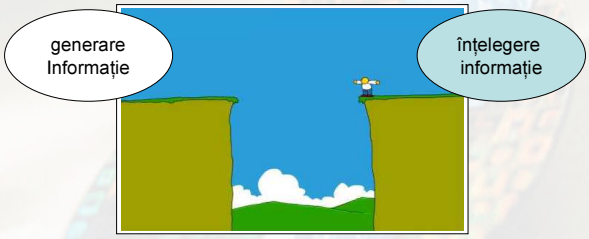
Internet (de facto)

[sursă imagine Patrick Barry, Flickr]

- știri
- rețele sociale
- divertisment
- securitate
- date financiare
- date medicale
- …



Punerea problemei



> "**bridge the gap**", înțelegere și descoperire informație ascunsă (șabloane) ce poate fi utilă și care nu este exploatată;

[o problemă veche de când lumea: vânătorii încercau să înțeleagă comportamentul de migrare al animalelor, fermierii încercau să înțeleagă comportamentul culturilor, etc]

Obiective clasificare informație

> **reducerea volumului informațional**:
- regruparea unui ansamblu de date în grupuri omogene și determinarea informație reprezentative;
- eliminarea redundanței informaționale.

> **punerea în evidență a "cunoașterii"**:
- localizarea într-un volum mare de date a unor grupuri de informații ce prezintă anumite caracteristici de interes;
- o nouă înțelegere a relațiilor existente între date.

> **punerea în evidență a datelor atipice**:
- localizarea datelor ce nu corespund criteriilor considerate, în particular interesante prin natura acestora.

> **rezolvarea unor probleme de calcul.**

Un exemplu, "weather problem"

[I.H. Witten, E. Frank, M.A. Hall, "*Data Mining: Practical Machine Learning Tools and Techniques*"]

Date vreme

nr.	*vreme*	*temperatură*	*umiditate*	*vânt*	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	ridicată	da	Nu
3	înnorat	cald	ridicată	da	Da
4	ploioasă	cald	normală	nu	Da
5	ploioasă	rece	normală	nu	Da

> dispunem de patru atribute măsurabile: *vreme* (3), *temperatură* (3), *umiditate* (2) și *vânt* (2) și trebuie să determinăm automat dacă putem practica o activitate;

> conform datelor avem 3 x 3 x 2 x 2 = 36 de combinații posibile, din care dispunem doar de 5 seturi de date.

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "*Data Mining: Practical Machine Learning Tools and Techniques*"]

Date vreme

nr.	*vreme*	*temperatură*	*umiditate*	*vânt*	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	ridicată	da	Nu
3	înnorat	cald	ridicată	da	Da
4	ploioasă	rece	normală	da	Nu
5	ploioasă	rece	normală	nu	Da

Dacă (vreme==însorită) && (umiditate==ridicată) -> sport=Nu;
Dacă (vreme==ploioasă) && (vânt==da) -> sport=Nu;
Dacă (vreme==înnorat) -> sport=Da;
Dacă (umiditate==normală) -> sport=Da;
altfel -> sport=Da;

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "*Data Mining: Practical Machine Learning Tools and Techniques*"]

Dacă (vreme==însorită) && (umiditate==ridicată) -> sport=Nu;
Dacă (vreme==ploioasă) && (vânt==da) -> sport=Nu;
Dacă (vreme==înnorat) -> sport=Da;
Dacă (umiditate==normală) -> sport=Da;
altfel -> sport=Da;

> analizate în ordine clasifică corect toate exemplele din tabel?

> scoase din context nu mai sunt adevărate ceea ce înseamnă că un set de reguli depinde de modul în care este interpretat!

Dacă (umiditate==normală) -> sport=Da;

nr.	*vreme*	*temperatură*	*umiditate*	*vânt*	sport
4	ploioasă	rece	normală	da	Nu
5	ploioasă	rece	normală	nu	Da

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "*Data Mining: Practical Machine Learning Tools and Techniques*"]

Date vreme

nr.	*vreme*	*temperatură*	*umiditate*	*vânt*	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	ridicată	da	Nu
3	înnorat	cald	ridicată	da	Da
4	ploioasă	rece	normală	da	Nu
5	ploioasă	rece	normală	nu	Da

> se poate merge mai departe, pe baza datelor să determinăm reguli de asociere care corelează atributele:

Dacă (temperatură==rece) -> umiditate=normală;
Dacă (umiditate==normală) && (vânt==nu) -> sport=da;
Dacă (vreme==însorită) && (sport==nu) -> umiditate=ridicată;

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "*Data Mining: Practical Machine Learning Tools and Techniques*"]

Date vreme

nr.	*vreme*	*temperatură*	*umiditate*	*vânt*	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	ridicată	da	Nu
3	înnorat	cald	ridicată	da	Da
4	ploioasă	rece	normală	da	Nu
5	ploioasă	rece	normală	nu	Da

(1) **date incomplete**: sistemul trebuie să fie capabil de generalizare pentru exemple noi, exemplu folosind cele 5 să putem prezice restul de 31 de situații?

nr.	*vreme*	*temperatură*	*umiditate*	*vânt*	sport
6	însorită	cald	normală	nu	?

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "*Data Mining: Practical Machine Learning Tools and Techniques*"]

Date vreme

nr.	*vreme*	*temperatură*	*umiditate*	*vânt*	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	-	da	Nu
3	înnorat	-	ridicată	da	Da
4	ploioasă	-	-	da	Nu
5	ploioasă	rece	normală	nu	Da

(2) **date incomplete**: datele de intrare pot fi incomplete, sistemul trebuie să fie capabil de generalizare și în acestă situație:

nr.	*vreme*	*temperatură*	*umiditate*	*vânt*	sport
6	însorită	cald	normală	nu	?

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "*Data Mining: Practical Machine Learning Tools and Techniques*"]

Date vreme

nr.	*vreme*	*temperatură*	*umiditate*	*vânt*	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	0	da	Nu
3	înnorat	%&#	ridicată	da	Da
4	ploioasă	&@##	0	da	Nu
5	ploioasă	rece	normală	nu	Da

(3) **date eronate**: regulile clasifică corect exemplele dar datorită erorilor datelor (ex. zgomot) în realitate clasificatorul nu este capabil să clasifice corect chiar datele pe baza cărora a fost definit.

Aplicații concrete

> accentul se pune pe abilitatea de a generaliza pe **date noi** despre care nu avem nici o informație a priori;

- **Motoarele de căutare de pe Internet**:

  - învață folosind datele de la utilizator, ex. cuvintele folosite la căutare și gradul de satisfacție al utilizatorului;

  - învață care pagini sunt mai relevante, ex. PageRank de la Google care definește "prestigiul" unei pagini în funcție de cât de corelată este cu alte pagini care la rândul lor sunt "prestigioase" etc;

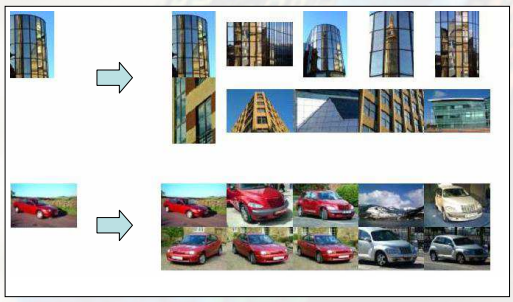
  - învață din istoricul de navigare pentru a recomanda produse și reclame, ex. platforme de comercializare cărți, filme, rețele sociale, etc.

[I.H. Witten, E. Frank, M.A. Hall, "*Data Mining: Practical Machine Learning Tools and Techniques*"]

Aplicații concrete (cont.)

- **Motoarele de căutare de pe Internet (cont.)**:

  > căutare după conținut date multimedia (audio, imagini, video);



Aplicații concrete (cont.)

- **Sisteme decizionale**: ex. acordarea unui credit

  - *procedura tradițională*: solicitantul furnizează o serie de date personale și financiare iar creditorul folosește metode statistice și decizia umana pentru "accept" sau "refuz";

    - în X% din cazuri, datele nu se încadrează clar în cele două clase iar decizia este luată de un operator uman ("**bordeline cases**");

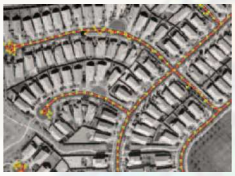
    - în ~½ din cele X% cazuri, decizia se dovedește greșită solicitantul creditat eșuând să returneze creditul -> problemă $;

  - *folosind clasificare*: predicție comportament al celor X% cazuri limită prin antrenare folosind date etichetate: cazuri limită din trecut pentru care se știe rezultatul -> *îmbunătățire decizie*.

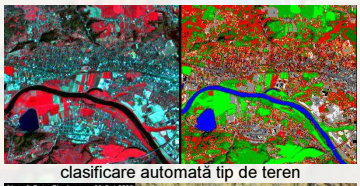
Aplicații concrete (cont.)

- **Analiză imagistică**:

  - *satelitară*;



detectare automată străzi



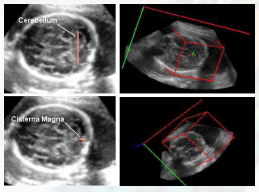
clasificare automată tip de teren

detecție automată incendii

Aplicații concrete (cont.)

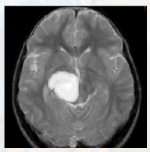
- **Analiză imagistică (cont.)**:

  - *medicală*;

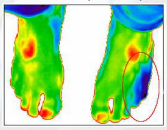


analiză automată (ultrasunete)

detecție automată fractură (raze X)



detecție tumori

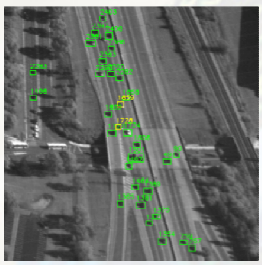


detecție automată probleme circulație (termografie)

Aplicații concrete (cont.)

- **Analiză imagistică (cont.)**:

  - *securitate*;



monitorizare automată trafic

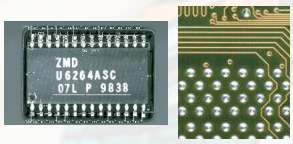
detecție automată persoane

detecție automată acțiuni

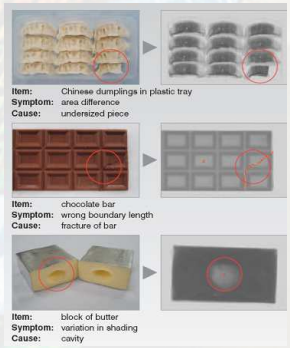
Aplicații concrete (cont.)

- **Analiză imagistică (cont.)**:

  - *automatizări industriale*;



inspecție automată contacte
puncte de sudură (optic)



detecție automată defecte

Aplicații concrete (cont.)

- **Robotică:**

roboți divertisment

roboți umanoizi

roboți electrocasnici

roboți manipulatori

Aplicații concrete (cont.)

- **Diagnoză de sistem:**

  - de regulă diagnoza defectelor este realizată de experți pe baza observației "manuale" a corelației dintre anumiți parametri și defectul în cauză (cunoștințe dobândite în timp ~ani de zile);

  - *exemplu*: mentenanță sisteme electro-mecanice (motoare, generatoare) – monitorizare vibrații pentru diagnoză rulmenți defecți, dezaliniere, slăbire componente, etc:

    - volum mare de date și echipamente, o fabrică are ~1000 de dispozitive de monitorizat, ~600 tipuri posibile de defecte;

    - pe baza datelor furnizate de experți (cumulate în timp), antrenare sistem de clasificare automată.

Aplicații concrete (cont.)

- **Marketing, vânzări și finanțe:**

  - volum imens de date de prelucrat – corelare și predicție date = $;
  - aplicații:

    - *domeniu bancar* - determinare profil de încredere pentru acordare credite, identificare clienți ce pot părăsi banca prin analiza tranzacțiilor realizate, identificare date atipice, etc;

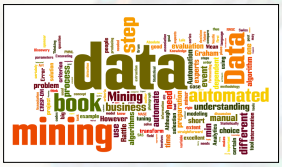
    - *consum* - determinare în funcție de produsele cumpărate a corelației între produse, ex. de regulă persoanele care cumpără bere cumpără și cipsuri, etc;

    - *marketing* – analiză date demografice și feedback potențiali cumpărători pentru determinarea automată a publicului țintă pentru un anumit produs (mai eficient decât "bulk" email/mail);

> **orice sistem actual include decizii și clasificare!**

Două abordări conexe

> **"machine learning"**: schimbarea comportamentului unui sistem astfel încât acesta să obțină performanțe mai bune în viitor – concept legat mai mult de performanță decât de cunoaștere.



> **"data mining"**: implică procesul de învățare în sensul practic, non teoretic – tehnici capabile să identifice și să descrie tipare structurale ale datelor ca mijloc de explicare a acelor date și pentru a face predicții pe baza lor.

Experimentul 1.1.

date de intrare

> să se regrupeze datele similare:

Experimentul 1.1. (cont.)

Q1: care a fost criteriul de decizie al numărului de clase rezultate?

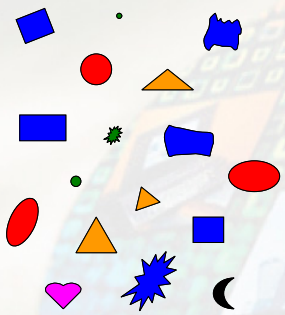
Q2: cum am decis că două date sunt similare sau nu?

Q3: există o singură partiționare "optimală"?

Q4: ce se întamplă cu datele care nu aparțin practic niciunei clase?

date de intrare

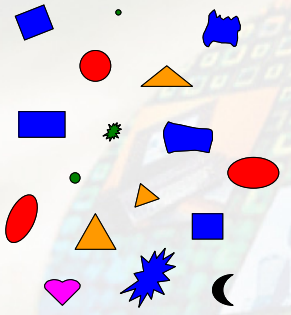
Experimentul 1.2.



> să se regrupeze datele similare:

date de intrare

Experimentul 1.2. (cont)



Q1: care a fost criteriul de decizie al numărului de clase rezultate?

Q2: cum am decis că două date sunt similare sau nu?

Q3: ce s-a schimbat față de experimentul anterior (1.1.)?

Q4: a fost mai ușor să partiționăm "optimal" datele în această variantă?

date de intrare

Concluzii experimentul 1.1. și 1.2.

- **clasificare ne-supervizată**, nu avem cunoștințe "a priori" despre semnificația și apartenența datelor;

- definitorii pentru partiționarea datelor au fost parametrii de conținut ai acestora (= descriptori) – aumite proprietăți se dovedesc a fi mai importante (= discriminatorii) decât altele (ex. culoare vs. formă);

- un număr mai mare de descriptori tinde să fie mai relevant pentru succesul (optimizarea) partiționării datelor;

- procesul este guvernat de definirea unei metrici (măsuri de distanță) pe baza cărei se evaluează similaritatea datelor;

- există date atipice, ce nu aparțin niciunei clase.

Experimentul 2.1.

clasa 1

clasa 2

clasa 3

clasa 4

date etichetate

NB: notați caracteristicile claselor pe hârtie.

Experimentul 2.2.

> să se regrupeze datele similare:

clasa 1

clasa 2

clasa 3

clasa 4

date de intrare

Experimentul 2.2. (cont)

date de intrare

Q1: care a fost criteriul de decizie al numărului de clase rezultate?

Q2: cum am decis că două date sunt similare sau nu?

Q3: există o singură partiționare "optimală"?

Q4: ce se întamplă cu datele care nu aparțin practic niciunei clase?

Concluzii experimentul 2.1. și 2.2.

- **clasificare supervizată**, sistemul este "antrenat" în prealabil să răspundă la anumite clase de date;

- se cunoaște numărul de clase de ieșire (sau se determină "a priori" în funcție de datele de antrenare);

- definitorii pentru învățare (și astfel clasificare) au fost parametrii de conținut ai acestora (= descriptori);

- procesul este guvernat de definirea unei metrici (măsuri de distanță) pe baza cărei se evaluează similaritatea datelor;

- învățarea nu este perfectă, clasificarea acelorași date de antrenare nu conduce la rezultate perfecte;

- există date atipice, acesta sunt asociate obligatoriu unei clase.

Definiție învățare

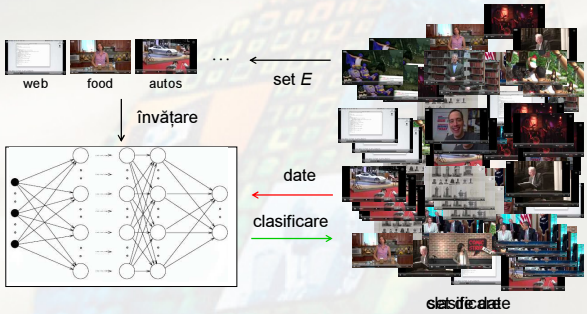
Definiție "learning": a dobândi cunoștințe sau aptitudini prin studiu, practică, experimentare sau prin intermediul altor persoane.
[dicționar Merriam-Webster]

Definiție "machine learning": un sistem învață din experiența *E* cu privire la o anumită clasă de cerințe *T* și o anumită măsură de performanță *P*, dacă performanța acestuia la cerințele din *T*, așa cum este măsurată de *P*, se îmbunătățește pe baza experienței din *E*.
[Rossen Dimov, Seminar A.I. Tools]

Exemplu:
- T = joc de șah;
- P = procentul de partide câștigate;
- E = 1000 de înregistrări a unor jocuri de șah.

> posibilitatea unui sistem de a "învăța" pe baza unor date;

Definiție învățare (cont.)

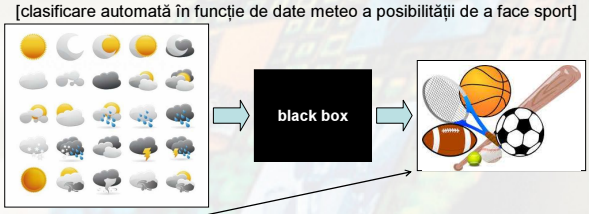


Terminologie

[sursă imagini Wikipedia]

concept = ceea ce se dorește a fi învățat de către sistem;

[clasificare automată în funcție de date meteo a posibilității de a face sport]



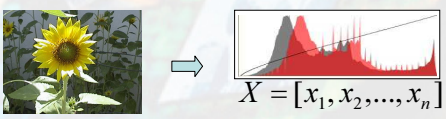
descriere concept = ceeea ce produce sistemul de învățare (ieșire sistem – reprezentare concept prin sistem);

Terminologie (cont.)

[sursă imagini Wikipedia]

dată = o entitate de informație unitară, exemplu: o imagine, o secvență video, un document, o înregistrare a unor parametri etc.

descriptor (observație, vector de caracteristici) = reprezentarea unei date într-o formă mai compactă, de regulă vectorială. Valorile vectorului reprezintă măsurători ale unor proprietăți definitorii ale datei respective:



dată descriptor (histogramă)

Terminologie (cont.)

[sursă imagine Wikipedia]

atribut (caracteristică, trăsătură) = o componenta a vectorului descriptor ce definește practic una dintre dimensiunile acestuia:

$$X = [x_1, x_2, ..., x_n] \Rightarrow$$

descriptor

atribut 1: x_1
atribut 2: x_2
...
atribut n: x_n

> descriptor = ansamblu de valori ale atributelor;

dată

$\Rightarrow X = [10, 5, ..., 12]$

$culoare_1$ $culoare_n$

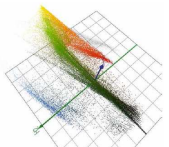
descriptor (ex. histogramă)

Terminologie (cont.)

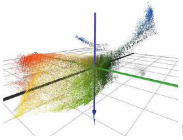
spațiu de caracteristici
= spațiul definit de descriptorii datelor; axele acestuia sunt definite de atributele descriptorului:

În exemplu:
> data = pixel imagine;
> atribute = proiectiile acestuia pe diferite spații de culoare;

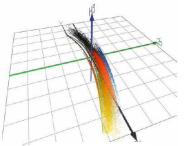
date



$x=\{R,G,B,X,Y,L,M,S\}$



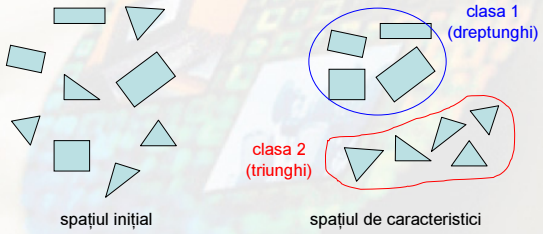
$x=\{L^*,a^*,H_{LCH},C_{LCH},H_{HSV},S_{HSV}\}$



$x=\{Y,L^*,L_{LHC},I_{HSI}\}$

Terminologie (cont.)

clasă = o sursă de date a căror distribuție în spațiul de caracteristici este guvernată de o anumită densitate de probabilitate specifică; astfel, o clasă definește un anumit tip de date cu proprietăți similare din punct de vedere al unor anumite criterii:



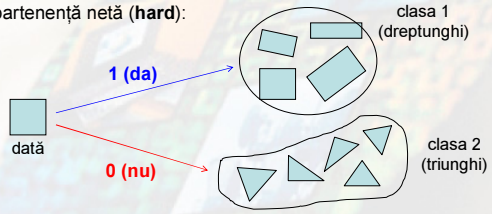
spațiul inițial

spațiul de caracteristici

Terminologie (cont.)

apartenență la clasă = asocierea unei date la o anumită clasă și astfel determinarea faptului că proprietățile acesteia sunt reprezentative pentru specificul clasei respective;

> apartenență netă (**hard**):

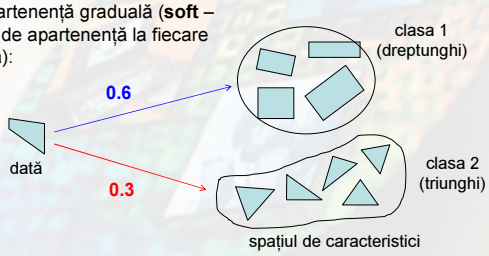


spațiul de caracteristici

Terminologie (cont.)

apartenență la clasă = asocierea unei date la o anumită clasă și astfel determinarea faptului că proprietățile acesteia sunt reprezentative pentru specificul clasei respective;

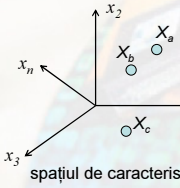
> apartenență graduală (**soft** – grad de apartenență la fiecare clasă):



spațiul de caracteristici

Terminologie (cont.)

metrică (distanță, măsură de similaritate) = o măsură de evaluare a gradului de similaritate între date diferite. De regulă returnează o valoare mică când datele sunt similare (ex. aparțin aceleiași clase) și o valoare semnificativă când sunt diferite:



spațiul de caracteristici

$$X_i = [x_{i,1}, x_{i,2}, ..., x_{i,n}],\ i \in \{a, b, c\}$$

descriptori

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{n} (x_{i,k} - x_{j,k})^2}$$

metrică (ex. distanța Euclidiană)

$$d(X_a, X_b) \lessgtr d(X_b, X_c) \lessgtr d(X_a, X_c)$$

Terminologie (cont.)

set de date (bază de date) = ansamblul datelor ce urmează să fie analizate și clasificate;

date etichetate (ground truth) = o colecție de date pentru care se cunoaște "a priori" apartenența la clase; acestea sunt de regulă obținute pe baza expertizei umane:

set de date

$X_1 \in$ clasa 1
$X_2 \in$ clasa 1
$X_4 \in$ clasa 2
$X_5 \in$ clasa 2
$X_7 \in$ clasa 3

$X_3 \in$ clasa ?
$X_6 \in$ clasa ?
...

Terminologie (cont.)

> rezumat: set de date

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicată	nu	Nu
3	ploioasă	cald	ridicată	nu	?
4	înnorat	rece	normală	da	?

set de date

Terminologie (cont.)

> rezumat (cont.): atribute

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicată	nu	Nu
3	ploioasă	cald	ridicată	nu	?
4	înnorat	rece	normală	da	?

atribute

Terminologie (cont.)

> rezumat (cont.): atribut clasă

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicată	nu	Nu
3	ploioasă	cald	ridicată	nu	?
4	înnorat	rece	normală	da	?

atribut clasă

Terminologie (cont.)

> rezumat (cont.): instanță

nr.	vreme	temperatură	umiditate	vânt	play
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicată	nu	Nu
3	ploioasă	cald	ridicată	nu	?
4	înnorat	rece	normală	da	?

instanță (dată)

Terminologie (cont.)

> rezumat (cont.): date etichetate

nr.	vreme	temperatură	umiditate	vânt	play
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicată	nu	Nu
3	ploioasă	cald	ridicată	nu	?
4	înnorat	rece	normală	da	?

date etichetate

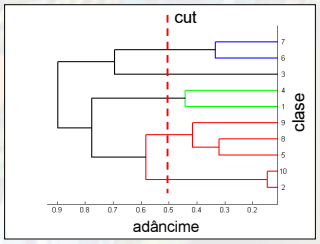
Tehnici existente

[sursă imagine Wikipedia]

1. Tehnici de clasificare nesupervizată (**clustering**)

- **metode ierarhice:**

[datele de intrare sunt grupate într-un număr variabil de clase sub forma unui arbore (dendrogram) pornind de la toate elementele într-o clasă și finalizând cu fiecare element într-o clasa separată]



Metode: Hierarchical Clustering cu diferite variante, agglomerative - "bottom up" sau divisive - "top down".

Tehnici existente (cont.)

1. Tehnici de clasificare nesupervizată (cont.)

- **metode partiționale:**

[produc o singură partiție și depind de alegerea numărului de clase de ieșire. Partiționarea se face folosind un criteriu de optimizare căutându-se prin încercări partiționarea optimală. Algoritmul este de regulă rulat repetitiv cu diferite puncte de plecare a partițiilor alegându-se în final varianta cea mai bună]

Metode:
- square error: k-means, ISODATA;
- graph-theoretic: Minimal Spanning Tree (MST);
- mixture resolving: Expectation Maximization (EM);
- nearest neighbor;
- fuzzy: fuzzy c-means (FCM).

Tehnici existente (cont.)

1. Tehnici de clasificare nesupervizată (cont.)

- o altă clasificare globală:

• **acumulative vs. partiționale**: clasificarea pornește de la o anumită partiție în clase, clase care pe parcursul algoritmului sunt fuzionate iterativ până când este satisfăcut un anumit criteriu de convergență vs. clasificarea pornește de la o singură clasă care este divizată iterativ până când criteriul de convergență este satisfăcut;

• **politetice vs. monotetice**: la stabilirea claselor sunt folosite toate atributele de intrare vs. atributele de intrare sunt folosite în mod secvențial pentru a constitui progresiv clasele, ex. atributul x_1 este folosit pentru a diviza datele în două clase, mai departe, atributul x_2 este folosit pentru divizarea claselor anterioare, și așa mai departe;

Tehnici existente (cont.)

1. Tehnici de clasificare nesupervizată (cont.)

- o altă clasificare globală (cont.):

• **nete vs. fuzzy**: datele sunt alocate unei singure clase, apartenența fiind binară (1 sau 0) vs. datele au un grad de apartenență la una sau mai multe clase - cu cât valoarea este mai mare cu atât este mai probabil să aparțină clasei respective;

• **deterministe vs. stohastice**: optimizarea claselor este deterministă pe baza unui algoritm determinist vs. se folosește o căutare aleatoare în spațiul format de toate clasificările posibile;

• **incrementale vs. non-incrementale**: volum foarte mare de date (ex. Big Data) - minimizare număr de citiri al datelor, reducere număr de repartiții în clase analizate, reducere date, partiționare progresivă crescând setul de date.

Tehnici existente (cont.)

2. Tehnici de clasificare supervizată (**classification**)

- **bazate pe criteriul Bayes:**

[clasificator probabilistic, de regulă binar (două clase), ce se bazează pe o ipoteză de independență a atributelor de intrare (naivă); fiecărei clase i se asociază o probabilitate, $p(C_k \mid x_1,...,x_n)$ unde C_k – clasa k iar x sunt datele de clasificat; ieșirea clasificatorului este clasa cea mai probabilă (optimizare în funcție de datele de antrenare)]

Metode:
- Naive Bayes;
- Bayes Networks;
- AODE, etc.

$$p(C_k \mid x) = \frac{p(C_k)p(x \mid C_k)}{p(x)}$$

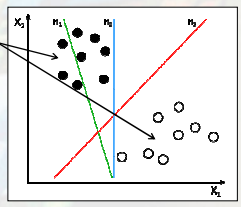
Tehnici existente (cont.)

[sursă imagine Wikipedia]

2. Tehnici de clasificare supervizată (cont.)

- **bazate pe funcții:**

[problema clasificării este modelată cu ajutorul unei reprezentări funcționale a datelor de intrare, reprezentare ce este optimizată folosind datele de antrenare]



Metode:
- Support Vector Machines – optimizează un hiperplan de separație a datelor din spațiul de caracteristici;
- Radial Basis Function network – rețea neuronală a cărei ieșire este o combinație funcțională a intrării;
- Linear Regression – asocierea optimală a unor funcții liniare perechilor de date de intrare-ieșire (date de antrenare), etc.

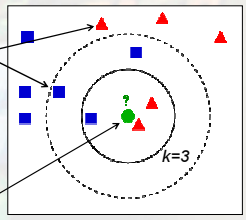
Tehnici existente (cont.)

[sursă imagine Wikipedia]

2. Tehnici de clasificare supervizată (cont.)

- **metode "leneșe":**

[antrenarea propriu-zisă este de fapt realizată cu întârziere de abia în momentul clasificării unei date noi; clasificarea este optimizată local, pentru fiecare dată nouă, ceea ce le face adaptabile la modificarea domeniului datelor]



Metode:
- k-Nearest Neighbours;
- Locally Weighted Learning (LWL);
etc.

Tehnici existente (cont.)

2. Tehnici de clasificare supervizată (cont.)

- **bazate pe reguli de decizie:**

[se bazează pe generarea și optimizarea unui set de reguli de decizie de tip "dacă – atunci" folosind datele de antrenare; regulile nu sunt neapărat exclusive]

vârstă=16, prescripție="miopie", astigmatism=0, lacrimi="reduse" - lentile=0;
vârstă=14, prescripție="miopie", astigmatism=1, lacrimi="reduse" - lentile=0;
…

dacă (vârstă>14 && vârstă<16) && (prescripție=="miopie") && (lacrimi=="reduse") atunci lentile=0;

Metode:
- Decision Table;
- RIpple-DOwn Rule learner (RIDOR);
etc.

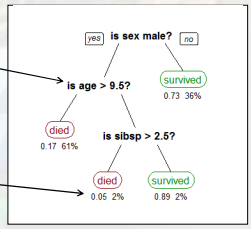
Tehnici existente (cont.)

[sursă imagine Wikipedia]

2. Tehnici de clasificare supervizată (cont.)

- **bazate pe arbori:**

[reprezintă informația sub formă arborescentă, unde clasele sunt reprezentate de frunzele arborelui, nodurile corespund atributelor datelor iar ramurile reprezintă relaționarea valorilor atributelor pentru a ajunge la un anumit nod]



Metode:
- Functional Trees;
- Random Trees;
- Random Forests;
- C4.5, etc.

Tehnici existente (cont.)

2. Tehnici de clasificare supervizată (cont.)

- **meta-metode:**

[reunesc mai mulți clasificatori existenți; de regulă învață în mod iterativ un set de clasificatori "slabi" și îi adaugă progresiv la clasificatorul global; pe măsură ce sunt adăugați, datele sunt re-ponderate, datelor clasificate eronat li se crește ponderea în timp ce datelor clasificate corect le scade ponderea; astfel încât următorul clasificator slab se va focaliza pe datele clasificate greșit]

Metode:
- AdaBoost;
- LogitBoost;
- AnyBoost, etc.

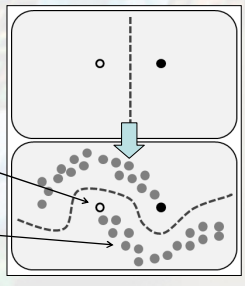
Tehnici existente (cont.)

[sursă imagine Wikipedia]

3. Tehnici de clasificare semi-supervizată

- se poziționează între tehnicile de clasificare nesupervizată și cele supervizate;

- **idee**: folosirea de date ne-etichetate în corelație cu date etichetate (datele etichetate sunt limitate și necesită un efort considerabil pentru a fi obținute);



Utilitare

- Neural Network Toolbox;
- Bioinformatics Toolbox (Support Vector Machines);
- Statistics Toolbox (Hierarchical Clustering, K-Means, Gaussian Mixture Models, Naive Bayes, Discriminant Analysis, Nearest Neighbors, Classification Trees, Ensemble Classifiers, etc);
- poate fi completat cu alți clasificatori (Internet).

- platformă specializată, "open source", în Java;
- furnizează tot lanțul de prelucrare: pregătire date, selecție atribute, clasificare supervizată și nesupervizată, evaluare performanțe, analiză rezultate, etc;
- poate rula în linie de comandă (batch processing);
- poate rula multi-procesor.

Utilitare (cont.)

- platformă generală de computer vision ce include și facilități de clasificare, “open source”, în C++;
- Machine Learning Library (MLL): Statistical Models, Bayes Classifier, K-Nearest Neighbors, Support Vector Machines, Decision Trees, Boosting, Expectation Maximization, Neural Networks;
- poate rula multi-procesor.

- platformă specializată, comercială;
- mod de operare vizuală, fără programare;
- folosită la nivel global de companii care prelucrează și analizează date (ex. financiare);
- pachet de utilitare “puternice”, validate.

> Sfârşit M1