Curs 4:

Tehnici de clasificare a datelor

Arbori de decizie și reguli de clasificare

Reminder

- Modele de clasificare:
  - Modele bazate pe reguli
  - Modele bazate pe instanțe
  - Modele probabiliste
  - Modele bazate pe funcții

- In construirea modelelor datele disponibile se împart în:
  - Set de antrenare – folosit în estimarea parametrilor modelului
  - Set de validare – folosit în alegerea caracteristicilor modelului (hiper-parametri)
  - Set de testare – folosit în evaluarea performanței

- Evaluarea performanței unui clasificator:
  - Construire matrice de confuzie
  - Calcul măsuri de performanță: acuratețe, senzitivitate / specificitate, precision / recall, F-score, AUC (area under receiver-operator-characteristic curve)

Structura

- Arbori de decizie
  - Criterii de ramificare
  - Algoritmi de construire
- Reguli de clasificare
  - Caracteristici ale unui set de reguli
  - Algoritmi de extragere a regulilor din date

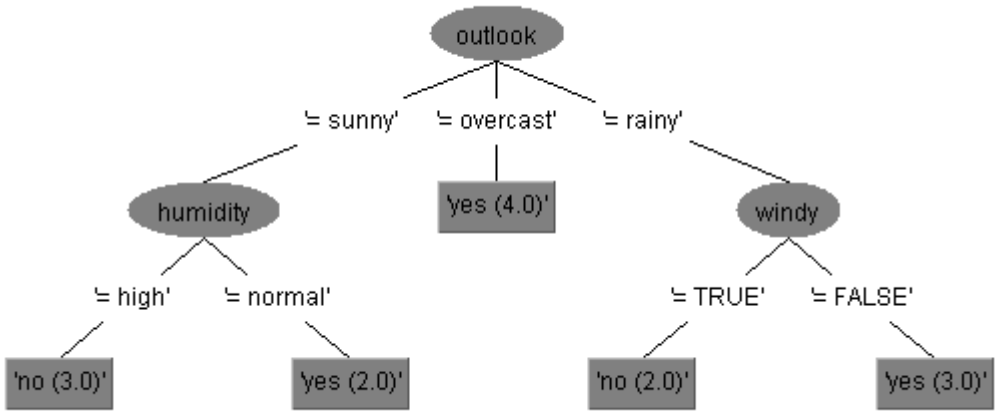
Arbori de decizie

Set de date: weather/play

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Arbore de decizie (construit folosind Weka)



Cum poate fi utilizat un arbore de decizie?

Ce clasă corespunde unei noi instanţe?
(outlook=sunny, temperature=mild, humidity=normal, windy=False)?

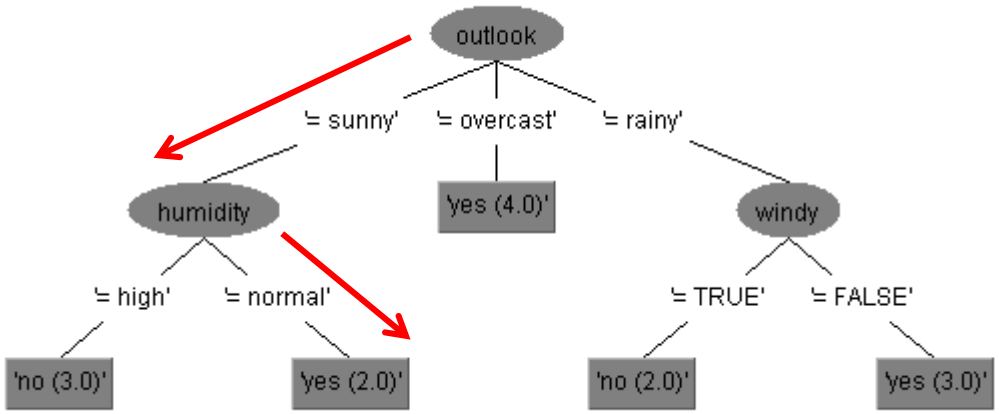
Arbori de decizie

Set de date: weather/play

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Arbore de decizie (construit folosind Weka)



Cum poate fi utilizat un arbore de decizie?

Ce clasă corespunde unei noi instanţe?
(outlook=sunny, temperature=mild, humidity=normal, windy=False)?
Clasa: Yes (clasa dominantă asociată nodului frunză)

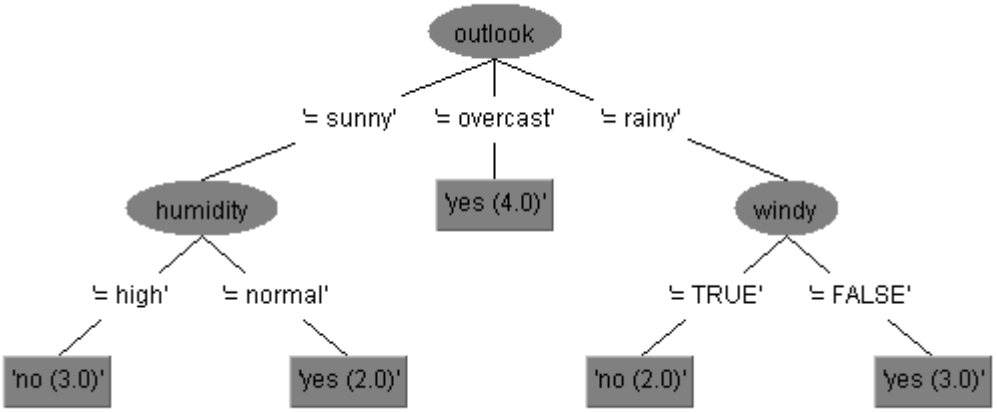
Arbori de decizie

Set de date: weather/play

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Arbore de decizie (construit folosind Weka)



Cum poate fi tradus într-un set de reguli de clasificare? Fiecare ramură conduce la o regulă

Regula 1: IF outlook=sunny and humidity=high THEN play=no
Regula 2: IF outlook=sunny and humidity=normal THEN play=yes
Regula 3: IF outlook=overcast THEN play=yes
Regula 4: IF outlook=rainy and windy=True THEN play=no
Regula 5: IF outlook=rainy and windy=False THEN play=yes

Arbori de decizie

Cum poate fi construit un arbore de decizie pornind de la date?

- Se alege un atribut şi se plasează în rădăcina arborelului
- Pt fiecare valoare posibilă a atributului (dintre cele prezente în setul de date) se construieşte o ramură
- Se partiţionează setul de date în subseturi corespunzătoare fiecărei ramuri
  - Dacă un subset conţine date ce aparţin unei singure clase atunci el va corespunde unui nod frunză (nu se mai ramifică) – nod pur
  - Dacă subsetul conţine date din mai multe clase atunci se continuă procesul de partiţionare până când
    - se ajunge la un nod pur
    - pe ramura respectivă au fost deja analizate toate atributele
    - subsetul corespunzător ramurii este vid

Weather/play dataset

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Problemă: în ce ordine ar trebui analizate atributele? Ce condiţie de testare ar trebui asociată cu fiecare nod?

Arbori de decizie

In ce ordine ar trebui analizate atributele?

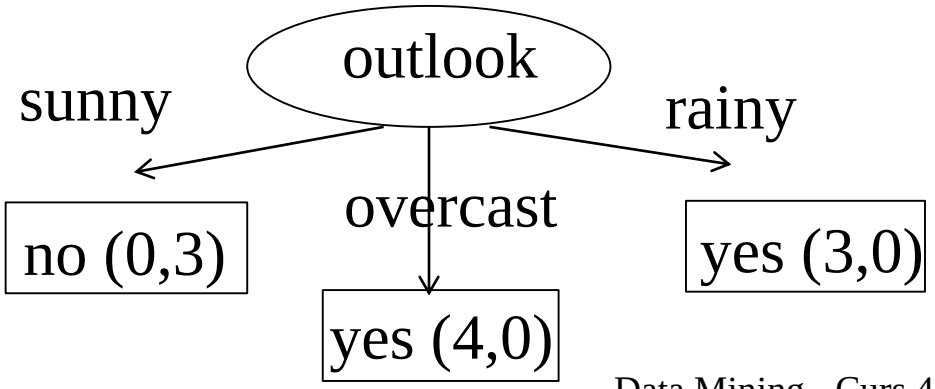
Ideea principală:

- Se selectează atributul care conduce la un arbore cât mai simplu adică un atribut cu grad de puritate cât mai mare (ideal ar fi ca pentru fiecare valoare a atributului datele corespunzătoare să aparţină aceleiaşi clase)

Set de date: weather/play (date selectate)

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	overcast	cool	normal	TRUE	yes
7	sunny	mild	high	FALSE	no
8	rainy	mild	normal	FALSE	yes
9	overcast	mild	high	TRUE	yes
10	overcast	hot	normal	FALSE	yes

Exemplu:



Obs:

- Toate nodurile frunză sunt “pure” (conţin date ce aparţin aceleiaşi clase)
- Conduce la reguli cu un singur atribut în membrul stâng
- O astfel de situaţie se întâmplă rar pentru date reale

Arbori de decizie

Principalele probleme ce trebuie soluţionate la construirea unui arbore de decizie

- Ce atribut ar trebui selectat pentru partiţionare?
  - Cel cu puterea cea mai mare de discriminare = cel ce asigură partiţionarea setului curent în subseturi cu grad mare de puritate
  - Criterii ce pot fi utilizate:
    - Bazate pe entropie (ex: câştig informaţional)
    - Index Gini
    - Măsură a erorii de clasificare

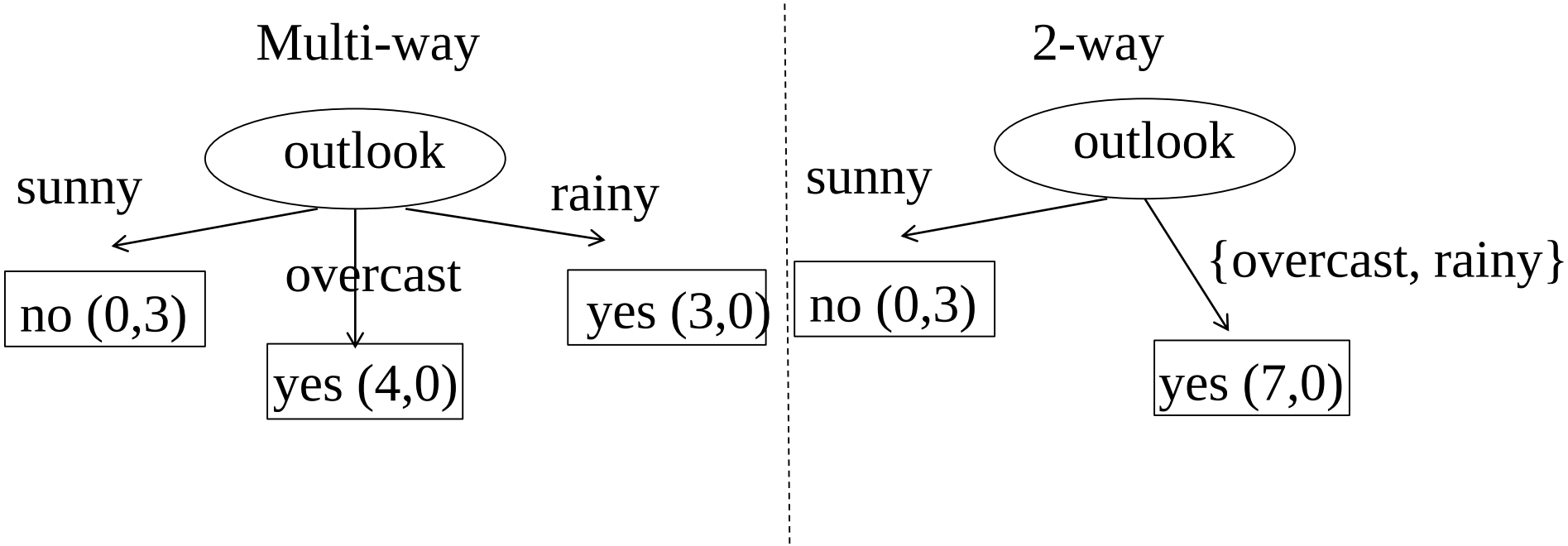
- Ce condiţii de test trebuie asignate ramurilor corespunzătoare unui nod ?
  - Depinde de tipul atributului
    - Nominal, ordinal: atribut = valoare
    - Continuu: atribut < valoare sau valoare1< atribut < valoare2
  - Depinde de gradul de ramificare dorit:
    - Ramificare binară (setul curent de date este împărţit în două subseturi)
    - Ramificare multiplă (setul curent de date este împărţit în mai multe subseturi)

Arbori de decizie

- Ce condiţii de test trebuie asignate ramurilor corespunzătoare unui nod ?

Atribute nominale şi ordinale:

- Ramificare multiplă (multi-way): atâtea ramuri câte valori posibile are atributul
- Ramificare binară (2-way): două ramuri



Arbori de decizie

Ce condiţii de test trebuie asignate ramurilor corespunzătoare unui nod?

Atribute numerice:

- Trebuie discretizate în prealabil, după care se aplică strategia specifică atributelor nominale sau ordinale

Ce atribut se selectează pentru partiţionare?

- Acel atribut care conduce la reducerea maximă în conţinutul de informaţie necesar pentru a lua decizia corectă

Exemplu: information gain

Câştig informaţional =
Entropia(distribuţia datelor înainte de partiţionare) –
EntropiaMedie (distribuţia datelor după partiţionare)

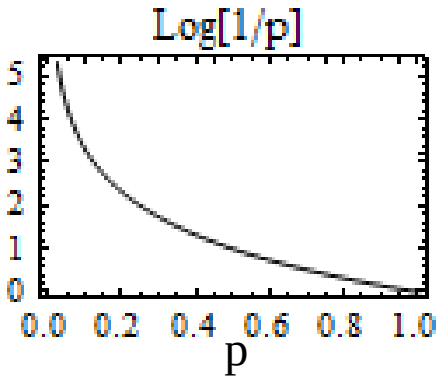
Reminder: entropie

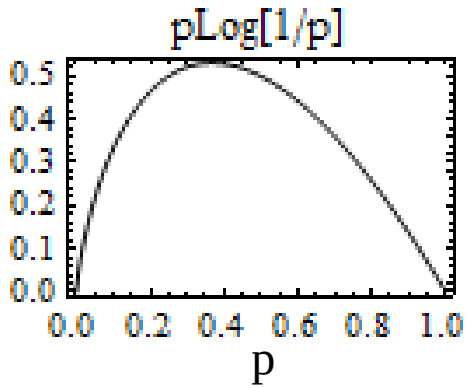
Fie D=(p_1,p_2, …, p_k) o distribuţie de probabilitate. Entropia asociată acestei distribuţii de probabilitate este caracterizată de:

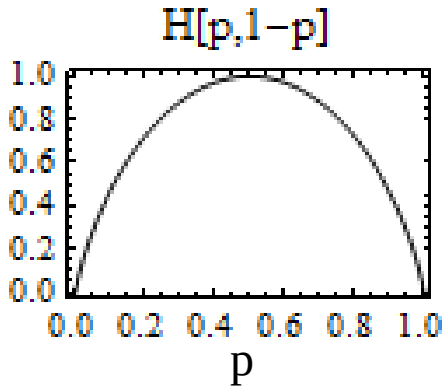
$$H(D) = H(p_1, p_2, ..., p_k) = -\sum_{i=1}^{k} p_i \log p_i$$

şi poate fi interpretată ca o măsură medie a incertitudinii (sau surprizei) când se generează/selectează o valoare pe baza acestei distribuţii

Caz particular: k=2 => p_1=p, p_2=1-p







Obs: Interpretare Log[1/p] : surpriza de a observa un eveniment caracterizat de o probabilitate mică (eveniment neaşteptat) este mai mare decât cea corespunzătoare unui eveniment de probabilitate mai mare (eveniment aşteptat)

Reminder: entropie

In contextul rezolvării problemelor de clasificare:

- $D=\{C_1,C_2,\ldots,C_k\}$ (set de date distribuit în k clase)
- Distribuţia de probabilitate $(p_1,p_2,\ldots,p_k)$, $p_i=card(C_i)/card(D)$
- Fie A un atribut şi $v_1,v_2,\ldots,v_{mA}$ valorile posibile ale acestui atribut
- Fie D_j=setul de date din D pt care atributul A are valoarea v_j şi P_j distribuţia datelor din D_j în cele k clase (C_{ji}= set de date din clasa C_i care au valoarea v_j pt atributul A)
- Câştigul informaţional obţinut prin partiţionarea setului de date folosind atributul A este:

$$IG(D,A)=H(D)-\sum_{j=1}^{m_A}P(D_j\mid A=v_j)H(D_j\mid A=v_j),\quad H(D)=-\sum_{i=1}^{k}p_i\log p_i$$

$$H(D_j\mid A=v_j)=-\sum_{i=1}^{k}p_{ij}\log p_{ij},\quad p_{ij}=\frac{card(C_{ji})}{card(C_i)}$$

$$P(D_j\mid A=v_j)=\frac{card(D_j)}{card(D)}$$

Alegerea atributului de ramificare

Exemplu

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

- Distribuţia claselor (C_1="yes", C_2="no"):
- p_1=9/14, p_2=5/14
- $H(p_1,p_2)$=0.94

Outlook

	C1 (yes)	C2(no)	Frequency
Sunny	2/5	3/5	5/14
Overcast	4/4	0/4	4/14
Rainy	3/5	2/5	5/14

H(sunny)=-2/5*log(2/5)-3/5*log(3/5)=0.97
H(overcast)=-1*log(1)-0=0
H(rainy)=-3/5*log(3/5)-2/5*log(2/5)=0.97

IG(outlook)=0.94-5/14*0.97-4/14*0-5/14*0.97=0.94-0.69=0.25

Alegerea atributului de ramificare

Exemplu

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

- Distribuţia claselor(C_1="yes", C_2="no"):
- p_1=9/14, p_2=5/14
- $H(p_1,p_2)$=0.94

Temperature

	C1 (yes)	C2(no)	Frequency
Hot	2/4	2/4	4/14
Mild	4/6	2/6	6/14
Cool	3/4	1/4	4/14

H(hot)=-2/4*log(2/4)-2/4*log(2/4)
H(mild)=-4/6*log(4/6)-2/6*log(2/6)
H(cool)=-3/4*log(3/4)-1/4*log(1/4)

IG(temperature)=0.03

Alegerea atributului de ramificare

Exemplu

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

- Distribuţia claselor (C_1="yes", C_2="no"):
- p_1=9/14, p_2=5/14
- $H(p_1,p_2)$=0.94

Humidity

	C1 (yes)	C2(no)	Frequency
High	3/7	4/7	7/14
Normal	6/7	1/7	7/14

H(high)=-3/7*log(3/7)-4/7*log(4/7)
H(normal)=-6/7*log(6/7)-1/7*log(1/7)

IG(humidity)=0.15

Alegerea atributului de ramificare

Exemplu

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

H(false)=-6/8*log(6/8)-2/8*log(2/8)
H(true)=-3/6*log(3/6)-3/6*log(3/6)

IG(windy)=0.05

- Distribuţia claselor(C_1="yes", C_2="no"):
- p_1=9/14, p_2=5/14
- H(p_1,p_2)=0.94

Windy

	C1 (yes)	C2(no)	Frequency
False	6/8	2/8	8/14
True	3/6	3/6	6/14

Alegerea atributului de ramificare

Exemplu

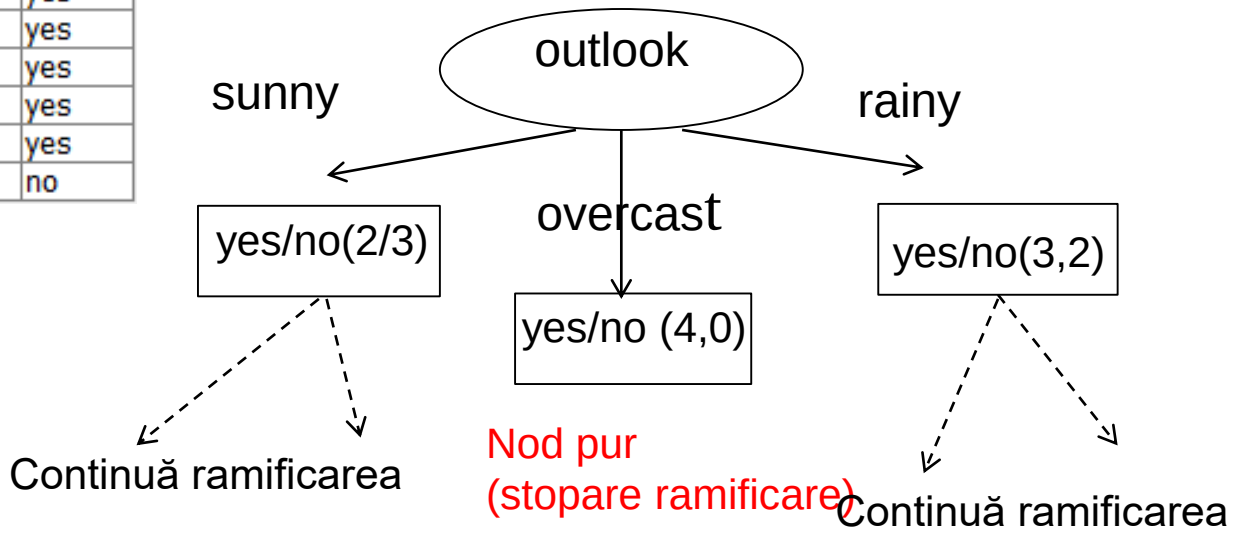
Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Câştigul informaţional al fiecărui atribut:

- IG(outlook)=0.25
- IG(temperature)=0.03
- IG(humidity)=0.15
- IG(windy)=0.05

Primul atribut selectat: outlook

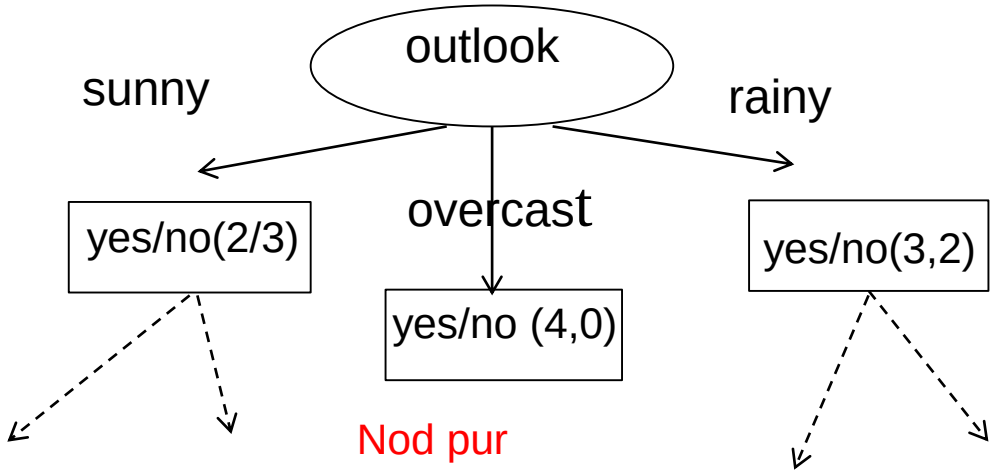


Alegerea atributului de ramificare

Exemplu

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no



Nod pur
(stopare ramificare)

Câştig informaţional pt atributele rămase:

- Entropia pt subsetul “sunny” :
- H(D(sunny))=-2/5*log(2/5)-3/5*log(3/5)=0.97
- H(hot)=0, H(mild)=1, H(cool)=0
- IG(temperature)=0.97-2/5=0.57

Temperature

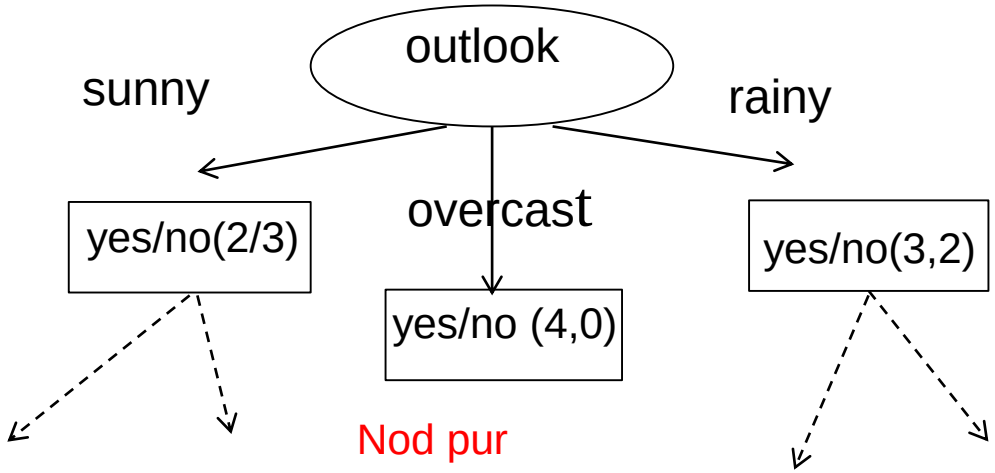
	C1 (yes)	C2 (no)	Freq.
Hot	0/2	2/2	2/5
Mild	1/2	1/2	2/5
Cool	1/1	0/1	1/5

Alegerea atributului de ramificare

Exemplu

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no



Nod pur
(stopare ramificare)

Câştig informaţional pt atributele rămase:

- Entropia pt subsetul “sunny” :

H(D(sunny))=-2/5*log(2/5)-3/5*log(3/5)=0.97

- H(high)=0, H(normal)=0

- IG(humidity)=0.97-0=0.97

Humidity

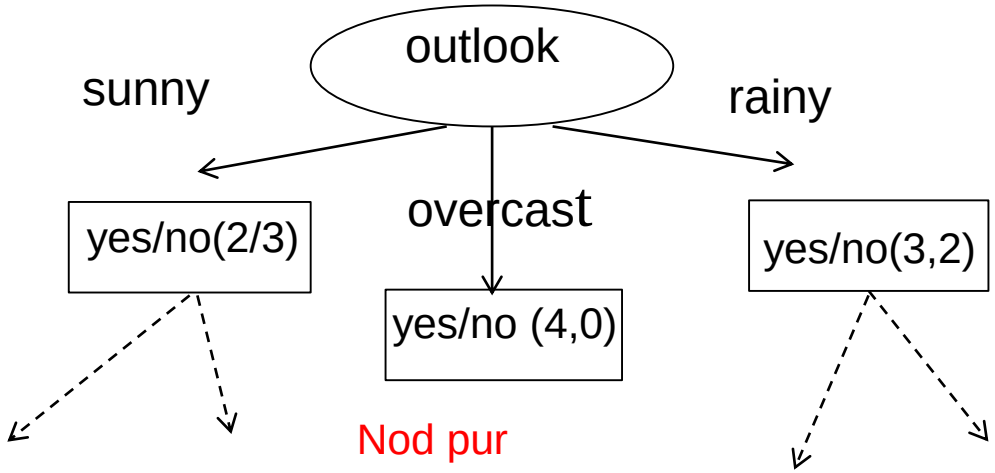
	C1 (yes)	C2 (no)	Freq.
High	0/3	3/3	3/5
Normal	2/2	0/2	2/5

Alegerea atributului de ramificare

Exemplu

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no



Nod pur
(stopare ramificare)

Câştig informaţional pt atributele rămase:

- Entropia pt subsetul “sunny” :
  H(D(sunny))=-2/5*log(2/5)-3/5*log(3/5)=0.97

- H(false)=0, H(true)=1

- IG(windy)=0.97-0.95=0.02

Windy

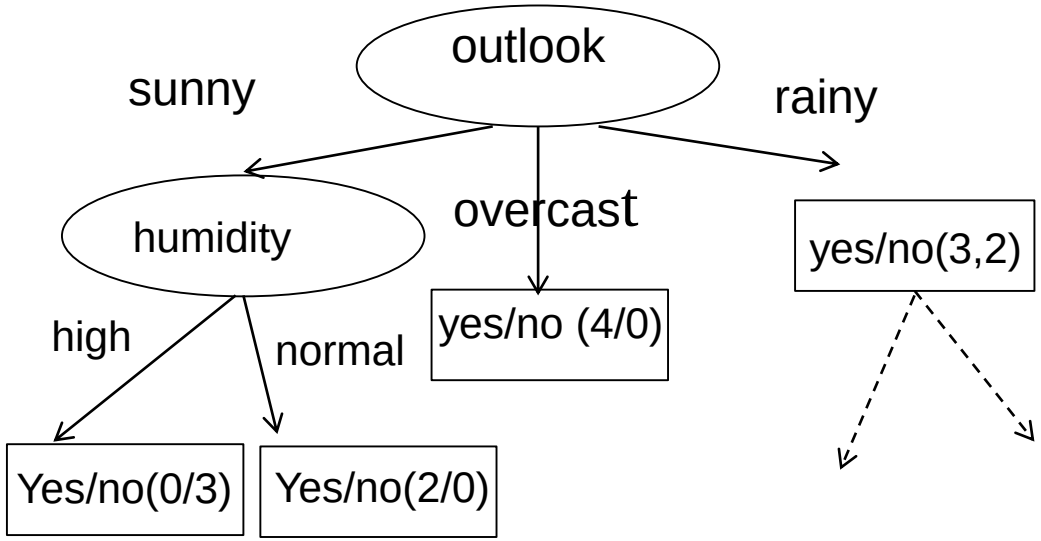
	C1 (yes)	C2 (no)	Freq.
false	1/3	2/3	3/5
true	1/2	1/2	2/5

Alegerea atributului de ramificare

Exemplu

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no



Câştig informaţional pt atributele rămase:

- IG(temperature)=0.97-2/5=0.57
- IG(humidity)=0.97-0=0.97
- IG(windy)=0.97-0.95=0.02

Alegerea atributului de ramificare

Obs:

- Câştigul informaţional favorizează atributele caracterizate printr-un număr mare de valori

- Pentru a limita această influenţă se poate utiliza raportul (Gain Ratio):

$$GainRatio(D, A) = \frac{IG(D, A)}{H(p_1^A, p_2^A, ..., p_{m_A}^A)}$$

$$p_j^A = \frac{card(D, A = v_j)}{card(D)}$$

(proportia de date care au valoarea v_j pt atributul A)

Alegerea atributului de ramificare

- Atributul de ramificare poate fi determinat folosind indexul Gini = măsură a gradului mediu de impuritate a subseturilor de date obținute prin ramificarea bazată pe atributul A (cu cât mai mică valoarea cu atât mai bună)

index Gini pt atributul A_i

$$G(A_i) = \frac{1}{N}\sum_{j=1}^{r_i} n_{ij} G(v_{ij}), \qquad G(v_{ij}) = 1 - \sum_{k=1}^{K} p_{ijk}^2$$

n_{ij} = numarul de instante pt care A_i are valoarea v_{ij}

$$p_{ijk} = \frac{\text{numar de instante in } C_k \text{ cu } A_i = v_{ij}}{\text{numar de instante cu } A_i = v_{ij}}$$

Algoritmi pentru construirea arborilor de decizie

ID3 (Iterative Dichotomiser):

- Intrare: set de date D
- Ieșire: arbore de decizie (noduri interne etichetate cu atribute, noduri frunză etichetate cu clase, muchii etichetate cu valori ale atributelor)

DTinduction (D, DT, N) /* D=set date, DT=arbore de decizie, N=nod */
    find the best splitting attribute A
    label node N with A
    construct the splitting predicates (branches) for N
    FOR each branch i from N DO
        construct the corresponding data set D_i
        create a new child node N_i
        IF <stopping condition>
            THEN label N_i with the dominant class in D_i (N_i is a leaf node)
            ELSE DTinduction(D_i,DT, N_i)

Algoritmi pentru construirea arborilor de decizie

C4.5 = îmbunătăţire a algoritmului ID3 pt a trata:

- Atribute continue:
  - Incorporează procedură de discretizare
- Valori absente:
  - Datele ce conţin valori absente sunt ignorate

sau

  - Valorile absente sunt imputate

- Atribut de ramificare:
  - Utilizează Gain Ratio pt selecţia atributului

- Simplificare sau fasonare (Pruning):
  - Anumiţi subarbori sunt înlocuiţi cu noduri frunză (dacă eroarea de clasificare nu creşte semnificativ) – abordare bottom-up

Obs:

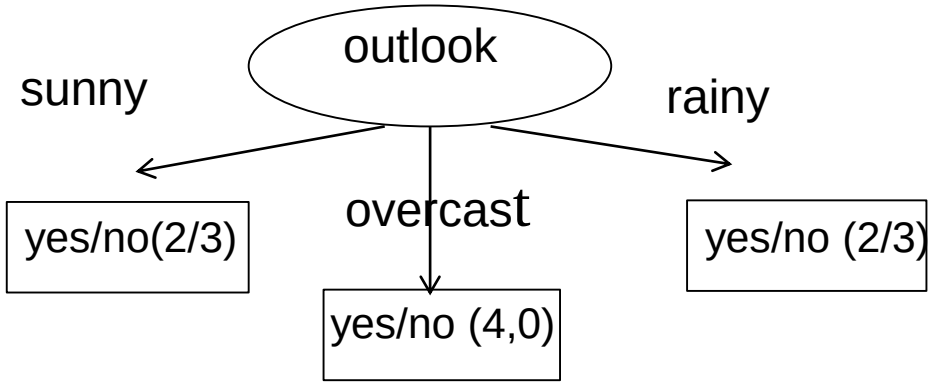
C5.0 – varianta comerciala a algoritmului C4.5 – accesibil pt analiza din 2011

J48 – implementarea din Weka a algoritmului C4.5

Algoritmi pentru construirea arborilor de decizie

Simplificare (pruning):

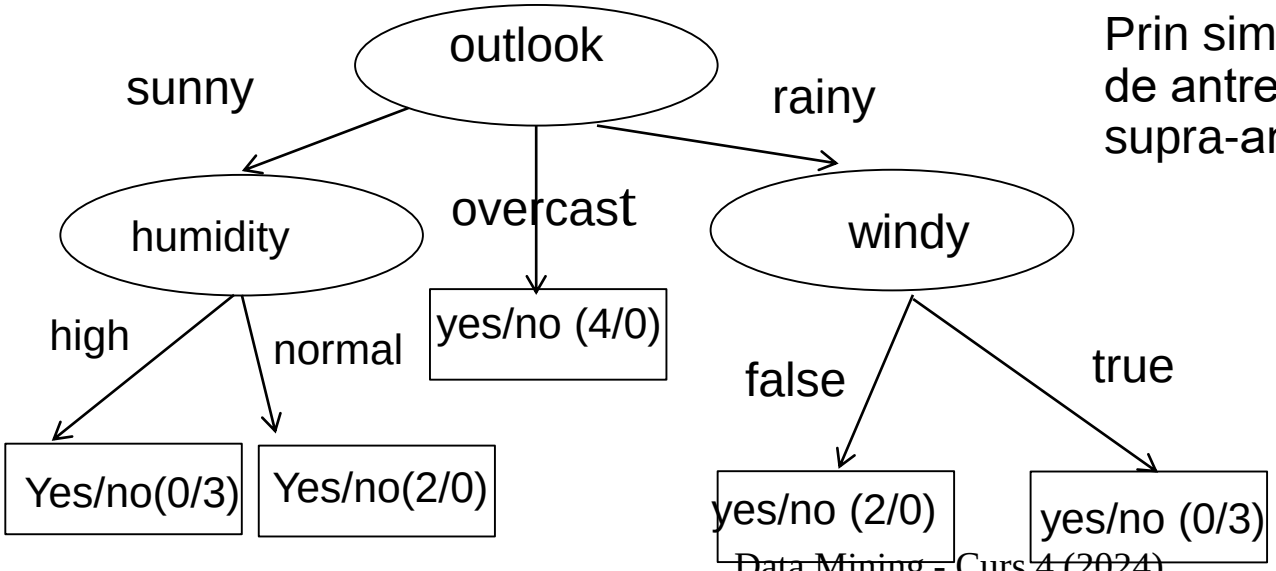
- Anumiţi subarbori sunt înlocuiţi cu noduri frunză (dacă eroarea de clasificare nu creşte semnificativ) – abordare bottom-up



Arbore : error = 0

Arbore simplificat: error = 4/14

Prin simplificare eroarea pe setul de antrenare creşte dar riscul de supra-antrenare poate să scadă



Extragerea regulilor de clasificare

Reminder: regulile de clasificare sunt structuri de tip IF … THEN care conţin:

- In partea de antecedent (membrul stâng): condiţii privind valorile atributelor (pot fi expresii logice care implică mai multe atribute)
- In partea de consecinţă (membrul drept): eticheta clasei

Exemple:

IF outlook=sunny THEN play=no

IF outlook=rainy THEN play=no

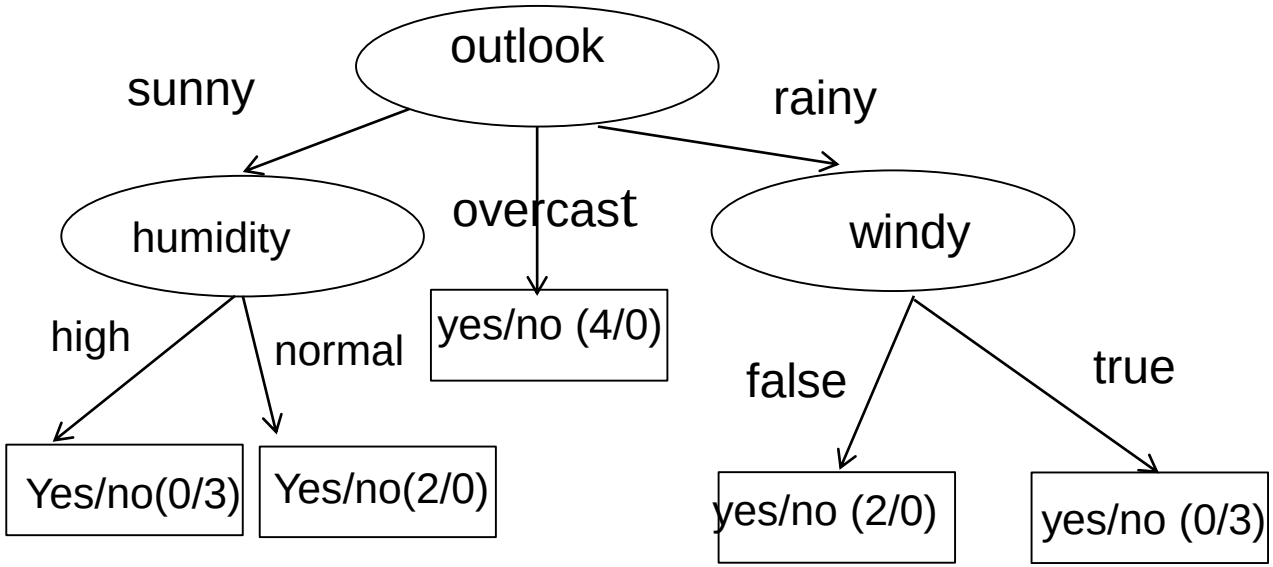
IF outlook=overcast THEN play=yes

Extragerea regulilor de clasificare

- Dacă regulile sunt extrase dintr-un arbore de decizie atunci fiecare ramură conduce la o regulă
- Condiţiile referitoare la noduri aflate pe aceaşi ramură se combină prin AND:

  IF (outlook=sunny) and (humidity=high) THEN play=no
- Regulile corespunzând unor ramuri diferite dar conducând la aceeaşi consecinţă (aceeaşi etichetă de clasă) pot fi reunite prin disjuncţie (OR) între părţile de antecedent:

IF (outlook=sunny and humidity=high) OR (outlook=rainy and windy =true) THEN play=no



Extragerea regulilor de clasificare

Alta variantă: Regulile de clasificare pot fi extrase direct din date printr-un proces de învăţare utilizând algoritmi de acoperire (covering algorithms)

Noţiuni:

- O regulă acoperă o dată dacă valorile atributelor se potrivesc cu condiţiile din antecedentul regulii
- Similar, despre o dată se spune că activează o regulă dacă valorile atributelor se potrivesc cu condiţiile din antecedentul regulii

- Suportul unei reguli (support) = fracţiunea din setul total de date care este acoperită de către regulă şi aparţine aceleiaşi clase ca cea din regulă = |cover(R) ∩class(R) |/|D|

- Gradul de încredere în regulă (rule confidence) = fracţiunea din datele acoperite de regulă care au aceeaşi clasă ca cea specificată de regulă = |cover(R) ∩class(R)|/|cover(R)|

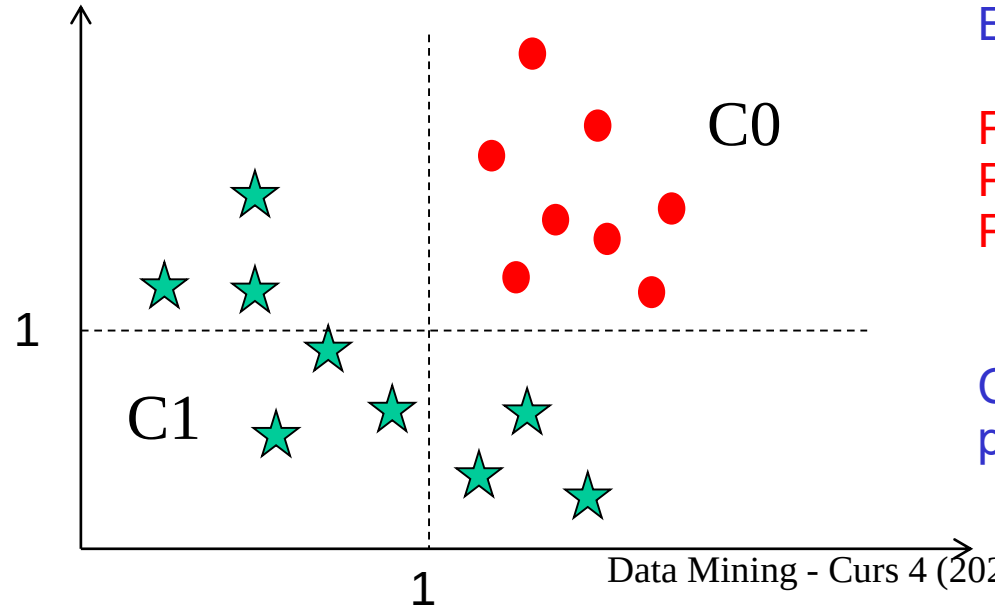
  - cover(R) = subsetul de date acoperit de R
  - class(R) = subsetul de date care au aceeaşi clasă cu cea specificată în R
  - D= setul de date

Extragerea regulilor de clasificare

Noţiuni:

- Reguli mutual exclusive = regiunile acoperite de reguli sunt disjuncte (o instanţă activează o singură regulă)
- Set complet de reguli = fiecare instanţă activează cel puţin o regulă

Obs: dacă setul de reguli e complet şi regulile sunt mutual exclusive atunci decizia privind apartenenţa unei date la o clasă este simplu de luat



Exemplu:

R1: IF x>1 and y>1 THEN C0
R2: IF x<=1 THEN C1
R3: IF x>1 and y<=1 THEN C1

Ce se întâmplă însă dacă aceste proprietăţi nu sunt satisfăcute?

Extragerea regulilor de clasificare

Obs: dacă regulile nu sunt mutual exclusive atunci pot să apară conflicte (o instanţă poate activa reguli care au asociate clase diferite)

Conflictele pot fi rezolvate în unul dintre următoarele moduri:

- Ordonarea regulilor (pe baza unui criteriu) şi decizia se ia conform primei reguli activate (prima regulă care se potriveşte cu instanţa).

- Criteriul de ordonare poate fi corelat cu:
  - calitatea regulii (e.g. grad de încredere) – regulile cu grad mai mare de încredere sunt mai bune

  - specificitatea regulii – o regulă este considerată mai bună dacă este mai specifică (e.g. reguli care corespund claselor rare)

  - complexitatea regulii (e.g. numărul de condiţii din partea de antecedent a regulii) – regulile mai simple sunt mai bune

Extragerea regulilor de clasificare

Obs: aceste criterii pot fi conflictuale (o regulă cu coeficient mare de încredere nu este neapărat o regulă simplă)

- Rezultatul se obţine considerând clasa dominantă pe baza tuturor regulilor activate de către instanţă

Extragerea regulilor de clasificare

Algoritm secvenţial de acoperire:

Intrare: set de date
Ieşire: set ordonat de reguli

Pas 1: se selectează una dintre clase şi se identifică cea mai “bună” regulă care acoperă datele din D care au clasa selectată. Se adaugă regula la lista de reguli.

Pas 2: Se elimină datele din D care activează regula adăugată. Dacă încă există clase netratate şi date în D se reia de la Pas 1

Obs:

- Aceasta este structura generală a algoritmilor secvenţiali de acoperire
- Algoritmii pot să difere în funcţie de strategia de selecţie a claselor

Extragerea regulilor de clasificare

Exemplu: algoritmul RIPPER (Repeated Incremental Pruning to Produce Error Reduction)

Particularităţi:

- Setul de date e divizat la început în growing set (folosit pentru construirea unui set de reguli care acoperă setul de date) şi pruning set (folosit pt simplificarea regulilor, de ex. prin eliminarea unor atribute din membrul stâng al regulii - se alege varianta de simplificare care reduce cel mai mult eroarea pe pruning set)

- Ordonare bazată pe clase: clasele sunt selectate crescător după dimensiune (clasele rare sunt selectate prima dată)

- Regulile corespunzătoare unei clase sunt plasate consecutiv în lista de reguli

Extragerea regulilor de clasificare

Exemplu: algoritmul RIPPER (Repeated Incremental Pruning to Produce Error Reduction)

Particularităţi:

- Adăugarea unei noi reguli corespunzătoare unei clase este stopată:
  - Când regula devine prea complexă
  - Când 'următoarea' regulă are o eroare de clasificare (pe setul de validare) mai mare decât un prag prestabilit

- Dacă la sfârşit rămân date "neacoperite" atunci se poate defini o regulă de tipul "catch all" căreia i se asociază clasa dominantă din setul de date "neacoperite"

Sumar

- Arbori de decizie
  - Criterii de ramificare: Information Gain, Gain Ratio, Gini Index
  - Algoritmi de construire: ID3, C4.5
  - Simplificare (pruning)
  - Avantaje și dezavantaje
- Reguli de clasificare
  - Caracteristici: suport, încredere
  - Algoritmi de extragere a regulilor din date (RIPPER)
- Modelele bazate pe reguli sunt interpretabile
  - https://christophm.github.io/interpretable-ml-book/tree.html
  - interpretable models: https://github.com/csinva/imodels

Cursuri următoare

Curs 5:

- Clasificatori bazați pe prototipuri (k Nearest Neighbours)
- Modele probabiliste (Naive Bayes classifiers)

Curs 6:

- Reţele neuronale (Multilayer Perceptrons)
- Clasificatori bazaţi pe vectori suport (Support Vector Machines)