UNIVERSITATEA BABEŞ-BOLYAI
Facultatea de Matematică şi Informatică

INTELIGENȚĂ ARTIFICIALĂ

Sisteme inteligente

Sisteme care învaţă singure

– SVM, kNN, arbori de decizie –

Laura Dioşan

Sumar

A. Scurtă introducere în Inteligenţa Artificială (IA)

B. Rezolvarea problemelor prin căutare
- Definirea problemelor de căutare
- Strategii de căutare
  - Strategii de căutare neinformate
  - Strategii de căutare informate
  - Strategii de căutare locale (Hill Climbing, Simulated Annealing, Tabu Search, Algoritmi evolutivi, PSO, ACO)
  - Strategii de căutare adversială

C. Sisteme inteligente
- Sisteme care învaţă singure
  - Mașini cu Suport Vectorial
  - Cel mai apropiat vecin (kNN – k nearest neighbour)
  - Arbori de decizie
  - Reţele neuronale artificiale
  - Maşini cu suport vectorial
  - Algoritmi evolutivi
- Sisteme bazate pe reguli
- Sisteme hibride

Materiale de citit şi legături utile

- capitolul VI (18) din *S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, 1995*
- capitolul 10 şi 11 din *C. Groşan, A. Abraham, Intelligent Systems: A Modern Approach, Springer, 2011*
- capitolul V din *D. J. C. MacKey, Information Theory, Inference and Learning Algorithms, Cambridge University Press, 2003*
- capitolul 3 din *T. M. Mitchell, Machine Learning, McGraw-Hill Science, 1997*

Conținut

- Sisteme inteligente
  - Sisteme care învaţă singure (SIS)
    - Sisteme
      - Logistic regression
      - kNN
      - Arbori de decizie
      - SVM

Sisteme inteligente – SIS – Învățare automată

- Tipologie
  - În funcţie de experienţa acumulată în timpul învăţării:
    - SI cu învăţare supervizată
    - SI cu învăţare nesupervizată
    - SI cu învăţare activă
    - SI cu învăţare cu întărire
  - În funcţie de modelul învăţat (algoritmul de învăţare):
    - Arbori de decizie
    - Reţele neuronale artificiale
    - **Maşini cu suport vectorial (MSV)**
    - Algoritmi evolutivi
    - Modele Markov ascunse

Sisteme inteligente – SIS – MSV

- Maşini cu suport vectorial (MSV)
  - Definire
  - Tipuri de probleme rezolvabile
  - Avantaje
  - Dificultăţi
  - Tool-uri

Sisteme inteligente – SIS – MSV

- Definire
  - Dezvoltate de Vapnik în 1970
  - Popularizate după 1992
  - Clasificatori liniari care identifică un hiperplan de separare a clasei pozitive de clasa negativă
  - Au o fundamentare teoretică foarte riguroasă
  - Funcţionează foarte bine pentru date de volum mare (analiza textelor, analiza imaginilor)

  - **Reamintim**
    - Problemă de învăţare supervizată în care avem un set de date de forma:
      - (x^d, t^d), cu:
      - $x^d \in \mathbf{R}^m$ ➔ $x^d=(x^d_1, x^d_2,..., x^d_m)$
      - $t^d \in \mathbf{R}$ ➔ $t^d \in \{1, -1\}$, 1 → clasă pozitivă, -1 → clasă negativă
      - cu d = 1,2,...,n,n+1,n+2,...,N
    - Primele n date (se cunosc x^d şi t^d) vor fi folosite drept bază de antrenament a MSV
    - Ultimele N-n date (se cunosc doar x^d, fără t^d) vor fi folosite drept bază de testare a MSV

Sisteme inteligente – SIS – MSV

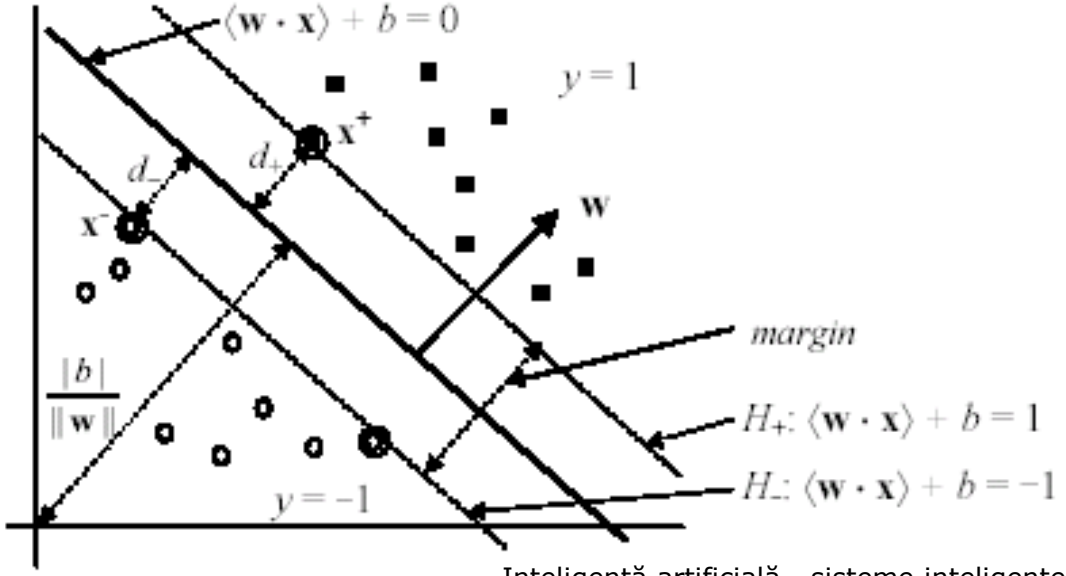
- Definire
  - MSV găseşte o funcţie liniară de forma $f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$, ($\mathbf{w}$ -vector pondere) a.î.

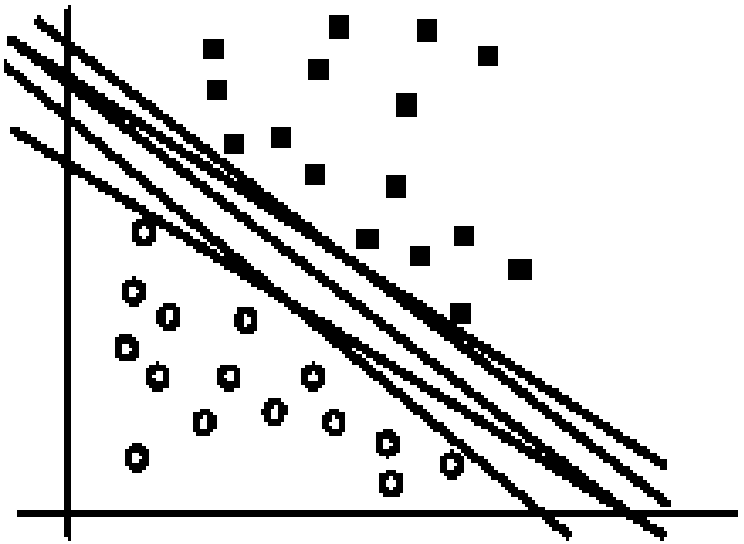
$$y_i = \begin{cases} 1 & if \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq 0 \\ -1 & if \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b < 0 \end{cases}$$

  - $\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$ → hiperplanul de decizie care separă cele 2 clase

Sisteme inteligente – SIS – MSV

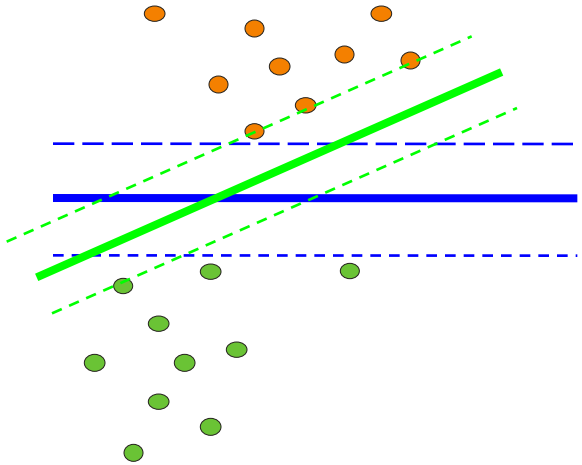
- Definire
  - Pot exista mai multe hiperplane
    - Care este cel mai bun hiperplan?
  - MSV caută hiperplanul cu cea mai largă margine (cel care micşorează eroarea de generalizare)
    - Algoritmul SMO (*Sequential minimal optimization*)



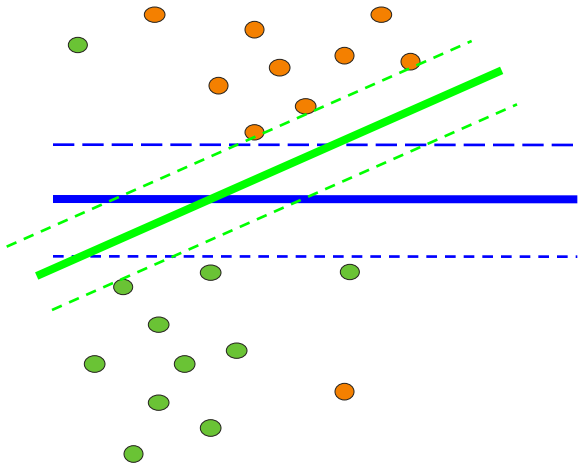


Sisteme inteligente – SIS – MSV

- Tipuri de probleme rezolvabile
  - Probleme de clasificare → Cazuri de date
    - Liniar separabile
      - Separabile
        - Eroarea = 0

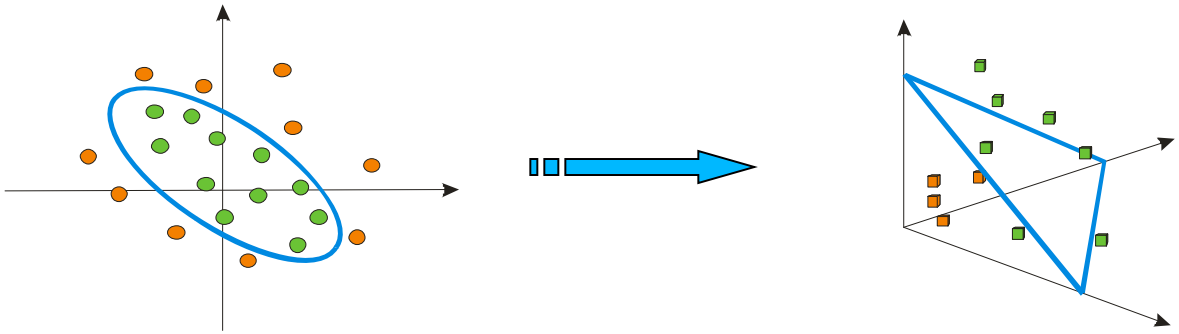


      - Ne-separabile
        - Se relaxează constrângerile → se permit unele erori
        - C – coeficient de penalizare



Sisteme inteligente – SIS – MSV

- Cazuri de date
  - Non-liniar separabile
    - Spaţiul de intrare se transformă într-un spaţiu cu mai multe dimensiuni (*feature space*), cu ajutorul unei funcţii kernel, unde datele devin liniar separabile
    - În MSV, funcţiile kernel calculează distanţa între 2 puncte
      - → kernelul ~ funcţie de similaritate



Sisteme inteligente – SIS – MSV

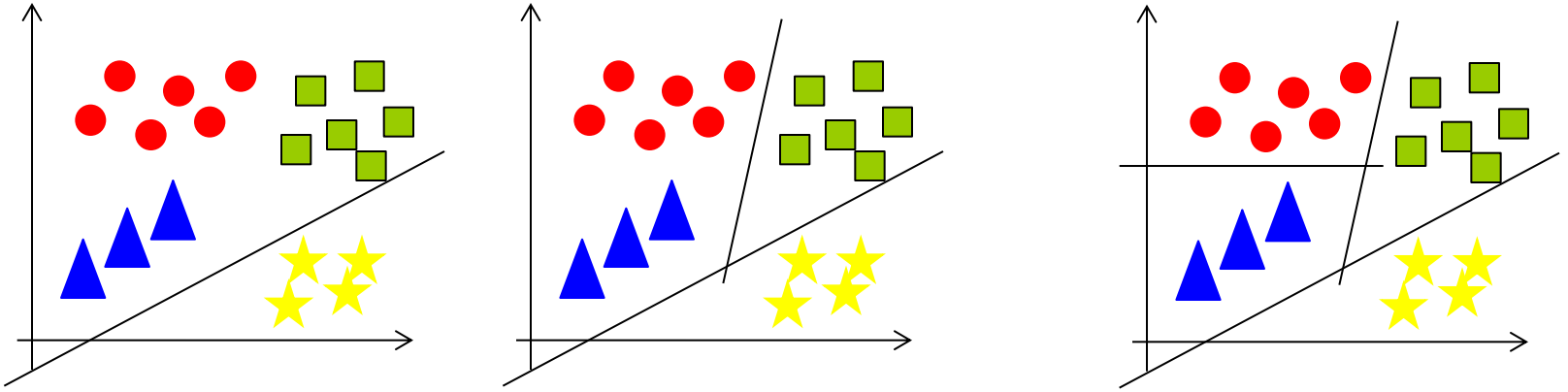
- Cazuri de date
  - Non-liniar separabile → Kernele posibile
    - Clasice
      - Polynomial kernel: $K(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) = (\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2} + 1)^p$
      - RBF kernel: $K(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) = exp(- ||\boldsymbol{x}^{d1} - \boldsymbol{x}^{d2}||^2/2\sigma^2)$
    - Kernele multiple
      - Liniare: $K(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) = \sum w_i K_i (\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2})$
      - Ne-liniare
        - Fără coeficienţi: $K(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) = K_1(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) + K_2(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) * exp(K_3(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}))$
        - Cu coeficienţi: $K(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) = K_1 (\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) + c_1 * K_2 *(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) exp(c_2 + K_3(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}))$
    - Kernele pentru stringuri
    - Kernele pentru imagini
    - Kernele pentru grafe

Sisteme inteligente – SIS – MSV

- Configurarea MSV
  - Parametrii unei MSV
    - Coeficientul de penalizare C
      - C – mic → convergenţă lentă
      - C – mare → convergenţă rapidă
    - Parametrii funcţiei kernel (care kernel şi cu ce parametri)
      - Dacă m (nr de atribute) este mult mai mare decât n (nr de instanţe)
        - MSV cu kernel liniar (MSV fără kernel) → $K(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) = \boldsymbol{x}^{d1} \cdot \boldsymbol{x}^{d2}$
      - Dacă m (nr de atribute) este mare, iar n (nr de instanţe) este mediu
        - MSV cu kernel Gaussian $K(\boldsymbol{x}^{d1}, \boldsymbol{x}^{d2}) = \exp(-||\boldsymbol{x}^{d1} - \boldsymbol{x}^{d2}||^2/2\sigma^2)$
        - σ – dispersia datelor de antrenament
        - Atributele instanţelor trebuie normalizate (scalate la (0,1))
      - m (nr de atribute) este mic, iar n (nr de instanţe) este mare
        - Se adaugă noi atribute, iar apoi
        - MSV cu kernel liniar

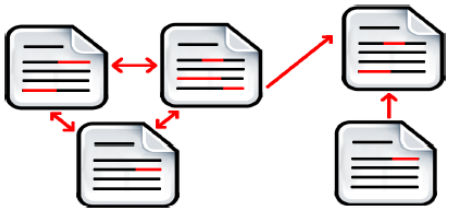
Sisteme inteligente – SIS – MSV

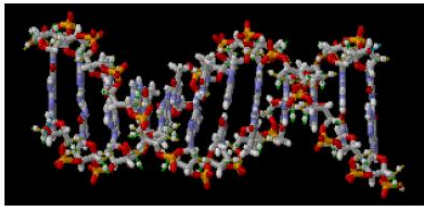
- MSV pentru probleme de clasificare supervizate cu mai mult de 2 clase
  - Una vs. restul (one *vs.* all)

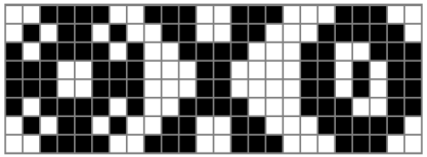


Sisteme inteligente – SIS – MSV

- MSV structurate
  - Învăţare automată
    - Normală f: $\mathcal{X}$ → $\mathbf{R}$
      - Intrări de orice fel
      - Ieşiri numerice (naturale, întregi, reale)
    - Structurată: $\mathcal{X}$ → $\mathcal{Y}$
      - Intrări de orice fel
      - Ieşiri de orice fel (simple sau structurate)
  - Informaţii structurate
    - Texte şi hiper-texte
    - Molecule şi structuri moleculare
    - Imagini







Sisteme inteligente – SIS – MSV

- MSV structurate
  - Aplicaţii
    - Procesarea limbajului natural
      - Traduceri automate (ieşiri → propoziţii)
      - Analiza sintactică şi/sau morfologică a propoziţiilor (ieşiri → arborele sintactic şi/sau morfologic)
    - Bioinformatică
      - Predicţia unor structuri secundare (ieşirile → grafe bi-partite)
      - Predicţia funcţionării unor enzime (ieşirile → *path*-uri în arbori)
    - Procesarea vorbirii
      - Transcrieri automate (ieşiri → propoziţii)
      - Transformarea textelor în voce (ieşiri → semnale audio)
    - Robotică
      - Planificare (ieşirile → secvenţe de acţiuni)

Sisteme inteligente – SIS – MSV

- Avantaje
  - Pot lucra cu orice fel de date (liniar separabile sau nu, distribuit uniform sau nu, cu distribuţie cunoscută sau nu)
    - Funcţiile kernel care crează noi atribute (features) → straturile ascunse dintr-o RNA
  - Dacă problema e convexă oferă o soluţie unică → optimul global
    - RNA pot asocia mai multe soluţii → optime locale
  - Selectează automat mărimea modelului învăţat (prin vectorii suport)
    - În RNA straturile ascunse trebuie configurate de către utilizator *apriori*
  - Nu învaţă pe derost datele (overfitting)
    - RNA se confruntă cu problema overfitting-ului chiar şi cand modelul se învaţă prin validare încrucişată

- Dificultăţi
  - Doar atribute reale
  - Doar clasificare binară
  - Background matematic dificil

- Tool-uri
  - LibSVM → http://www.csie.ntu.edu.tw/~cjlin/libsvm/
  - Weka → SMO
  - SVMLight → http://svmlight.joachims.org/
  - SVMTorch → http://www.torch.ch/
  - http://www.support-vector-machines.org/

Sisteme inteligente – SIS

- kNN (cei mai apropiați k vecini)
  - Cel mai simplu algoritm de clasificare
  - În etapa de antrenament, algoritmul doar citeşte datele de intrare (atributele şi clasa fiecărei instanţe)
  - În etapa de testare, pentru o nouă instanţă (fără clasă) se caută (printre instanţele de antrenament) cei mai apropiaţi *k* vecini şi se preia clasa majoritară a acestor *k* vecini
  - Căutarea vecinilor se bazează pe:
    - distanţa Minkowski (Manhattan, Euclidiană) – atribute continue
    - distanţa Hamming, Levensthein – analiza textelor
    - alte distanţe (funcţii kernel)

Sisteme inteligente – SIS – Arbori de decizie

- Arbori de decizie
  - Scop
  - Definire
  - Tipuri de probleme rezolvabile
  - Exemplu
  - Proces
  - Tool-uri
  - Avantaje şi limite

Sisteme inteligente – SIS – Arbori de decizie

- Scop
  - Divizarea unei colecţii de articole în seturi mai mici prin aplicarea succesivă a unor reguli de decizie → adresarea mai multor întrebări
    - Fiecare întrebare este formulată în funcţie de răspunsul primit la întrebarea precedentă
  - Elementele se caracterizează prin informaţii non-metrice
- Definire
  - Arborele de decizie
    - Un graf special → arbore orientat bicolor
    - Conţine noduri de 3 tipuri:
      - Noduri de decizie → posibilităţile decidentului (ex. Diversele examinări sau tratamente la care este supus pacientul) şi indică un test pe un atribut al articolului care trebuie clasificat
      - Noduri ale hazardului – evenimente aleatoare în afara controlului decidentului (rezultatul examinărilor, efectul terapiilor)
      - Noduri rezultat – situaţiile finale cărora li se asociază o utilitate (apreciată aprioric de către un pacient generic) sau o etichetă
    - Nodurile de decizie şi cele ale hazardului alternează pe nivelele arborelui
    - Nodurile rezultat – noduri terminale (frunze)
    - Muchiile arborelui (arce orientate) → consecinţele în timp (rezultate) ale decizilor, respectiv ale realizării evenimentelor aleatoare (pot fi însoţite de probabilităţi)
  - Fiecare nod intern corespunde unui atribut
  - Fiecare ramură de sub un nod (atribut) corespunde unei valori a atributului
  - Fiecare frunză corespunde unei clase

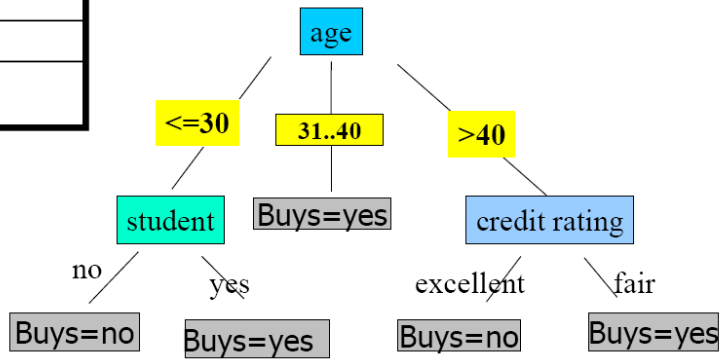
Sisteme inteligente – SIS – Arbori de decizie

- Tipuri de probleme
  - Exemplele (instanţele) sunt reprezentate printr-un număr fix de atribute, fiecare atribut putând avea un număr limitat de valori
  - Funcţia obiectiv ia valori de tip discret
  - AD reprezintă o disjuncţie de mai multe conjuncţii, fiecare conjuncţie fiind de forma atributul a_i are valoarea v_j
  - Datele de antrenament pot conţine erori
  - Datele de antrenament pot fi incomplete
    - Anumitor exemple le pot lipsi valorile pentru unele atribute
  - Probleme de clasificare
    - Binară
      - exemple date sub forma [($atribut_{ij}$, $valoare_{ij}$), $clasă_i$, i=1,2,...,n, j=1,2,...,m, $clasă_i$ putând lua doar 2 valori]
    - Multi-clasă
      - exemple date sub forma [($atribut_{ij}$, $valoare_{ij}$), $clasă_i$, i=1,2,...,n, j=1,2,...,m, $clasă_i$ putând lua k valori]
  - Probleme de regresie
    - AD se construiesc similar cazului problemei de clasificare, dar în locul etichetării fiecărui nod cu eticheta unei clase se asociază nodului o valoare reală sau o funcţie dependentă de intrările nodului respectiv
    - Spaţiul de intrare se împarte în regiuni de decizie prin tăieturi paralele cu axele Ox şi Oy
    - Are loc o transformare a ieşirilor discrete în funcţii continue
    - Calitatea rezolvării problemei
      - Eroare (pătratică sau absolută) de predicţie

Sisteme inteligente – SIS – Arbori de decizie

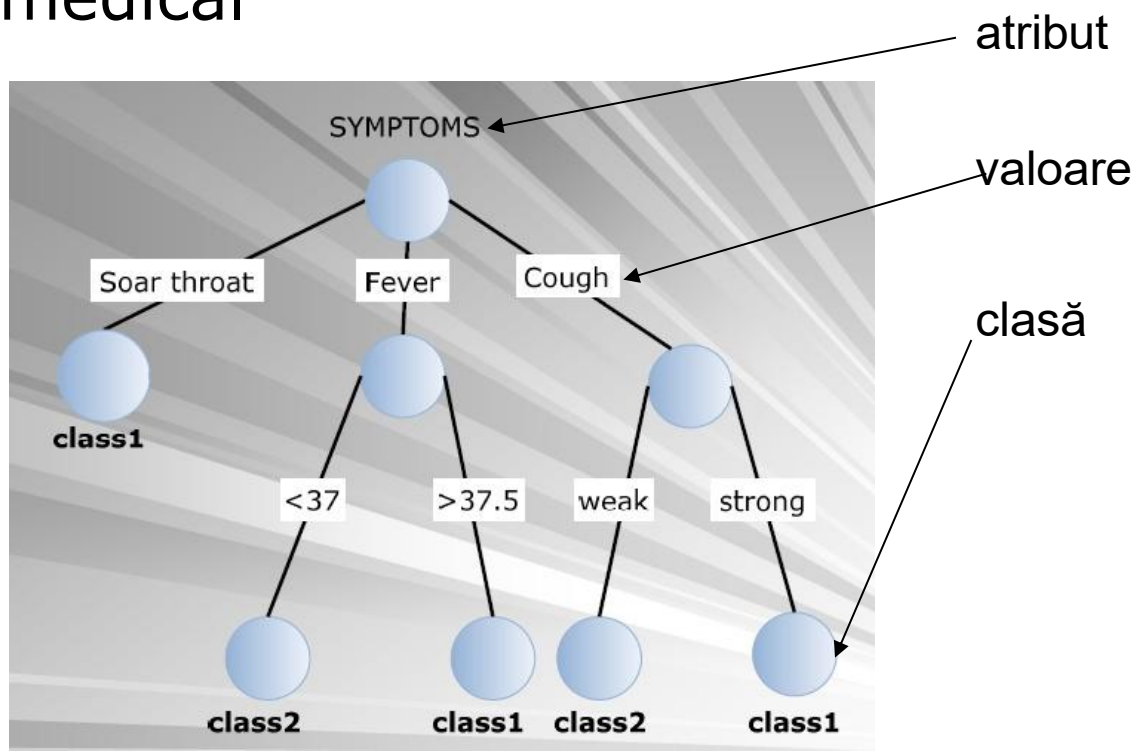
- Exemplu

rec	Age	Income	Student	Credit_rating	Buys_computer(CLASS)
r1	<=30	High	No	Fair	No
r2	<=30	High	No	Excellent	No
r3	31…40	High	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31…40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	<=30	Low	Yes	Fair	Yes
r10	>40	Medium	Yes	Fair	Yes
r11	<=30	Medium	Yes	Excellent	Yes
r12	31…40	Medium	No	Excellent	Yes
r13	31…40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No



Sisteme inteligente – SIS – Arbori de decizie

- Exemplu
  - Sistem medical



- Exemplu
  - Acordarea de credite

aprobat sau nu

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

Sisteme inteligente – SIS – Arbori de decizie

- Proces
  - Construirea (creşterea, inducţia) arborelui
    - Se bazează pe un set de date de antrenament
    - Lucrează de jos în sus sau de sus în jos (prin divizare – *splitting*)
  - Utilizarea arborelui ca model de rezolvare a problemelor
    - Ansamblul deciziilor efectuate de-a lungul unui drum de la rădăcină la o frunză formează o regulă
    - Regulile formate în AD sunt folosite pentru etichetarea unor noi date
  - Tăierea (curăţirea) arborelui (pruning)
    - Se identifică şi se mută/elimină ramurile care reflectă zgomote sau excepţii

Sisteme inteligente – SIS – Arbori de decizie

- Proces → Construirea AD
  - Divizarea datelor de antrenament în subseturi pe baza caracteristicilor datelor
    - Un nod → întrebare legată de o anumită proprietate a unui obiect dat
    - Ramurile ce pleacă din nod → etichetate cu posibilele răspunsuri la întrebarea din nodul curent
    - La început toate exemplele sunt plasate în rădăcină
      - La pornire, un atribut va fi rădăcina arborelui, iar valorile atributului vor deveni ramuri ale rădăcinii
    - Pe următoarele nivele exemplele sunt partiţionate în funcţie de atribute ➔ ordinea considerării atributelor
      - Pentru fiecare nod se alege în mod recursiv câte un atribut (cu valorile lui pe ramurile descendente din nodul curent)
    - Divizarea → greedy în luarea decizilor
  - Proces iterativ
    - Reguli de oprire
      - toate exemplele aferente unui nod fac parte din aceeaşi clasă → nodul devine frunză şi este etichetat cu Ci
      - Nu mai sunt exemple → nodul devine frunză şi este etichetat cu clasa majoritară în setul de date de antrenament
      - nu mai pot fi considerate noi atribute

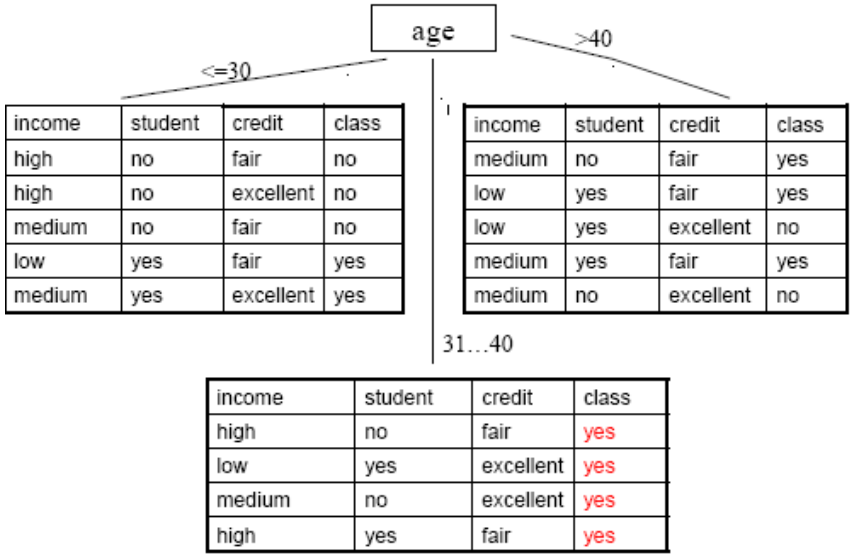
Sisteme inteligente – SIS – Arbori de decizie

- Proces → Construirea AD
  - Exemplu

rec	Age	Income	Student	Credit_rating	Buys_computer(CLASS)
r1	<=30	High	No	Fair	No
r2	<=30	High	No	Excellent	No
r3	31…40	High	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31…40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	<=30	Low	Yes	Fair	Yes
r10	>40	Medium	Yes	Fair	Yes
r11	<=30	Medium	Yes	Excellent	Yes
r12	31…40	Medium	No	Excellent	Yes
r13	31…40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No

Sisteme inteligente – SIS – Arbori de decizie

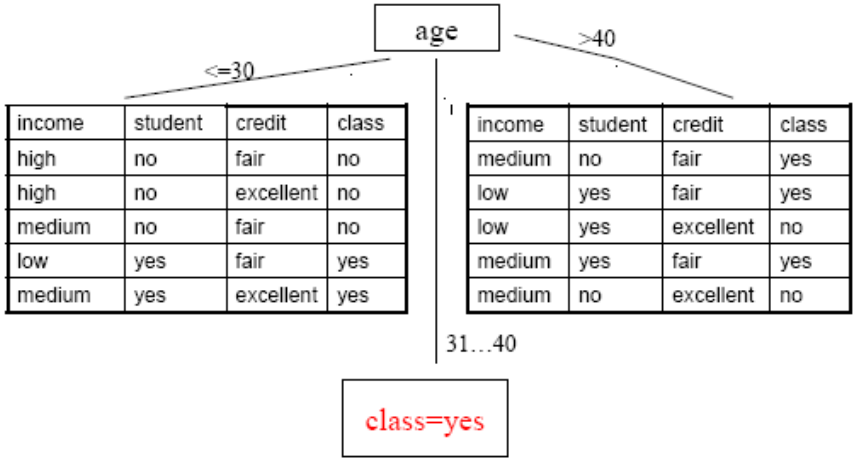
- Proces → Construirea AD
  - Exemplu
    - Pentru rădăcină se alege atributul *age*



income	student	credit	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	student	credit	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

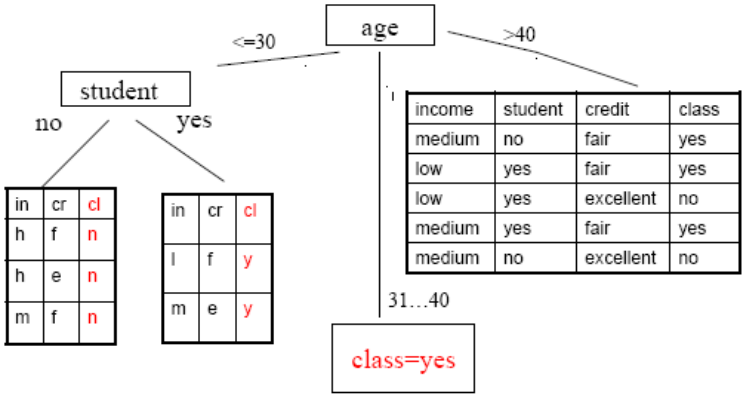


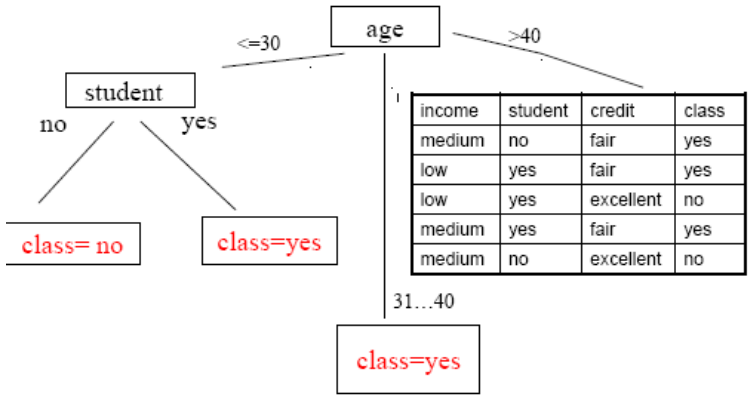
income	student	credit	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

Sisteme inteligente – SIS – Arbori de decizie

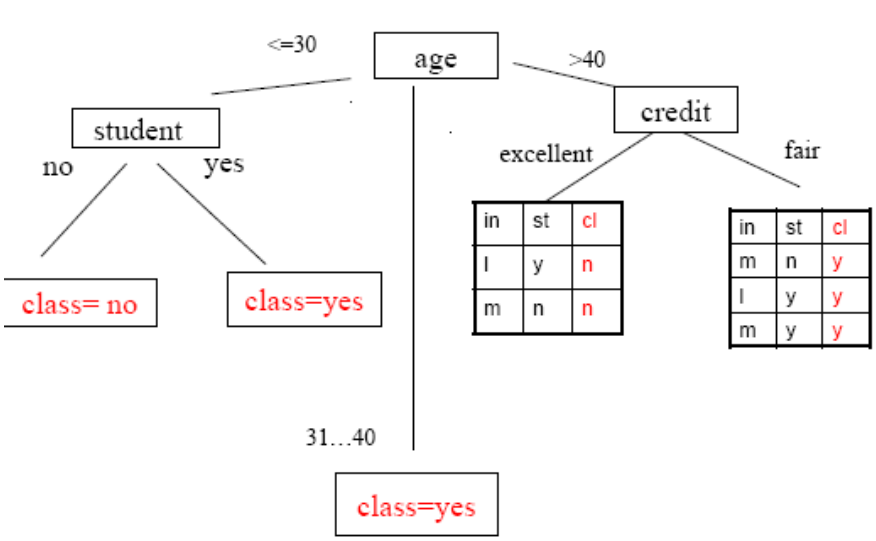
- Proces → Construirea AD
  - Exemplu
    - Pentru rădăcină se alege atributul *age*
    - Pe ramura <=30 se alege atributul student

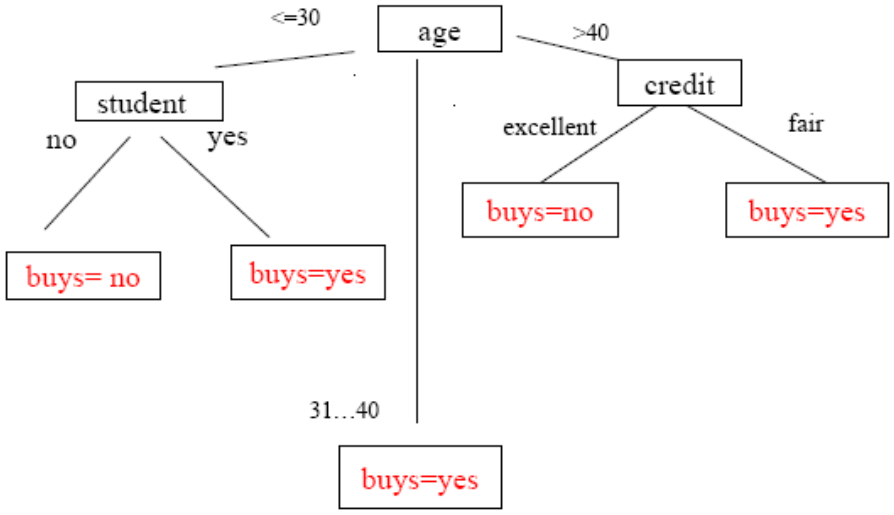




Sisteme inteligente – SIS – Arbori de decizie

- Proces → Construirea AD
  - Exemplu
    - Pentru rădăcină se alege atributul *age*
    - Pe ramura <=30 se alege atributul student
    - Pe ramura > 40 se alege atributul credit





Sisteme inteligente – SIS – Arbori de decizie

- Proces → Construirea AD → Algoritmul ID3/C4.5
  - Greedy, recursiv, top-down, divide-and-conquer

```
generare(D, A){      //D – partiţionare a exemplelor de antrenament, A – lista de atribute
    Crearea unui nod nou N
    Dacă exemplele din D fac parte dintr-o singură clasă C atunci
        nodul N devine frunză şi este etichetat cu C
        returnează nodul N
    Altfel
        Dacă A=∅ atunci
            nodul N devine frunză şi este etichetat cu clasa majoritară în D
            returnează nodul N
        Altfel
            atribut_separare = Selectează_atribut(D, A)
            Etichetează nodul N cu atribut_separare
            Pentru fiecare valoare posibilă vj a lui atribut_separare
                Fie Dj mulţimea exemplelor din D pentru care atribut_separare = vj
                Dacă Dj =∅ atunci
                    Ataşează  nodului N o frunză etichetată cu clasa majoritară în D
                Altfel
                    Ataşează nodului N un nod returnat de generare(Dj, A –atribut_separare)
            Returnează nodul N
}
```

Sisteme inteligente – SIS – Arbori de decizie

- Proces → Construirea AD → Algoritmul ID3/C4.5
  - Selectează_atribut(D, A) → Alegerea atributului aferent unui nod (rădăcină sau intern)
    - Aleatoare
    - Atributul cu cele mai puţine/multe valori
    - Pe baza unei ordini prestabilite a atributelor
      - Câştigul de informaţie
      - Rata câştigului
      - Indicele Gini
      - Distanţa între partiţiile create de un atribut

Sisteme inteligente – SIS – Arbori de decizie

- Proces → Construirea AD → Algoritmul ID3/C4.5 → Selectare atribut
  - Câştigul de informaţie
    - O măsură de impuritate
      - 0 (minimă) dacă toate exemplele fac parte din aceeaşi clasă
      - 1 (maximă) dacă avem număr egal de exemple din fiecare clasă
    - Se bazează pe entropia datelor
      - măsoară impuritatea datelor
      - numărul sperat (aşteptat) de biţi necesari pentru a coda clasa unui element oarecare din setul de date
      - clasificare binară (cu 2 clase): $E(S) = - p_+ log_2 p_+ - p_- log_2 p_-$, *unde*
        - p_+ - proporţia exemplelor pozitive în setul de date S
        - p_- - proporţia exemplelor negative în setul de date S
      - clasificare cu mai multe clase: $E(S) = \sum_{i=1, 2, ..., k} - p_i log_2 p_i$ – entropia datelor relativ la atributul ţintă (atributul de ieşire), unde
        - p_i – proporţia exemplelor din clasa i în setul de date S
    - câştigul de informaţie (*information gain*) al unei caracterisitici a (al unui atribut al) datelor
      - Reducerea entropiei setului de date ca urmare a eliminării atributului *a*
      - $Gain(S, a) = E(S) - \sum_{v \in valori(a)} |S_v| / |S| \; E(S_v)$
      - $\sum_{v \in valori(a)} |S_v| / |S| \; E(S_v)$ – *informaţia scontată*

Sisteme inteligente – SIS – Arbori de decizie

- Proces → Construirea AD → Algoritmul ID3/C4.5 → Selectare atribut
  - Câştigul de informaţie
    - exemplu

	a1	a2	a3	Clasa
d1	mare	roşu	cerc	clasa 1
d2	mic	roşu	pătrat	clasa 2
d3	mic	roşu	cerc	clasa 1
d4	mare	albastru	cerc	clasa 2

$S = \{d1, d2, d3, d4\} \rightarrow p_+ = 2/4,\ p_- = 2/4 \rightarrow E(S) = -p_+ \log_2 p_+ - p_- \log_2 p_- = 1$

$S_{v=mare} = \{d1, d4\} \rightarrow p_+ = ½,\ p_- = ½ \rightarrow E(S_{v=mare}) = 1$

$S_{v=mic} = \{d2, d3\} \rightarrow p_+ = ½,\ p_- = ½ \rightarrow E(S_{v=mic}) = 1$

$S_{v=rosu} = \{d1, d2, d3\} \rightarrow p+ = 2/3,\ p- = 1/3 \rightarrow E(S_{v=rosu}) = 0.923$

$S_{v=albastru} = \{d4\} \rightarrow p+ = 0,\ p- = 1 \rightarrow E(S_{v=albastru}) = 0$

$S_{v=cerc} = \{d1, d3, d4\} \rightarrow p+ = 2/3,\ p- = 1/3 \rightarrow E(S_{v=cerc}) = 0.923$

$S_{v=patrat} = \{d2\} \rightarrow p+ = 0,\ p- = 1 \rightarrow E(S_{v=patrat}) = 0$

$Gain(S, a) = E(S) - \sum_{v \in valori(a)} |S_v| / |S| \, E(S_v)$

$Gain(S, a_1) = 1 - (|S_{v=mare}| / |S| \, E(S_{v=mare}) + |S_{v=mic}| / |S| \, E(S_{v=mic})) = 1 - (2/4 * 1 + 2/4 * 1) = 0$

$Gain(S, a_2) = 1 - (|S_{v=rosu}| / |S| \, E(S_{v=rosu}) + |S_{v=albastru}| / |S| \, E(S_{v=albastru})) = 1 - (3/4 * 0.923 + 1/4 * 0) = 0.307$

$Gain(S, a_3) = 1 - (|S_{v=cerc}| / |S| \, E(S_{v=cerc}) + |S_{v=patrat}| / |S| \, E(S_{v=patrat})) = 1 - (3/4 * 0.923 + 1/4 * 0) = 0.307$

Sisteme inteligente – SIS – Arbori de decizie

- Proces → Construirea AD → Algoritmul ID3/C4.5 → Selectare atribut
  - Rata câştigului
    - Penalizează un atribut prin încorporarea unui termen – *split information* – sensibil la gradul de împrăştiere şi uniformitate în care atributul separă datele
      - *Split information* – entropia relativ la valorile posibile ale atributului *a*
      - *Sv* – proporţia exemplelor din setul de date *S* care au atributul *a* evaluat cu valoarea *v*
      - *splitInformation(S,a)=* $$-\sum_{v=value(a)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$$

- Studiu de caz - Problema vampirilor

Items	Shadow	Complexion	Garlic	Accent	Vampire
i1	?	Pale	Yes	None	No
i2	Yes	Ruddy	Yes	None	No
i3	?	Ruddy	No	None	Yes
i4	No	Average	No	Heavy	Yes
i5	?	Average	No	Odd	Yes
i6	Yes	Pale	No	Heavy	No
i7	Yes	Average	No	Heavy	No
i8	?	Ruddy	Yes	Odd	No

Sisteme inteligente – SIS – Arbori de decizie

- Proces
  - Construirea arborelui
  - Utilizarea arborelui ca model de rezolvare a problemelor
    - Ideea de bază
      - Se extrag regulile formate în arborele anterior construit → Reguli extrase din arborele dat în exemplul anterior:
        - IF *age* = "*<=30"* AND *student* = *"no"* THEN *buys_computer* = *"no"*
        - IF *age* = "*<=30"* AND *student* = *"yes"* THEN *buys_computer* = *"yes"*
        - IF *age* = *"31…40"* THEN *buys_computer* = *"yes"*
        - IF *age* = "*>40"* AND *credit_rating* = *"excellent"* THEN *buys_computer* = *"no"*
        - IF *age* = "*>40"* AND *credit_rating* = *"fair"* THEN *buys_computer* = *"yes"*
      - Regulile sunt folosite pentru a clasifica datele de test (date noi). Fie *x* o dată pentru care nu se ştie clasa de apartenenţă → Regulile se pot scrie sub forma unor predicate astfel:
        - IF *age(x, <=30)* AND *student(x, no)* THEN *buys_computer (x, no)*
        - IF *age(x, <=30)* AND *student (x, yes)* THEN *buys_computer (x, yes)*

Sisteme inteligente – SIS – Arbori de decizie

- Proces
  - Construirea arborelui
  - Utilizarea arborelui ca model de rezolvare a problemelor
    - Dificultăţi
      - *Underfitting* (sub-potrivire) → AD indus pe baza datelor de antrenament este prea simplu ➔ eroare de clasificare mare atât în etapa de antrenare, cât şi în cea de testare
      - *Overfitting* (supra-potrivire, învăţare pe derost) → AD indus pe baza datelor de antrenament se potriveşte prea accentuat cu datele de antrenament, nefiind capabil să generalizeze pentru date noi
      - Soluţii:
        - fasonarea arborelui (pruning) → Îndepărtarea ramurilor nesemnificative, redundante ➔ arbore mai puţin stufos
        - validare cu încrucişare

Sisteme inteligente – SIS – Arbori de decizie

- Proces
  - Construirea arborelui
  - Utilizarea arborelui ca model de rezolvare a problemelor
  - Tăierea (fasonarea) arborelui
    - Necesitate
      - Odată construit AD, se pot extrage reguli (de clasificare) din AD pentru a putea reprezenta cunoştinţele sub forma regulilor *if-then* atât de uşor de înţeles de către oameni
      - O regulă este creată (extrasă) prin parcurgerea AD de la rădăcină până la o frunză
      - Fiecare pereche (atribut,valoare), adică (nod, muchie), formează o conjuncţie în ipoteza regulii (partea dacă), mai puţin ultimul nod din drumul parcurs care este o frunză şi reprezintă consecinţa (ieşirea, partea atunci) regulii
    - Tipologie
      - Prealabilă (*pre-pruning*)
        - Se opreşte creşterea arborelui în timpul inducţiei prin sistarea divizării unor noduri care devin astfel frunze etichetate cu clasa majoritară a exemplelor aferente nodului respectiv
      - Ulterioară (*post-pruning*)
        - După ce AD a fost creat (a crescut) se elimină ramurile unor noduri care devin astfel frunze ➔ se reduce eroarea de clasificare (pe datele de test)

Sisteme inteligente – SIS – Arbori de decizie

- Tool-uri
  - http://webdocs.cs.ualberta.ca/~aixplore/learning/DecisionTrees/Applet/DecisionTreeApplet.html
  - WEKA → J48
  - http://id3alg.altervista.org/
  - http://www.rulequest.com/Personal/c4.5r8.tar.gz
- Biblio
  - http://www.public.asu.edu/~kirkwood/DAStuff/decisiontrees/index.html

Sisteme inteligente – SIS – Arbori de decizie

- Avantaje
  - Uşor de înţeles şi interpretat
  - Permit utilizarea datelor nominale şi categoriale
  - Logica deciziei poate fi urmărită uşor, regulile fiind vizibile
  - Lucrează bine cu seturi mari de date

- Dezavantaje
  - Instabilitate → modificarea datelor de antrenament
  - Complexitate → reprezentare
  - Greu de manevrat
  - Costuri mari pt inducerea AD
  - Inducerea AD necesită multă informaţie

Sisteme inteligente – SIS – Arbori de decizie

- Dificultăţi
  - Existenţa mai multor arbori
    - Cât mai mici
    - Cu o acurateţe cât mai mare (uşor de "citit" şi cu performanţe bune)
    - Găsirea celui mai bun arbore → problemă NP-dificilă
  - Alegerea celui mai bun arbore
    - Algoritmi euristici
    - ID3 → cel mai mic arbore acceptabil
      - → teorema lui Occam: "always choose the simplest explanation"
  - Atribute continue
    - Separarea în intervale
      - Câte intervale?
      - Cât de mari sunt intervalele?
  - Arbori prea adânci sau prea stufoşi
    - Fasonarea prealabilă (pre-pruning) → oprirea construirii arborelui mai devreme
    - Fasonarea ulterioară (post-pruning) → înlăturarea anumitor ramuri

Recapitulare

- Sisteme care învaţă singure (SIS)
  - Regresie logistică
  - kNN
  - Arbori de decizie
    - Fiecare nod intern corespunde unui atribut
    - Fiecare ramură de sub un nod (atribut) corespunde unei valori a atributului
    - Fiecare frunză corespunde unei clase (etichete) – conţine toate datele din acea clasă
  - Mașini cu suport vectorial
    - Hiperplan de separare
    - Funcții kernel

Cursul următor

A. Scurtă introducere în Inteligenţa Artificială (IA)

B. Rezolvarea problemelor prin căutare
- Definirea problemelor de căutare
- Strategii de căutare
  - Strategii de căutare neinformate
  - Strategii de căutare informate
  - Strategii de căutare locale (Hill Climbing, Simulated Annealing, Tabu Search, Algoritmi evolutivi, PSO, ACO)
  - Strategii de căutare adversială

C. Sisteme inteligente
- Sisteme care învaţă singure
  - Arbori de decizie
  - Reţele neuronale artificiale
  - Maşini cu suport vectorial
  - **Algoritmi evolutivi**
- Sisteme bazate pe reguli
- Sisteme hibride

Cursul următor – Materiale de citit şi legături utile

- Capitolul VI (19) din *S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, 1995*
- capitolul 8 din *Adrian A. Hopgood, Intelligent Systems for Engineers and Scientists, CRC Press, 2001*
- capitolul 12 şi 13 din *C. Groşan, A. Abraham, Intelligent Systems: A Modern Approach, Springer, 2011*
- Capitolul V din *D. J. C. MacKey, Information Theory, Inference and Learning Algorithms, Cambridge University Press, 2003*
- Capitolul 4 din *T. M. Mitchell, Machine Learning, McGraw-Hill Science, 1997*

- Informaţiile prezentate au fost colectate din diferite surse de pe internet, precum şi din cursurile de inteligenţă artificială ţinute în anii anteriori de către:
  - Conf. Dr. Mihai Oltean – www.cs.ubbcluj.ro/~moltean
  - Lect. Dr. Crina Groşan - www.cs.ubbcluj.ro/~cgrosan
  - Prof. Dr. Horia F. Pop - www.cs.ubbcluj.ro/~hfpop