

Etapele procesului de DM (3)

- Evaluarea rezultatului:** nu tot ce ieșe dintr-un proces de data mining este valoare.
- Unele rezultate sunt adevaruri pur statistică care se poate deduce și fără aplicarea algoritmilor, iar altele nu prezintă interes pentru utilizator.
- De aceea este necesară evaluarea rezultatelor de către expert pentru a separa cunoștințele valuroase de celelalte lucruri obținute în urma rularii algoritmului.

25 

Principiul lui Bonferroni

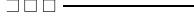


O cunoștință descoperită în urma procesului de data mining poate fi un adevar pur statistic. Exemplu (din [Ullman 03]):

- În 1950 David Rhine, un parapsiholog de la universitatea Duke a testat studentul pentru a afla dacă au sau nu percepție extrasenzorială (ESP).
- Pentru acesta l-a cerut să ghicească culoarea a 10 carti de joc successive - rosu sau negru. Rezultatul a fost că 1/1000 din participanți au ghicit toate cele 10 carti - în consecință el a declarat că având ESP.

26 

Algoritmi



Algoritmi de predicție:

- Clasificare
 - Regresie
 - Detectia deviatiei
- Algoritmi de descriere:

- Grupare - Clustering
- Gasirea de reguli de asociere
- Descoperirea de sabloane secențiale

31 

Clasificare



Intrare:

- Un set având un număr k de clase $C = \{c_1, c_2, \dots, c_k\}$.
 - Un set de n articole etichetate $D = \{(d_1, c_1), (d_2, c_2), \dots, (d_n, c_n)\}$. Articolele sunt $d_i \in D$, fiecare articol d_i fiind etichetat cu clasa $c_i \in C$. D este numit și setul (multimea) de antrenare (training set).
- Pentru calibrarea unor algoritmi este necesar și un set (multime) de validare (validation set). Acesta conține de asemenea articole etichetate, neniechis în setul de antrenare.
- lesire:
- Un model sau metoda de a clasifica noi articole (clasificator). Seul conținând noile articole se numește set de test (test set)

32 

Principiul lui Bonferroni

- Re-testarea efectuată doar asupra acestora nu a mai avut aceleși rezultate, el considerand că în momentul în care au alături ca ESP au pierdut această capacitate.
- David Rhine nu a realizat ca statistica spune că probabilitatea de a ghici 10 carti consecutive este $1/10^{10}$ deoarece probabilitatea de a ghici o carte este $1/2$ (rosu sau negru)!

27 

Principiul lui Bonferroni



- Astfel de rezultate pot fi regăsite în lecția unui algoritm de data mining și trebuie recunoscute ca adevaruri statistică și nu ca rezultate reale ale procesului de data mining.

- Astfel de rezultate sunt obiectul principiului lui Bonferroni. Aceasta poate fi sintetizat astfel:

Daca sunt prea multe concluzii, unele vor fi adevărate din motive pur statistică.

Data Mining, note: "Tuning data until it converges... and if you torture it enough, it will confess to anything" - Jeff Jones, 1994 (<http://www.cs.cmu.edu/~mason/courses/Deployment.html>)

28 

Exemplu



• Să considerăm că articolele sunt pacienti ai unui serviciu de urgențe medicale.

• Există 5 clase, $C_1, C_{10}, C_{20}, \dots, C_{100}$, unde eticheta C_i înseamnă că pacientul poate fi lăsat să aștepte maxim k minute fară că starea sa se înrăutătească.

• Vom reprezenta datele setului de antrenare sub forma tabelară.

• Iesirea unui algoritm de creare a unui clasificator poate fi să completeze un arbore de decizie sau un set ordonat de reguli.

• Modelul obținut poate fi utilizat apoi pentru a clasifica noi pacienți asignându-le o etichetă din multimea celor 5 de mai sus.

33 

Multimea de antrenare UPU



Name (or ID)	Vital risk?	Danger if water?	0 resource needed	1 resource needed	>1 resource needed	>1 resource needed and vital resources affected	Waiting time before new label?
John	No	No	Yes	No	No	No	00
Maria	No	No	No	No	Yes	No	010
Natalia	Yes	Yes	Yes	No	No	No	010
Oliver	No	No	No	No	Yes	Yes	030
Kiril	No	No	No	Yes	No	Yes	060
Denis	No	No	No	No	Yes	No	010
Jean	No	No	Yes	Yes	No	No	010
Patricia	Yes	Yes	No	No	Yes	Yes	030

34 

Cuprins

- Ce este data mining
- Etapele procesului de data mining
- Metode si subdomenii in data mining
- Preprocesarea datelor
- Sumar

29 

2 tipuri de metode



- Metode predictive:** Aceste metode utilizează anumite variabile pentru a prezice valoarea altor variabile. Un exemplu din acesta categoria este clasificarea; bazat pe date cunoscute și de etichetă (clasificate), algoritmii de clasificare crează modele care pot fi folosite pentru predicție.
- Metode descriptive:** Algoritmii din această categorie găsește sabioane / modele care descriu structura internă a unui set de date. De exemplu algoritmii de grupare (clustering) găsește grupuri de obiecte similară în setul de date (grupuri numite cluster) și de asemenea detectează posibile obiecte izolate, separate de oricare dintre cluster - acea numită "outliers".

30 

Regresia (2)

Există multe tipuri de regresie, de exemplu:

- Regresie liniara
- Regresie liniara simplă
- Regresie logistică
- Regresie neliniera
- Regresie nonparametrică
- Regresie robustă

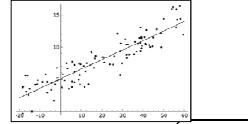
37 

Exemplu



Exemplu de regresie liniara

(sursa: http://en.wikipedia.org/wikifile/Linear_regression.svg)

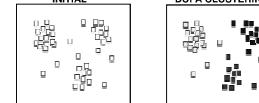


38 

Exemplu



• Dacă avem un set de 2 puncte într-un spațiu 2D, să se găsească grupurile/clusterurile naturale formate de ele



43 

Regresia (1)



• Regresia provine din statistică.

• Însemnată: precizarea (predicția) valorii unei anumite variabile continuu pe baza valorilor altor variabile, considerând un model de dependență liniară sau neliniară (Tian, Steinbach, Kumar 06).

• Utilizată în predicție sau în proiecție;

• Analizează bazate pe regresie sunt folosite pentru înțelegerea relațiilor dintre variabile dependente și independente.

Intrebare: care e deosebita între predicție și proiecție?

44 

Detectia deviatiei

- Detectia deviatiei (sau a anomalilor) presupune descompunerea de detalii non-relevante de la comportamentul normal. Putem să identificăm care sunt o categorie semnificativa de astfel de date anormale.
- Detectia deviatiei poate fi utilizată în multe situații:
 - În fază de rulare a algoritmilor de data mining, datele anormale pot indica că există un atac.
 - Audit: astfel de informații pot arăta existența unor probleme sau practici gresite.
 - Detectia fraudei: cercetă frauduloase contin desori informații false.
 - Detectia intruziunilor: într-o rețea de calculatoare poate fi facuta să detecteze datele anormale.
 - Corectarea datelor (data cleaning): astfel de informații anormale pot reprezenta datele oronice cără trebuie corectate

39 

Metode de detectie a deviatiei



- Tehnici bazate pe distanțe (ex.: k-nearest neighbor).
- Folosirea "One Class Support Vector Machines" (Ca UCM) este o metodă de detectie exemplu de antrenare din aceeași clasă și de asemenea detectă ca de-a doua clasă.
- Metode predictive (arbore de decizie, rețele neurale).
- Detectia punctelor izolate folosind clustering.
- Înregistrari care nu respectă regulile de asociere
- Analize Hotspot (clusterare având o valoare ridicată a anumitor parametri; de exemplu poliția folosește astfel de analize pentru analiza zonelor unde criminalitatea are valori ridicate)

40 

Reguli de asociere



- O regula** este o construcție de tipul $X \rightarrow Y$ unde X și Y sunt multimi de articole (itemsets).

• **Suportul** unei regule $X \rightarrow Y$ este numarul de aparitii ale $X \cup Y$ în totalul acestor reuniri ca itemset:

$$\text{support}(X \rightarrow Y) = \text{support}(X \cup Y)$$

• **Confidența** unei regule este proporția transacțiilor continându-l pe Y în multimea transacțiilor continându-l pe X :

$$\text{confidence}(X \rightarrow Y) = \text{support}(X \cup Y) / \text{support}(X)$$

• Acceptăm o regula ca validă doar dacă suportul și confidența ei sunt mai mari sau egale cu pragurile date.

45 

Exemplu



• Sa considerăm următorul set de tranzacții:

Transacție ID	Items
1	Bread, Milk, Butter, Orange Juice, Onion, Beer
2	Pen, Ink, Baby diapers
3	Milk, Butter, Onion, Garlic, Beer
4	Orange Juice, Shirsh, Sheesh, Bread, Milk
5	Butter, Onion, Beer, Orange Juice

• Dacă $s = 60\%$ atunci (Bread, Milk, Orange Juice) sau (Onion, Garlic, Beer) sunt multimi frecvente de articole.

• De asemenea, dacă $s = 60\%, c=70\%$ atunci regula (Onion, Beer) → (Garlic) și validă având suport de 60% și incidență de 75%.

46 

Algoritmi

- Algoritmi de predicție:**
 - Clasificare
 - Regresie
 - Detectia deviatiei
- Algoritmi de descriere:**
 - Grupare - Clustering
 - Gasirea de reguli de asociere
 - Descoperirea de sabloane secențiale

41 

Clustering



Intrare:

- Un set de obiecte $D = \{d_1, d_2, \dots, d_n\}$ (numite uzual puncte). Obiectele nu sunt etichetate și nu există definitiv reuniune de clase.
- Metoda de distanță (măsură de similaritate) care poate fi utilizată pentru a calcula distanța dintre oricare două puncte. O distanță mica înseamnă "aproape" iar una mare "departe".
- Un algoritm necesită introducerea unei valori pentru numărul de clusteruri de către utilizator.
- Iesire:
- Multimea de obiecte/puncte, numite cluste, unde punctele din același cluster sunt "aproape" unele de altele, considerând funcția de distanță existentă.

42 

Descoperire sabloane secențiale



Intrare:

- O mulțime de m articole $I = \{i_1, i_2, \dots, i_m\}$. O seconță este o listă ordonată de elemente din articolele I .
- Mulțimea de seconde este S_m (numită "sequence database").
- O funcție booleană care poate testa dacă o seconță S_m este inclusă (este o subseconță) în seconța S_n . În acest caz S_n este numita o superseconță a lui S_m .
- Un prag (percent sau valoare absolută) necesar pentru a gasi secondele frecvente.
- Un prag (percent sau valoare absolută) necesar pentru a gasi secondele rare.

47 

Exemplu



- Într-o librărie putem găsi secondele de tipul: `Book_on_C, Book_on_C++` → `Book_on_Perl`

• Din această seconță se poate deduce o regula de tipul: după cumpărarea unor cărți de C și C++, clientul cumpără cărți de Perl:

`Book_on_C, Book_on_C++ → Book_on_Perl`

48 

Cuprins

- Ce este data mining
- Etapele procesului de data mining
- Metode si subdomenii in data mining
- Preprocesarea datelor
- Sumar

49

Prezentare Noile cursuri UBD-7

Tipuri de date

- Categorice vs. Numerice
- Scale de masurare
 - Nominala
 - Ordinata
 - Interval
 - Proportionala (eng.: ratio scale)

51

Prezentare Noile cursuri UBD-7

Date categorice si numerice

- Datele categorice constau in numere apartinand unei multimi continue sau discrete de valori numerice.
- Valorile sunt ordonate, astfel incat se poate testa ordinea ($<$, $<=$, $>$, $>=$).
- Uneori este necesara conversia datelor categorice in date numerice asignand o valoare numerica (sau cod) pentru fiecare eticheta (categorie).

52

Prezentare Noile cursuri UBD-7

Nominala

- Valurile aparținând scălei nominale sunt caracterizate prin etichete (nume).
- Valurile sunt neordonate și au importanță (pondere) egală.
- Nu putem calcula valoarea medie sau mediana a unui astfel de set.
- Putem totuși calcula valoarea modală – valoarea cea mai frecventă.
- Datele nominale sunt de tip categoric și uneori este nevoie să fie convertite în valori numerice.

55

Prezentare Noile cursuri UBD-7

Ordinala

- Valurile de acest tip sunt ordonate dar diferența / distanța între două valori nu poate fi calculată.
- Putem specifica doar ordinea / poziția în sirul ordonat pentru aceste valori.
- Exemplu: lista gradelor militare, lista gradelor didactice, lista în ordinea sosirii a concurenților la o cursă sportivă (doar lista numelor, fără timpul de sosire), etc.
- Pentru astfel de valori putem calcula valoarea modală sau mediana (valoarea mijlocie) dar nu media.
- Valurile sunt în esență categorice dar este mai ușor sa le assignăm valori numerice în caz de conversie.

56

Prezentare Noile cursuri UBD-7

Interval

- Aceste valori sunt numerice.
- În acest caz diferența între două valori are semnificație.
- Exemplu: temperatură în grade Celsius: diferența între 10 și 20 de grade este aceeași ca cea între 40 și 50 de grade (aceeași cantitate de energie în cazul încălzirii unui corp).
- Valoarea zero (0) nu înseamnă "nimic" ci este o valoare arbitrară aleasă. De aceea sunt permise și valori negative.
- Putem calcula media, mediana, valoarea modală, deviația standard și putem utiliza regresia pentru a proгnoza noi valori.

57

Prezentare Noile cursuri UBD-7

Proportionala (ratio)

- Aceste valori sunt numerice, ca și în cazul interval, dar zero (0) înseamnă "nimic".
- Valoare negativă nu are sens.
- Raportul dintre două valori are semnificație.
- Exemplu: greutatea în kg. Un copil de 10 kg este de două ori mai gros decât unul de 5 kg.
- Alte exemple: temperatură în grade Kelvin, lungimea în metri, varsta în ani, etc.
- Toate operațiile matematice sunt permise.

58

Prezentare Noile cursuri UBD-7

Date categorice si numerice

- Datele numerice constau în numere apartinând unei multimi continue sau discrete de valori numerice.
- Valorile sunt ordonate, astfel încât se poate testa ordinea ($<$, $<=$, $>$, $>=$).
- Uneori este necesara conversia datelor categorice în date numerice asignând o valoare numerică (sau cod) pentru fiecare etichetă (categorie).

53

Prezentare Noile cursuri UBD-7

Scale de masurare

- Stanley Smith Stevens, directorul laboratorului de psihो-acustică de la Universitatea Harvard a afirmat într-un articol publicat în 1946 în revista Science că toate masurările din stiință utilizează patru tipuri diferenți de scale de masurare:
 - Nominala
 - Ordinata
 - Interval
 - Proportionala (ratio scale)

54

Prezentare Noile cursuri UBD-7

Date binare

- Uneori un atribut poate avea doar 2 valori: 0 sau 1, masculin sau feminin, da sau nu, etc. În acest caz datele se numesc **binare**. Avem două cazuri:
 1. Binare simetrice: cele două valori au importanță / pondere egală (ca în cazul genului).
 2. Binare asymetrice: una dintre valori are o importanță mai mare decât cealaltă. De exemplu în cazul unei analize pentru depistarea HIV valoarea Pozitiv are o greutate mai mare decât Negativ.

59

Prezentare Noile cursuri UBD-7

Date binare

- Atributele binare pot fi tratate uneori ca fiind de tip interval sau proporțional dar în cea mai mare parte a cazurilor ele vor fi considerate nominale (cele binare simetrice) sau ordinaile (cele binare asymetrice).
- Există o serie de funcții de distanță (dissimilaritate) care pot fi utilizate în acest caz.

60

Prezentare Noile cursuri UBD-7

Preprocesarea datelor

- Tipuri de date
- Masurarea datelor**
- Curătarea datelor
- Integrarea datelor
- Transformarea datelor
- Reducerea datelor
- Discretizarea datelor

61

Prezentare Noile cursuri UBD-7

Masurarea datelor

- Masurarea tendinței valorii centrale:
 - Media
 - Mediana
 - Valoare modală
 - Mijlocul intervalului
- Masurarea dispersiei:
 - Interval de valori
 - Procentul k
 - IQR
 - Sumarul de 5 numere
 - Deviația standard și varianta

62

Prezentare Noile cursuri UBD-7

Dispersia valorilor (1)

- Intervalul de valori. Este diferența între cea mai mare și cea mai mică valoare.
- Exemplu: Ior {1, 2, 3, 5, 7, 1001, 2002, 9999} valoarea este 9999 – 1 = 9998.
- Procentul k. Aceasta este o valoare x, apărând setul de date și care are proprietatea că un procent de k valori din set sunt mai mici sau egale cu ea.
- Exemplu: Mediana este procentul 50.
- Cele mai utilizate procente sunt 25 și 75, numite și quartile (eng. Quartiles). Notație: Q1 pentru 25% and Q3 pentru 75%.

67

Prezentare Noile cursuri UBD-7

Dispersia valorilor (2)

- Intervalul între cuartile (eng. Interquartile range sau IQR) este diferența între Q3 și Q1:
$$IQR = Q3 - Q1$$
- Potențiale puncte izolate (outliers) sunt valori la distanță de cel puțin 1,5 x IQR sub Q1 sau peste Q3.
- Sumarul de 5 numere. Uneori mediana și cuartile nu sunt suficiente pentru a reprezenta dispersia valorilor. În acest caz adăugăm și valoarea minimă și maximă din setul de date.
- (Min, Q1, Median, Q3, Max) este ceea ce se numește sumarul de 5 numere (eng. five-number summary).

68

Prezentare Noile cursuri UBD-7

Tendinta centrala - Medie

- Fie n valori ale unui atribut: x_1, x_2, \dots, x_n .
- Media:** Media aritmetică sau valoarea medie este:
$$\mu = (x_1 + x_2 + \dots + x_n) / n$$
- Dacă valoare lui x are ponderile w_1, w_2, \dots, w_n , atunci **media aritmetică ponderată** este:
$$\mu = (w_1 x_1 + w_2 x_2 + \dots + w_n x_n) / (w_1 + w_2 + \dots + w_n)$$
- În unele cazuri se practică eliminarea celor mai mici și celor mai mari valori (de exemplu cele mai mici 1% și cele mai mari 1%).

63

Prezentare Noile cursuri UBD-7

Tendinta centrala - Mediana

- Mediana:** Valoarea mediană a unui set de n valori ordonate este valoarea din mijlocul seventele de valori.
- Exemplu: Mediana pentru {1, 3, 5, 7, 1001, 2002, 9999} este 7.
- Dacă n este par (și datele sunt numerice), mediana este media valorilor din mijlocul seventelor:
 - mediana pentru {1, 3, 5, 7, 1001, 2002} este 6 (media valorilor 5 și 7).

64

Prezentare Noile cursuri UBD-7

Dispersia valorilor (3)

- Exemplu:**
Pentru {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11} Intervalul de valori = 10; Mijloc interval = 6; Q1 = 3; Q2 = 6; Q3 = 9; IQR = 9 - 3 = 6
Pentru {1, 2, 3, 4, 5, 6, 7, 8, 8} Intervalul de valori = 7; Mijloc interval = 5,5; Q1 = 3; Q2 = 5,5 = (5+6)/2; Q3 = 7; IQR = 7 - 3 = 4
Pentru {1, 3, 5, 7, 8, 10, 11, 13} Intervalul de valori = 12; Mijloc interval = 7; Q1 = 4; Q2 = 7,5; Q3 = 10,5; IQR = 10,5 - 4 = 6,6

69

Prezentare Noile cursuri UBD-7

Dispersia valorilor (4)

- Deviația standard** (abaterea standard). Pentru n valori (observați) aceasta este:
$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}, \text{ unde } \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i.$$
- Patratul deviației standard se numește **varianță** (dispersie).
- Deviația standard măsoară raspândirea valorilor în jurul mediului.
- Valoarea 0 este obținuta doar dacă toate valorile sunt egale.

70

Prezentare Noile cursuri UBD-7

Tendinta centrala – Valoare modală

- Valoare modală:** Cea mai frecventă valoare din multimea respectivă.
- Un set de date poate avea mai multe valori mode. Pentru 1, 2 și 3 setul este denumit ca unimodal, bimodal sau trimodal.
- Cand fiecare valoare apare o singura data setul nu are valoare modală.
- Pentru un set unimodal, valoarea modală măsoară tendința centrală a acestuia. În acest caz există relația empirică:
$$\text{medie} - vmod = 3 \times (\text{medie} - \text{mediana})$$

65

Prezentare Noile cursuri UBD-7

Tendinta centrala – Mijlocul intervalului

- Mijlocul intervalului:** este media aritmetică a celei mai mari și celei mai mici valori.
- Exemplu: pentru {1, 3, 5, 7, 1001, 2002, 9999} valoarea de mijloc a intervalului este 5000 (media lui 1 și 9999).

66

Prezentare Noile cursuri UBD-7

Preprocesarea datelor

- Tipuri de date
- Masurarea datelor
- Curătarea datelor
- Integrarea datelor
- Transformarea datelor
- Reducerea datelor
- Discretizarea datelor

71

Prezentare Noile cursuri UBD-7

Curătarea datelor

- Obiectivele principale ale acestei etape sunt:
 - Înlăturarea (sau eliminarea) valorilor lipsă
 - Înlăturarea zgomotului
 - Eliminarea sau doar identificarea punctelor izolate (eng.: outliers - valori aberante)

72

Prezentare Noile cursuri UBD-7

Valori lipsa

- Anumite atribute pot contine valori lipsa / valori nule (NULL).
- Pot sa apară din diverse motive:
 - Probleme hardware, software sau erori ale operatorului de introducere date.
 - Date care nu au fost colectate într-o anumita perioadă (nu au fost considerate importante).
 - Valori eliminate pentru ca erau inconsistente cu restul.
- Există doar soluții în acest caz:
 - Ignorarea linierelor din tabela de date de intrare care contin valori lipsă. Așa se poate face doar în cazul în care numărul lor este mic și eliminarea nu induce erori în calculele ulterioare.

Foto: Wikipedia/Nume de curs UBD-7

Valori lipsa

- Pot apărea din diverse motive:
 - Probleme hardware, software sau erori ale operatorului de introducere date.
 - Date care nu au fost colectate într-o anumita perioadă (nu au fost considerate importante).
 - Valori eliminate pentru ca erau inconsistente cu restul.
- Există două soluții în acest caz:
 - Ignorarea linierelor din tabela de date de intrare care contin valori lipsă. Așa se poate face doar în cazul în care numărul lor este mic și eliminarea nu induce erori în calculele ulterioare.

Foto: Wikipedia/Nume de curs UBD-7

Puncte izolate

- Punctele izolate / valori abnormale (outliers) sunt valori numerice de atribut aflate la o distanță mare de restul datelor.
- Uneori sunt valori corecte: salariajul unui CEO poate fi mult mai mare decât al restului angajaților.
- De cele mai multe ori însă sunt date eronate / zgomot.
- Trebuie identificate și eliminate sau înlocuite, deoarece multi algoritmi sunt sensibili la astfel de puncte – de exemplu k-means da clustere eronate în prezentă lor.

Foto: Wikipedia/Nume de curs UBD-7

Identificare puncte izolate

- Folosind IQR: am mai spus că potențiale puncte izolate (outliers) sunt valori aflate la o distanță de cel puțin 1.5 * IQR sub Q1 sau peste Q3.
- Folosind deviația standard: valori care sunt la distanță mai mare de 2 de valoarea medie sunt potențiale valori izolate.
- Folosind clustering: după găsirea clusterelor, punctele aflate în afara oricărui cluster (sau foarte departe de orice centru de cluster) sunt potențiale puncte izolate.

Foto: Wikipedia/Nume de curs UBD-7

Netezire zgromot

- Zgromot poate fi definit ca o eroare aleatoare sau variată pentru o valoare măsurată ([Han, Kamber 06]).
- Pentru eliminarea zgromotului se pot folosi diverse tehnici de netezire ca:
 - Regresia (prezentată mai devreme)
 - Binning = Partitionarea în trame

Foto: Wikipedia/Nume de curs UBD-7

Binning

- Aceasta tehnica poate fi folosita pentru netezirea unui set ordonat de valori. Este bazata pe valoare din vecinătate.
- Sunt 2 pasi:
 - Partitionarea setului ordonat în mai multe trame
 - Netezirea fiecarei trame de date pe baza mediei, medianei sau capetelor intervalului.

Foto: Wikipedia/Nume de curs UBD-7

Preprocesarea datelor

- Tipuri de date
- Măsurarea datelor
- Curătarea datelor
- Integrarea datelor
- Transformarea datelor
- Reducerea datelor
- Discretizarea datelor

Foto: Wikipedia/Nume de curs UBD-7

Integrarea datelor

- După ce datele provenind din sursele de date individuale au fost curătate ca mai înainte, are loc integrarea acestora într-un unic ansamblu de date.
- Activități din această categorie includ:
 - 1. Integrarea schemelor.** Problemele sunt de tip sinonime (același lucru apare în două surse cu nume diferite) și omónime (cu același nume se găsesc date de tip diferit în diverse surse).
 - 2. Duplicate și redundanță.** În surse diferite se regăsesc același informații. Duplicatele și redundanța trebuie eliminate în această fază.

Foto: Wikipedia/Nume de curs UBD-7

Exemplu

- Fie urmatorul set ordonat de valori: 1, 2, 4, 6, 9, 12, 16, 17, 18, 23, 34, 56, 78, 79, 81.

Tramele inițiale	Se utilizează media	Se utilizează mediana	Se utilizează capetele intervalului
1, 2, 4, 6, 9	4, 4, 4, 4, 4	4, 4, 4, 4, 4	1, 1, 9, 9
12, 16, 17, 18, 23	17, 17, 17, 17, 17	17, 17, 17, 17, 17	12, 12, 12, 23, 23
34, 56, 78, 79, 81	66, 66, 66, 66, 66	78, 78, 78, 78	34, 34, 81, 81, 81

Foto: Wikipedia/Nume de curs UBD-7

Rezultat

- Rezultatul netezirii este:
 - Initial: 1, 2, 4, 6, 9, 12, 16, 17, 18, 23, 34, 56, 78, 79, 81
 - Cu medie: 4, 4, 4, 4, 17, 17, 17, 17, 17, 66, 66, 66, 66
 - Cu mediana: 4, 4, 4, 4, 17, 17, 17, 17, 78, 78, 78, 78
 - Cu capetele intervalului: 1, 1, 1, 9, 9, 12, 12, 12, 23, 23, 34, 34, 81, 81, 81

Foto: Wikipedia/Nume de curs UBD-7

Integrarea datelor

- Inconsistențe.
 - Acum sunt valori aflate în conflict în același set de date.
 - Din exemplu Data nasterii=1.1.1980 și Varsta=30 pentru o același persoană reprezintă o inconsistenta.
 - Pentru detectarea acestora sunt necesare informații suplimentare despre date.

Foto: Wikipedia/Nume de curs UBD-7

Transformarea datelor

- Aceasta etapa cuprinde procese de transformare și sumarizare a datelor ca:
 - Normalizare
 - Construcție de noi atribute
 - Sumarizare folosind funcțiile statistice

Foto: Wikipedia/Nume de curs UBD-7

Normalizare

- Toate atributele numerice sunt scalate pentru a avea valori între-un acelaș interval specificat:
 - De la 0 la 1,
 - De la -1 la 1 sau, mai general: $|v| \leq r$ unde r este o valoare specificată.
- Normalizarea este necesara cand anumite atribute sunt mai importante decât altale doar pentru ca plaja lor de valori este mai întinsă.
- Exemplu: Distanța euclidiană între A(0,5, 10) și B(0,01, 2111) este ≈ 2010 , determinată aproape exclusiv pe baza celei de-a doua componente.

Foto: Wikipedia/Nume de curs UBD-7

Reducerea datelor

- Nu toate informații colectate sunt necesare pentru anumite analize asupra datelor.
- Reducerea înseamnă extragerea unei parti din datele colectate și preprocesate, să anume doar a aceliei parti care conțin informații necesare procesului de analiză respectiv.
- Se realizează de exemplu prin selectarea doar a anumitor atribute (coloane) sau exemple (linii) din matricea de date supusă analizei.

Foto: Wikipedia/Nume de curs UBD-7

Discretizare

- Nu există algoritmi sănătoși concepuți pentru analiza datelor discrete.
- Pentru a ii folosi, este necesar ca valorile continue să fie înlocuite cu unele discrete.
- Din exemplu Varsta care are valori numerice va fi înlocuită cu un atribut categoric având doar valorile Copil, Adult, Batran.
- Există diverse metode de a efectua operația de discretizare, folosind împărțirea în trame (binning), histograme, intervale bazate pe entropie, clustering, ierarhii de concepte, etc.

Foto: Wikipedia/Nume de curs UBD-7

Normalizare

- Putem folosi pentru normalizare:
 - Normalizarea min-max:** $v_{new} = (v - v_{min}) / (v_{max} - v_{min})$
 - Pentru valori pozitive** putem folosi formula: $v_{new} = v / v_{max}$
 - Normalizarea de tip z-score (o deviație standard):** $v_{new} = (v - v_{mean}) / \sigma$
 - Scalarea decimală:** $v_{new} = v \cdot 10^n$
- Unde n e cel mai mic întreg pentru care toate numerele devin în modul mai mic decât o valoare dată ($v_{new} \leq 1$).

Foto: Wikipedia/Nume de curs UBD-7

Constructie noi atribute

- Este o tehnică prin care se adaugă noi atribute (în engleză se numește **feature construction**).
- Exemplu: dacă setul de date are un atribut Culoare cu valori (Rosu, Verde, Albastru) putem construi 3 noi atribute numite Rosu, Verde, Albastru conținând doar valori binare (0 sau 1) cu 1 singur pe atributul (culoarea) corespunzător colorii curente.
- Alt exemplu: se poate utiliza un arbore de decizie sau set de reguli pentru a construi noi atribute din cele existente – ca de exemplu clasa.

Foto: Wikipedia/Nume de curs UBD-7

Sumar

- În acest curs am prezentat:
 - O listă de definiții alternative pentru Data Mining și cetera exemple pentru a înțelege ce este Data Mining și cum se face Data Mining.
 - O descriere a principalelor implicații ale Data Mining și despre faptul că Data Mining este de fapt o reunire heterogenă de subdomenii.
 - Pasi principali ai Data Mining process de colectarea de date și a locației diferențiale (deosebit de date, arhive sau sisteme operaționale), para la pasul final de evaluare
 - O scurtă descriere a subdomeniilor principale ale Data Mining cu exemple pentru fiecare dintre ele.
 - O descriere cuprinzătoare a procesului de preprocesare a datelor.

Lectia viitoare: Multimi frecvente de articole și reguli de asociere

Foto: Wikipedia/Nume de curs UBD-7

Bibliografie și referințe

- [Liu 11] Bing Liu, 2011. Web Data Mining, Exploring Hyperlinks, Content, and Queries in Web Data. Second Edition, Springer.
- [Tan, Steinbach, Kumar 06] Tan, Steinbach, Kumar, 2006. Introduction to Data Mining, Addison-Wesley, 1-18.
- [Kimbala, Ross 02] Ralph Kimball, Margy Ross, 2002. The Data Warehouse Toolkit: Practical Techniques for Building Data Warehouses and OLAP Applications, 2nd ed., John Wiley and Sons, 1-18, 396.
- [Miers, Fischetti 11] Reid Mifflin and Thomas Fischetti, Data mining tools, 2011, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, DOI: 10.1002/widm.24/pdf
- [Ullman] Jeffrey Ullman, Data Mining Lecture Notes, 2003-2009, web page: <http://infolab.stanford.edu/~ullman/mining/>
- What is the difference between data mining, machine learning and AI? <http://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai>

Foto: Wikipedia/Nume de curs UBD-7

Sumarizare

- Se folosesc funcții statistică (min, max, sum, avg, count, etc) pentru a adăuga date sumare datelor originale.
- Exemplu: se adaugă suma varianțelor pe zi sau luna sau an, numărul mediu sau total de clienți sau înfrântăci, etc.
- Aceste date vor fi folosite pentru creșterea vitezei de procesare a unor cereri în procesul de data mining.
- Rezultatul este un cub de date și fiecare data sumară este atașată unui nivel de granularitate.

Foto: Wikipedia/Nume de curs UBD-7

Preprocesarea datelor

- Tipuri de date
- Măsurarea datelor
- Curătarea datelor
- Integrarea datelor
- Transformarea datelor
- Reducerea datelor
- Discretizarea datelor

Foto: Wikipedia/Nume de curs UBD-7

Bibliografie și referințe

- [Akash Mitra, Data Mining - a simple guide for beginners, 2012, <http://www.dataminingtips.com/data-warehouse/11-data-mining-definitions.html>
- [Wikipedia] Wikipedia, the free encyclopedia, en.wikipedia.org
- [Han, Kamber 06] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann Publishers Inc., 2006.
- [Stevens June 1946] Stevens, S.S. On the Theory of Scales of Measurement. Science June 1946, 103 (2884): 677–680.
- [Wikipedia] Wikipedia, the free encyclopedia, en.wikipedia.org
- [Liu 11] Bing Liu, 2011. CS 593 Data mining and text mining course notes, <http://www.cs.uic.edu/~iub/teach/cs593-fall-11/cs593.html>

Foto: Wikipedia/Nume de curs UBD-7

<p>Capitolul 8</p> <p>Data mining: Reguli de asociere si multimi frecvente de articole (frequent itemsets)</p> <p>F. Radulescu, Curs Utilizarea bazelor de date, anul IV CS.</p>	<p>Problema</p> <ul style="list-style-type: none"> □ Problema cosului de produse presupune ca avem un mare numar de articole, e.g. "paine", "lalte". □ Cumparatorii pun in cosul lor de produse anumite submultimi de articole iar noi vom afla ce articole sunt cumporate impreuna chiar daca nu si de cine. <p>F. Radulescu, Curs Utilizarea bazelor de date, anul IV CS.</p>	<p>Reguli de asociere</p> <ul style="list-style-type: none"> □ Regulele de asociere: Daca X si Y sunt multimi de articole, o regula este o propozitie de forma $X \rightarrow Y$ inseamnand ca daca gasim toate articolele din X intr-un cos atunci sunt mari sanse sa gasim in acel cos si articolele din Y. □ Probabilitatea de a-l gazi pe Y pentru a accepta aceasta regula este numita <i>incredere</i> acelei reguli. □ In mod normal vom cauta doar reguli care au o incredere peste un anumit prag. <p>F. Radulescu, Curs Utilizarea bazelor de date, anul IV CS.</p>	<p>Reguli de asociere</p> <ul style="list-style-type: none"> □ Putem de asemenea cere ca increderea sa fie semnificativ mai mare decat in cazul in care articolele ar fi plaseate aleator in cos. □ De exemplu putem gasi o regula ca {lalte, urb} \rightarrow paine doar pentru ca foarte multa lume cumpara paine. □ Totusi exemplul bere/scutice arata ca regula {scutice} \rightarrow {bere} este verificata cu o incredere semnificativ mai mare decat a submultimii de cosuri continand bere. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>
<p>Problema</p> <ul style="list-style-type: none"> □ Vanzatorii utilizeaza aceste informatii pentru asezarea articolelor in magazin, in felul acesta putand sa controleze modul in care anumite clase de cumparatori parcurg raioanele magazinului. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Alte utilizari</p> <ul style="list-style-type: none"> □ Cos = documente; articole = cuvinte. □ Cuvintele aparand frecvent impreuna in documente pot reprezenta fraze sau concepte legate intre ele. □ Poate fi utilizat pentru colectarea de informatii (intelligence). <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Scutice si bere</p> <ul style="list-style-type: none"> □ În 1992, Thomas Elschnig, managerul unui grup de consultanță pentru vânzări cu ambițiile de la Teradac, și personalul său au proiectat și analizat 1,2 milioane de cosuri de produse de la aproximativ 75 de magazine Oscar Drap. □ Au fost răsuza interogări alezi de date pentru a identifica afinitățile. □ Analiza a descoperit că între orele 17:00 și 19:00 consumatorii cumpără bere și scutice. □ Managerii Oscar NU au exploatat asocierea dintre bere și scutice multănd produsele mai aproape unei de celelalte pe rafturi. □ Acest studiu pentru suportul decizional a fost realizat folosind instrumente de interrogare pentru a găsi asociările. □ Povestea adevarată este foarte faidă în comparare cu legenda. <p>Sursa: http://www.dssresources.com/newsletters/66.php</p> <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Cauzalitate</p> <ul style="list-style-type: none"> □ Cauzalitate, Ideal: am vrea sa stim daca intr-o regula de asociere prezenta elementele X efectiv "cauzeaza" (determina) cumpararea lui Y. □ Cauzalitatea este insa un concept echivoc. □ Exemplu: <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>
<p>Alte utilizari</p> <ul style="list-style-type: none"> □ Cos = propozitii, articole = documente. □ Doua documente continand aceleasi propozitii pot reprezenta un plagiat sau "mirror sites" pe web. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Obiective pentru aceasta clasa de probleme</p> <ul style="list-style-type: none"> □ Gasirea de: <ul style="list-style-type: none"> □ Reguli de asociere □ Cauzalitate □ Multimi frecvente de articole <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Cauzalitate</p> <ul style="list-style-type: none"> □ Daca scadem pretul scutecelor si crestem pretul berii putem ademeneii cumparatorii de scutice care au inabilitatea de a cumpara bere din magazin, acoperind astfel pierderile din vanzarea scutecelor. □ Aceasta strategie este valabila deoarece "scutice determina bere". <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Cauzalitate</p> <ul style="list-style-type: none"> □ Actiunea inversa, micsorarea pretului la bere si marirea pretului scutecelor nu va determina cumparatorii de bere sa cumpere scutice in numar mare si vom pierde bani deoarece regula "bere determina scutice" nu este adevarata. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>
<p>Multimi frecvente</p> <ul style="list-style-type: none"> □ Multimi frecvente de articole (frequent itemsets): in multe situatii (dar nu in toate) ne intereseaza doar regulile de asociere si cauzalitatea in ceea ce priveste multimi de articole care apar <i>frecvent</i> in cosuri. □ De exemplu, nu putem conduce o buna strategie de marketing care implica produse pe care oricum nu le cumpara nimeni. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Multimi frecvente</p> <ul style="list-style-type: none"> □ Astfel, cautarile in date portesc de la premisa ca ne intereseaza doar multimi de articole cu un larg suport; □ Larg suport = ele apar impreuna in multe cosuri de produse. □ Gasim apoi reguli de asociere sau cauzalitatea implicand doar articolele cu larg suport (i.e. $\{X, Y\}$ trebuie sa apară in cel putin un anumit procent din cosuri, numit <i>prag de suport</i>) <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Abordari</p> <ul style="list-style-type: none"> □ Pentru a gasi multimi frecvente de articole avem doua abordari: <ol style="list-style-type: none"> 1.Nivel cu nivel (aplicand principiul a-priori) 2.Totale multimile de articole - de orice dimensiune - in cateva treceri (2-3 treceri) <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Nivel cu nivel</p> <ul style="list-style-type: none"> □ Procedam nivel cu nivel, gasind intai articolele frecvente (multimi de dimensiune 1), apoi perielele frecvente, tripletele frecvente, etc. □ In multe cazuri partea cea mai grea este gasirea perielelor frecvente; continuarea pe nivelele superioare necesita mai putin timp decat gasirea acestora. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>
<p>Cadru de cautare</p> <ul style="list-style-type: none"> □ Utilizam termenul <i>multimi frecvente de articole (frequent itemsets)</i> pentru "o multime de articole S care apare in cel putin a 's'-a parte din cosuri", unde s este o constanta aleasa, de obicei 0.01 sau 1%. □ Vom presupune ca avem o cantitate de date care nu incepe in memoria centrala a calculatorului. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Cadru de cautare</p> <ul style="list-style-type: none"> □ Stocare (pentru exemplele urmatoare): <ul style="list-style-type: none"> □ Fie sunt stocate intr-o baza de date relationala (BDR), de exemplu o relatie (tabela) <i>Cosuri(IdCos, articol)</i> □ Fie ca un fisier text cu linii de forma <i>(IdCos, articol1, articol2, ..., articol-n)</i>. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Nivel cu nivel</p> <ul style="list-style-type: none"> □ Algoritmii de acest tip utilizeaza o trecere per nivel. □ Există insă si abordări care nu sunt de tip nivel cu nivel: <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Toate multimile</p> <ul style="list-style-type: none"> □ Gasim toate <i>multimile frecvente de articole maximale</i> (i.e., multimi S a.i. nici o multime care include strict pe S nu este frecventa) intr-o singura trecere sau in cateva treceri. □ Tehnicile de acest tip includ fie esantionarea datelor fie citirea lor in transe □ De obicei sunt suficiente 2 treceri prin date (o trecere pentru esantionarea si inca una pentru verificarea finala) <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>
<p>Cadru de cautare</p> <ul style="list-style-type: none"> □ Cand evaluam timpul de rulare al algoritmilor parametrul optimizat este <i>numarul de treceri prin date</i>. □ Deoarece costul principal este dat adesea de timpul necesar citirii datelor de pe disc, numarul de citiri pentru fiecare data este adesea cea mai buna masura a timpului de rulare al algoritmului. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Principiul a-priori</p> <ul style="list-style-type: none"> □ Exista un principiu cheie, numit <i>monotonicitate</i> (eng. monotonicity) sau <i>principiul a-priori</i> care ne ajuta sa gasim multimi frecvente de articole: <ul style="list-style-type: none"> ▪ Daca o multime de articole S este frecventa (i.e., apare cel putin in a 's'-a parte a cosurilor), atunci orice submultime a lui S este de asemenea frecventa. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Algoritmul a-priori</p> <p>Acest algoritm procedeaza nivel cu nivel.</p> <ol style="list-style-type: none"> 1.Dandu-se pragul de suport s, in prima trecere gasim articolele care apar in cel putin a 's'-a parte a cosurilor. Aceasta multime este notata L_1, multimea articolelor frecvente. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>	<p>Algoritmul a-priori</p> <ol style="list-style-type: none"> 2. Perechile de articole din L_1 devin multimea C_2 a <i>perechilor candidate</i> pentru a doua trecere. Speram ca dimensiunea lui C_2 nu este atat de mare pentru ca altfel nu este suficient spatiu in memoria centrala pentru un contor numeric intreg al paritetiei fiecarei perechi. Perechile din C_2 al caror contor ajunge sau depaseste s formeaza multimea L_2 a <i>perechilor frecvente</i>. <p>F. Radulescu, Curs Utilizarea bazeelor de date, anul IV CS.</p>

Algoritmul a-priori

3. Tripletele candidate C_3 sunt multimele $\{A, B, C\}$ pentru care $\{A, B\}, \{A, C\}$ și $\{B, C\}$ sunt în L_2 . În a treia trecere sunt numărate aparițiile tripletelor din C_3 ; cele al căror conținut este cel putin s formează multimea L_3 a tripletelor frecvente.

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

25

Algoritmul a-priori

4. Se poate merge oricât de departe să dorescă (sau până multimile devinvide), L_i conține multimile frecvente de articole de dimensiune i , C_{i+1} este multimea candidatelor de dimensiune $i+1$ a.i. fiecare submultime a lor de dimensiune i este inclusă în L_i .

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

26

Efect a-priori

- A-priori "împinge clauza HAVING în jos pe arborele expresiei", determinându-ne în primul rand să înclocuim *Cosuri* cu rezultatul "cererii":

```
SELECT *
FROM Cosuri
GROUP BY articol
HAVING COUNT(*) >= s;
```

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

31

Efect a-priori

- Cererea corectă care returnează doar linile continând articolele frecvente din cosuri este:
- ```
SELECT * FROM COSURI
WHERE articol IN
(SELECT articol // articole
FROM Cosuri // frecvente
GROUP BY articol
HAVING COUNT(*) >= s);
```

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

32

## Algoritmul a-priori

- Oprea algoritmului se face:
1. În cazul în care nici una dintre multimile din  $C_{i+1}$  nu are un conținut de apariții mai mare decât pragul de suport.

Sau

2. Nu putem forma nici un element al multimii  $C_{i+1}$  din elementele lui  $L_i$ .
- De exemplu dacă  $L_2 = \{(1, 2), (1, 3)\}$  nu se poate găsi nici un triplet  $(a, b, c)$  cu toate submultimile de 2 elemente în  $L_2$

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

27

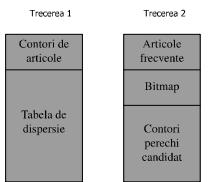
## Efect a-priori

- Sa considerăm cererea SQL din slide-ul anterior cu ipotezele:
- Utilizează tabela *Cosuri* (*IdCos*, *articol*)
  - având  $10^8$  tupluri
  - care contin date despre  $10^7$  cosuri
  - de către 10 articole fiecare,
  - Presupunem existența a 100,000 articole diferite (tipic pentru o reteaua ca Wal-Mart de exemplu).

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

29

## PCY



37

## Trecerea 1

- Se numără aparițiile fiecarui articol
- Pentru fiecare cos constant din articolele  $\{i_1, \dots, i_n\}$ , se aplică funcția de dispersie fiecărei perechi asociind-o unei intrări a unei tabele de dispersie și se incrementă conținutul acestora cu 1.
- La sfârșitul trecerii, se determină  $L_1$ , articolele cu conținut cel puțin  $s$ .

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

38

## Trecerea 2

- Oparece  $(i, j)$  poate fi candidat în  $C_2$  doar dacă toate conditiile următoare sunt adevărate:
1.  $i$  este în  $L_1$ ,
  2.  $j$  este în  $L_1$ ,
  3.  $(i, j)$  este dispersată într-o intrare frecventă (se utilizează bitmapul)
- Ultima condiție diferențiază PCY de a-priori clasice și reduce necesarul de memorie în trecerea 2.

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

43

## Trecerea 2

- În timpul trecerii 2 luăm în considerare fiecare cos și fiecare pereche de articole din el, efectuând testul de mai sus.
- Dacă o pereche îndeplinește toate cele trei condiții, se incrementă conținutul acesteia din memorie sau se crează unul dacă acesta nu există încă.
- În final perechile numărate care au un conținut cel puțin  $s$  formează  $L_2$ .

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

44

## Trecerea 1

- De asemenea la sfârșitul trecerii se determină celele intrări din tabela de dispersie care au conținutul cel puțin egal cu  $s$  (intrări frecvente).
- Punct cheie: o pereche  $(i, j)$  nu poate fi frecventă decât dacă este dispersată într-o intrare frecventă astfel încât perechile care sunt dispersate în alte intrări NU POT fi candidata în  $C_2$ .

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

39

## Trecerea 1

- Se înlocuiește tabela de dispersie cu un bitmap având un bit per intrare: 1 dacă intrarea a fost frecventă, 0 altfel.
- Acest bitmap va fi folosit la trecerea 2 prin date.

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

40

## Q & A

- Q: Când este mai performant PCY decât a-priori?
- A: Cand sunt prea multe perechi de articole din  $L_1$  pentru a încapea într-o tabela de perechi candidate și de conținut asociați în memoria centrală iar numărul de intrări frecvente din algoritmul PCY este suficient de mic pentru a reduce dimensiunea lui  $C_2$  suficient pentru a încapea în memoria centrală (chiar și fără 1/16 din ea consumată de bitmapul).

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

45

## Q & A

- Q: Cand o mare parte a intrărilor nu vor fi frecvente în PCY?
- A: Cand sunt puține perechi frecvente iar cea mai mare parte a perechilor sunt atât de puțin frecvente încât chiar dacă sumam conținutul tuturor perechilor care sunt dispersate într-o intrare data nu sunt mari sanse să se obțină o valoare egală sau mai mare ca  $s$ .

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

46

## Trecerea 2

- Memoria centrală conține o listă cu toate articolele frecvente, i.e.  $L_1$ .
- Tot memoria centrală conține un bitmap reprezentând rezultatele dispersiei din prima trecere.
- Punct cheie: intrările utiliză 16 sau 32 de biți pentru un conținut dar sunt comprimate la un singur bit.

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

41

## Trecerea 2

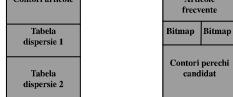
- Astfel, chiar dacă tabela de dispersie ocupă aproape întregă memoria centrală la prima trecere, bitmapul nu ocupă mai mult de 1/16 din memoria centrală la trecerea 2.
- În final, memoria centrală conține de asemenea o tabelă cu toate perechile candidat și conținutul asociați lor.

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

42

## Tabele de dispersie multiple

- Varianta a PCY. Se folosesc mai multe funcții de dispersie.
- Se împarte memoria între două sau mai multe tabele de dispersie, ca în figura următoare:



F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

47

## Tabele de dispersie multiple

- La trecerea 2 se retine în memoria cată un bitmap pentru fiecare dintre acestea; de notat că spațiul necesar pentru aceste bitmape este exact același cu cel necesar în bitmapul unic din PCY deoarece numărul total de intrări reprezentate este același.
- Pentru a fi candidata la  $C_2$  o pereche:
1. Conține articole din  $L_1$ , și
  2. Este dispersată într-o intrare frecventă în fiecare tabelă de dispersie.

F. Radulescu, Curse Utilizarea bazei de date, anul IV CS.

48

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Tabele de dispersie iterate</b></p> <ul style="list-style-type: none"> <li>□ Tabele de dispersie iterate <i>Multistage</i>:</li> <li>□ Ca la PCY dar in locul verificarii candidatelor in trecerea 2 se creaza o alta tabela de dispersie (alta functie de dispersie) si sunt dispersate doar acele perechi care indeplinesc conditiile pentru PCY; i.e., ambele articole sunt din L<sub>1</sub>, si sunt dispersate intr-un intrare frecventa in prima trecere.</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 49</p>                                        | <p><b>Tabele de dispersie iterate</b></p> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 50</p>                                                                                                                                                                                                                                                                                                                                                                                                              | <p><b>Abordarea simplă</b></p> <ul style="list-style-type: none"> <li>□ <i>Abordarea simplă</i>: Se ia un esantion de date de dimensiunea memoriei centrale.</li> <li>□ Se ruleaza un algoritm nivel cu nivel in memoria centrala (deci nu sunt costuri de I/O) si</li> <li>□ Se spera ca esantionul ne va conduce la adevaratele multimii frecvente.</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 55</p>                                                                         | <p><b>Abordarea simplă</b></p> <ul style="list-style-type: none"> <li>□ De notat ca pragul s trebuie scapat prin micsorare; e.g. daca esantionul este 1% din date, se utilizeaza s/100 ca prag de suport (in valoare absoluta).</li> <li>□ Se poate face o trecere completa prin date pentru a verifica daca multimile frecvente de articole ale esantionului sunt cu adevarat frecvente,</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 56</p> |
| <p><b>Tabele de dispersie iterate</b></p> <ul style="list-style-type: none"> <li>□ Intr-o treia trecere, pastram cate un bitmap pentru fiecare tabela de dispersie si tratam o pereche ca fiind candidata (in C<sub>3</sub>) doar daca:</li> <ul style="list-style-type: none"> <li>□ Ambele articole sunt in L<sub>1</sub>.</li> <li>□ Perechea a fost dispersata intr-o intrare frecventa in prima trecere.</li> <li>□ Perechea a fost dispersata de asemenea intr-o intrare frecventa in trecerea 2.</li> </ul> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 51</p> | <p><b>Q &amp; A</b></p> <ul style="list-style-type: none"> <li>□ Q: Cand sunt utile tabelele de dispersie multiple?</li> <li>□ A: Cand cele mai multe dintre intrari de la prima trecere a PCY au contori cu mult sub pragul s. Atunci putem dubla contorii intrarilor si totusi cele mai multe vor fi sub prag.</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 52</p>                                                                                                                            | <p><b>Abordarea simplă</b></p> <ul style="list-style-type: none"> <li>□ Vor fi pierdute insa multimile de articole care sunt frecvente in ansamblul datelor dar nu in esantion.</li> <li>□ Pentru a minimiza falsele negative se poate scadea putin pragul pentru esantion gasindu-se mai multe candidate pentru trecerea prin ansamblul datelor.</li> <li>□ Risc: prea multe candidate pentru a incepea in memoria centrala.</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 57</p> | <p><b>SON95</b></p> <ul style="list-style-type: none"> <li>□ <i>SON95</i> (Savasere, Omnickinski and Navathe in VLDB 1995; referit de Toivonen).</li> <li>□ Se citesc submultimi (transfe) ale datelor in memoria centrala aplicandu-se abordarea simpla pentru descoperirea multimilor candidat.</li> <li>□ Fiecare cos este parte a uneia dintre aceste submultimi.</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 58</p>                     |
| <p><b>Q &amp; A</b></p> <ul style="list-style-type: none"> <li>□ Q: Cand sunt utile tabelele de dispersie iterate?</li> <li>□ A: Cand numarul intrarilor frecvente din prima trecere este mare (e.g. 50%) - dar nu toate, Atunci, a doua dispersie cu unele dintre perechile ignorante poate reduce semnificativ numarul de intrari.</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 53</p>                                                                                                                                                                          | <p><b>Toate multimile in 2 treceri</b></p> <ul style="list-style-type: none"> <li>□ Metodele de mai sus sunt cele mai bune cand dorim perechile frecvente, cazul cel mai comun.</li> <li>□ Daca dorim toate multimile frecvente maximale de articole, inclusiv multimile mari, sunt necesari prea multi pasi.</li> <li>□ Există mai multe abordări pentru obținerea tuturor multimilor frecvente de articole în două treceri sau mai puțin.</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 54</p> | <p><b>SON95</b></p> <ul style="list-style-type: none"> <li>□ In a doua trecere o multime este candidata daca a fost identificata ca si candidata in una sau mai multe submultimi ale datelor.</li> <li>□ Punct cheie: O multime nu poate fi frecventa in ansamblul datelor daca nu este frecventa in cel putin o submultime a acestora (oare de ce?).</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 59</p>                                                                         | <p><b>Toivonen</b></p> <ul style="list-style-type: none"> <li>□ Se ia un esantion care incape in memoria centrala.</li> <li>□ Se ruleaza abordarea simpla pe aceste date dar cu un prag micsorat astfel incat sa fie imposibila pierderea vreunei adevarate multimii frecvente de articole (e.g. daca esantionul este de 1% din date se foloseste s/125 ca prag de suport).</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 60</p>               |
| <p><b>Toivonen</b></p> <ul style="list-style-type: none"> <li>□ Se adauga candidatelor din esantion <i>marginea negativa</i>: cele multimii de articole S astfel incat S nu este identificata ca frecventa in esantion dar orice submultime stricta maxima a lui S este identificata astfel.</li> <li>□ De exemplu, daca ABCD nu este frecventa in esantion dar ABC, ABD, ACD si BCD sunt frecvente in esantion, atunci ABCD este in marginea negativa.</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 61</p>                                                       | <p><b>Toivonen</b></p> <ul style="list-style-type: none"> <li>□ Se face o trecere prin date, numarand toate multimile frecvente de articole si marginea negativa.</li> <li>□ Daca nici o multime din marginea negativa nu este frecventa in ansamblul datelor, atunci multimile frecvente de articole sunt exact cele candidate care sunt deasupra pragului.</li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 62</p>                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <p><b>Bibliografie</b></p> <ul style="list-style-type: none"> <li>□ J.D.Ullman - CS345 --- Lecture Notes, primele doua capitolte (Overview of Data Mining, Association-Rules, A-Priori Algorithm)<br/><a href="http://infolab.stanford.edu/~ullman/cs345-notes.htm">http://infolab.stanford.edu/~ullman/cs345-notes.htm</a></li> </ul> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 63</p>                                                                                                                                                                                   | <p><b>Sfârșitul</b><br/><b>capitolului 8</b></p> <p>F Radulescu, Curs Utilizarea bazelor de date, anul IV CS. 64</p>                                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |

## Capitolul 9

### Data mining – date corelate

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

1

#### Reprezentarea datelor

- ❑Vom continua să considerăm modelul de date "coșuri de produse" și vom vizualiza datele ca o matrice booleană unde:
  - > linii=coșuri și
  - > coloane=articole.

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

2

#### Aplicații

- 2. Linile și coloanele sunt pagini de web:  
 $(r, c) = 1$  înseamnă că pagina corespunzătoare coloanei  $c$  conține legături către pagina liniei  $r$ .
- ❑Acum, coloane similare pot reprezenta copii multiple (mirror) ale unei pagini.

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

3

#### Aserțiuni

1. Matricea este foarte rară; aproape peste tot 0.
2. Numărul de coloane (articole) este suficient de mic pentru a putea stoca în memoria centrală ceea ceva per coloană dar suficient de mare astfel încât nu putem stoca ceva per perche de coloane în memoria centrală (aceeași așteptare pe care am făcut-o până acum privind regulile de asociere).

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

5

#### Aserțiuni

3. Numărul de linii este atât de mare încât nu putem stoca întreaga matrice în memoria chiar profitând de faptul că e rară și comprimând-o (din nou aceeași așteptare ca întotdeauna).
4. Nu suntem interesați de perchele sau mulțimile de coloane cu larg suport; în schimb dorim perchele de coloane puternic corelate (similare).

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

4

#### Aplicații

- 4. linile sunt propoziții iar coloanele sunt pagini web.
- ❑Coloane similare pot fi pagini despre același domeniu.

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

9

#### Similaritate

- ❑Am vorbit despre similaritatea coloanelor fără să dam o măsură cantitativă a acesteia.
- ❑Definiție: Să ne gândim la o coloană ca la multimea linilor pentru care coloana conține 1. Atunci **similaritatea** a două coloane  $C_1$  și  $C_2$  este  $\text{Sim}(C_1, C_2) = |C_1 \cap C_2| / |C_1 \cup C_2|$ .

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

10

#### Aplicații

- ❑Aplicațiile de marketing sunt interesante doar de produsele de larg consum (nu merită să încercăm promovarea obiectelor pe care oricum nu le cumpără mai nimănui).
- ❑Există însă un număr de aplicații care se potriveșc cu modelul de mai sus, de interes fiind în special problema **percheilor** de coloane / articole inclusiv de consum restrâns dar puternic corelate:

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

5

#### Aplicații

1. Linile și coloanele sunt pagini de web:  
 $(r, c) = 1$  înseamnă că pagina corespunzătoare liniei  $r$  conține legături către pagina coloanei  $c$ .
- ❑Coloane similare pot fi pagini despre același domeniu.
- ❑Obs: Coloane similare = au 1 în cam aceeași poziție.

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

#### Similaritate

- Unde:  
➢ $|x|$  reprezintă cardinalul multimii  $x$ , deci:  
➢ $|C_1 \cap C_2|$  reprezintă numărul de linii în care ambele coloane au valoarea 1  
➢ $|C_1 \cup C_2|$  reprezintă numărul de linii în care cel puțin una dintre coloane are valoarea 1.

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

11

#### Exemplu

|   |   |
|---|---|
| 0 | 1 |
| 1 | 0 |
| 1 | 1 |
| 0 | 0 |
| 1 | 1 |
| 0 | 1 |

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

12

#### Problema

- ❑Deoarece matricea conține foarte multe linii și coloane, testarea similarității este laborioasă (matricea nu încape în memoria centrală).
- ❑Ar fi preferabil ca fiecare coloană să fie reprezentată de o cantitate mult mai mică de informație având proprietatea că testul de similaritate efectuat pe această sa fie relevant pentru similaritatea coloanelor.

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

13

#### Signature

- ❑Idee principală: Se mapează ("disperează") fiecare coloană  $C$  într-o cantitate mică de date numita **signature** lui  $C$ , (notată  $\text{Sig}(C)$ ) astfel încât:
- $\text{Sig}(C)$  este suficient de mică pentru ca signaturele tuturor coloanelor să încapă în memoria centrală și să se poată efectua testul de similaritate.

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

#### Convenție utilă

- ❑De asemenea vom utiliza  $a$  pentru "numărul de linii de tip  $a$ ", ș.a.m.d.
- ❑De notat că  $\text{Sim}(C_1, C_2) = a / (a + b + c)$ .
- ❑Dar cum cele mai multe linii sunt de tip  $d$ , într-o selecție de, să spunem, 100 de linii alese aleator toate vor fi de tip  $d$ , deci similaritatea coloanelor doar pentru aceste 100 de linii nici nu este definită.

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

19

#### Ce vom prezenta în continuare:

- a. **Dispersia de tip Min** și **dispersia de tip k-min** – metode de calcul signature
- b. **Dispersia sensibilă la localizare (DSL)** – o metodă prin care se minimizează numărul de perechi de signature care se testează pentru similaritate

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

20

#### Signature

- ❑Când un mod de calcul pentru signature este bun?
- Când coloanele  $C_1$  și  $C_2$  sunt puternic similare dacă și numai dacă  $\text{Sig}(C_1) \approx \text{Sig}(C_2)$  sunt puternic similare (dar de notat că este nevoie să definim "similaritatea" pentru signature).

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

32

#### Exemplu (prost)

- ❑O idee - care însă nu funcționează:
  - > Se iau aleator 100 de linii și șiruul de 100 de biti ai coloanelor pentru acele linii este signaturea fiecărei coloane.
  - ❑De ce nu funcționează?

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

#### Dispersia de tip Min (Minhashing)

- ❑Este o metodă de calcul signature
- ❑Signature unei coloane va fi un sir de numere întregi.
- ❑Să ne imaginăm linile permute într-o ordine aleatoare. "Disperează" fiecare coloană  $C$  în  $\text{Sig}(C)$ , numărul primei linii în care coloana  $C$  are un 1.
- ❑În felul acesta obținem primul întreg al signaturei

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

21

#### Dispersia de tip Min (Minhashing)

- ❑Probabilitatea ca  $\text{H}(C_1) = \text{H}(C_2)$  este  $a / (a + b + c)$  deoarece valoarea de dispersie este aceeași dacă prima linie cu un 1 în vreuna din coloane este de tip  $a$  și este diferită dacă prima astfel de linie este de tip  $b$  sau  $c$ .
- ❑De notat că această probabilitate este aceeași cu  $\text{Sim}(C_1, C_2)$ .

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

22

#### Exemplu (prost)

- ❑Motivul pentru care ideea nu funcționează este că matricea este presupusă că fiind **foarte rară** deci multe coloane vor avea signature identice formate doar din 0 chiar dacă ele nu sunt deloc similare,

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

33

#### Convenție utilă

- ❑Dându-se două coloane  $C_1$  și  $C_2$ , ne vom referi la linile lor ca fiind de patru tipuri –  $a$ ,  $b$ ,  $c$ ,  $d$  – în funcție de bitii lor pe aceste coloane, după cum urmează:

| Tip | C1 | C2 |
|-----|----|----|
| a   | 1  | 1  |
| b   | 1  | 0  |
| c   | 0  | 1  |
| d   | 0  | 0  |

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

18

#### Dispersia de tip Min (Minhashing)

- ❑Dacă repetăm experimentul cu o nouă permutare a linilor de un număr mare de ori, să zicem 100, obținem o signature constantă din 100 de numere de linii pentru fiecare coloană.
- ❑"Similaritatea" acestor liste (fractiunea a pozitilorlor în care ele sunt egale) va fi foarte apropiată de similaritatea coloanelor.

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

22

#### Dispersia de tip Min (Minhashing)

- ❑Observație importantă: nu trebuie să permitem fizic linile, ceea ce ar duce la multe treceri prin întreaga cantitate de date.
- ❑În schimb cînd linile într-o ordine oarecare și dispersăm fiecare linie (numărul acesteia) utilizând (să zicem) 100 de funcții de dispersie diferite.

F.Rădulescu, Curs Universitatea Bucureşti de date și VCS

24



### DSL Hamming

- ❑ O a doua trecere prin datele originale confirmă care dintr-o candidată sunt întrăveăr similare.
- ❑ Aceasta metodă **exploatează** o idee care poate fi folosită și în alte parti: coloanele similare au un număr similar de 1, deci nu are rost compararea coloanelor al căror număr de 1 este foarte diferit

F. Radulescu, Curs Universitar  
de date, anul IV/CS

49

### Bibliografie

- ❑ J.D.Ullman - CS345 ---- Lecture Notes, Capitolul 3  
(Overview of Data Mining, Association-Rules, A-Priori Algorithm)  
<http://csail.mit.edu/~ullman/cs345notes.html>

F. Radulescu, Curs Universitar  
de date, anul IV/CS

51

50

### Sfârșitul capitolului 9

F. Radulescu, Curs Universitar  
de date, anul IV/CS

51

## Capitolul 10

### Algoritmi de clasificare

#### Cuprins

- ❑ Ce este învățarea supervizată
- ❑ Evaluare clasificatori
- ❑ Arbori de decizie: ID3 și C4.5
- ❑ Clasificator bayesien (Naive Bayes)
- ❑ Mașini cu vectori suport (Support vector machines)
- ❑ K-nearest neighbors
- ❑ Metode asambliste: Bagging, Boosting, Random Forest

 2

#### Formatul datelor de intrare

- ❑ În majoritatea cazurilor, setul de antrenament (precum și setul de test / validare sau de test) pot fi reprezentate ca un tabel având o coloană pentru fiecare atribut al lui X, iar ultima coloană conține eticheta clasei.
- ❑ Un exemplu foarte folosit este Play-tennis în care sunt folosite condițiile meteorologice pentru a decide dacă jucătorii pot sau nu începe un nou joc.
- ❑ Acest set de date va fi utilizat și în acest curs.

 3

#### Setul Play tennis

| Day      | Outlook  | Temperature | Humidity | Wind | Play Tennis |
|----------|----------|-------------|----------|------|-------------|
| Sunny    | Hot      | High        | Strong   | No   |             |
| Sunny    | Hot      | High        | Weak     | Yes  |             |
| Overcast | Hot      | High        | Weak     | Yes  |             |
| Rain     | Hot      | High        | Weak     | Yes  |             |
| Rain     | Moderate | High        | Weak     | Yes  |             |
| Rain     | Moderate | High        | Strong   | No   |             |
| Overcast | Cool     | Normal      | Strong   | Yes  |             |
| Rain     | Cool     | Normal      | Weak     | Yes  |             |
| Sunny    | Cool     | Normal      | Weak     | Yes  |             |
| Rain     | Moderate | Normal      | Weak     | Yes  |             |
| Rain     | Moderate | High        | Weak     | Yes  |             |
| Sunny    | Moderate | High        | Weak     | Yes  |             |
| Rain     | Cool     | Normal      | Strong   | Yes  |             |
| Sunny    | Cool     | Normal      | Strong   | Yes  |             |
| Rain     | Moderate | High        | Strong   | Yes  |             |
| Rain     | Moderate | High        | Weak     | No   |             |

 4

#### Obiective

- ❑ Învățarea supervizată este un subdomeniu foarte studiat din Data Mining
- ❑ În acest caz se construiesc noi modele din experiențe trecute (date)

 5

#### Definiții

- ❑ Învățarea supervizată include:
  - ❑ Clasificare: rezultatele sunt valori discrete (obiectiv: identificarea apartenenței la un grup / clasă).
  - ❑ Regresie: rezultatele sunt valori continue sau ordonate (obiectiv: estimare sau prezicerea unui răspuns).
- ❑ În acest capitol ne concentrăm pe clasificare

 6

#### Arbori de decizie

- ❑ Arbori de decizie: un exemplu este arborele de decizie UPU așa cum este descris în exemplul precedent.
- ❑ Într-un arbore de decizie nodurile interne conțin decizii bazate pe valorile atributelor exemplilor (atributul de argumentul  $x_j$  și fiecare frunză  $y_i$  este o clasă din C).
- ❑ ID3 și C4.5 sunt doi algoritmi cunoscuți pentru construirea arborilor de decizie și sunt prezenți în acest curs.

 7

#### Clasificatori bayesieni

- ❑ Clasificare bayesiană (Naive Bayes): pentru fiecare exemplu  $x_i$ , metoda calculează probabilitatea pentru fiecare clasă  $y_j$  din C.
- ❑ Clasificarea se face alegând cea mai probabilă clasă pentru fiecare exemplu.
- ❑ Cuvântul naiv este folosit deoarece se fac unele presupuneri simplificatoare.

 8

#### Clasificarea

- Întrare:**
- Un set de clase  $C = \{c_1, c_2, \dots, c_k\}$  în număr de  $k$ .
  - Un set de  $n$  articole etichetate  $D = \{(d_1, c_1), (d_2, c_2), \dots, (d_n, c_n)\}$ . Articolele sunt  $d_1, \dots, d_n$ , fiecare articol  $d_i$  fiind etichetat cu o clasa  $c_i$  din C, C se numește **etichetă de antrenare**.
  - Pentru că clasificarea folosește un algoritm, este necesara o mulțime de validate. Aceasta conține de asemenea elemente etichetate care nu sunt incluse în setul de antrenare.
- Ieșire:**
- Un model sau o metodă pentru clasificarea articolelor noi, Mulțimea de articole pe care vor fi clasificate folosind modelul / metoda obținută se numește **mulțimea/test**.

 9

#### Exemplu. Model: arbore de decizie



 10

#### Mașini cu vectori suport

- ❑ Mașini cu vectori suport: această metodă este utilizată pentru clasificarea binară.
- ❑ Exemplile sunt clasificate în doar două clase: fie pozitive, fie negative.
- ❑ Este o metodă foarte eficientă și poate fi folosită recursiv pentru a construi un clasificator cu mai mult de două clase.

 11

#### KNN

- ❑ Cei mai apropijați K vecini (K-nearest neighbors): o metodă foarte simplă, dar puternică pentru clasificarea exemplilor bazată pe etichetele vecinilor.

 12

#### Metode asambliste

- ❑ Metode asambliste: Random Forest, Bagging și Boosting.
- ❑ În aceste cazuri, se construiesc mai mulți clasificatori iar clasificarea finală se face prin agregarea rezultatelor acestora.
- ❑ De exemplu, un clasificator de tip Random Forest constă din mai mulți arbori de decizie, iar clasa de ieșire este valoarea modală (cea mai frecventă) a claselor returnate de arborii individuali.

 13

#### Cuprins

- ❑ Ce este învățarea supervizată
- ❑ Evaluare clasificatori
- ❑ Arbori de decizie: ID3 și C4.5
- ❑ Clasificator bayesien (Naive Bayes)
- ❑ Mașini cu vectori suport (Support vector machines)
- ❑ K-nearest neighbors
- ❑ Metode asambliste: Bagging, Boosting, Random Forest

 14

#### Alte măsuri

- ❑ Formula acurateței poate fi recrisă astfel:

$$\text{Accuracy} = \frac{\text{Recall} * \text{Pos}}{\text{Pos} + \text{Neg}} + \frac{\text{Specificity} * \text{Neg}}{\text{Pos} + \text{Neg}}$$

unde Pos și Neg sunt numărul total de exemple pozitive și negative.

 15

 16

#### Acuratețea și rata de eroare

- ❑ Pentru estimarea eficienței unui clasificator pot fi utilizate mai multe măsuri:
- ❑ Acuratețea (sau acuratețea predictivă) este proporția exemplilor de test corect clasificați de model (sau metodă).
- ❑ Numărul total de exemple de test
- ❑ Rata de eroare este proporția exemplilor de test clasificați incorrekt:

$$\text{Rata de eroare} = 1 - \text{Acuratețea}$$

 17

#### Alte măsuri

- ❑ În unele cazuri exemplii sunt clasificați în doar două clase (pozitive și negative). Atunci pot fi definite și alte măsuri.
- ❑ Să considerăm matricea de confuzie care conține numărul de exemple clasificate corect și incorrect (pentru exemplare pozitive, precum și pentru exemplare negative):

|                 | Classified as Positive | Classified as Negative |
|-----------------|------------------------|------------------------|
| Actual Positive | TP = True Positive     | FN = False Negative    |
| Actual Negative | FP = False Positive    | TN = True Negative     |

 18

#### Alte măsuri

- ❑ Atunci precizia  $p = 100\%$  dar sensibilitatea  $r = 30\%$ . Putem combina precizia și sensibilitatea obținând scorul  $F_1$ -score care este media armonică:

$$F_1\text{-score} = \frac{2}{\frac{1}{r} + \frac{1}{p}} = \frac{2pr}{p+r}$$

❑ Pentru exemplul anterior  $F_1$ -score = 46%; în general  $F_1$ -score este mai aproape de valoarea mai mică a celor două.

 19

 20

#### Alte măsuri

- Unde:
- ❑ TP = numărul de clasificări corecte pentru exemple pozitive.
  - ❑ TN = numărul de clasificări corecte pentru exemple negative.
  - ❑ FN = numărul de clasificări incorrecte pentru exemple negative.
  - ❑ FP = numărul de clasificări incorrecte pentru exemple pozitive.
  - ❑ Precizia este proporția exemplilor pozitivi clasificați corect din mulțimea exemplilor clasificate ca pozitive.
  - ❑ Precizia =  $TP / (TP + FP)$

 21

#### Alte măsuri

- ❑ Sensibilitatea (Recall, sensitivity) este proporția de exemple pozitive corect clasificate în mulțimea exemplilor pozitive (sau rata de recunoaștere a exemplilor pozitivi):
- ❑ Sensibilitatea =  $TP / (TP + FN)$
- ❑ Specificitatea (Specificity) este rata de recunoaștere a exemplilor negativi:
- ❑ Specificitatea =  $TN / (TN + FP)$

 22

#### Metoda Holdout

- ❑ În acest caz, setul de D se împarte în două: un set de antrenare și un set de test.
- ❑ Setul de test se numește mulțimea reținută - **holdout set** (de aici numele metodei).
- ❑ Clasificatorul obținut folosind setul de antrenare este utilizat pentru clasificarea exemplilor din setul de test.
- ❑ Deoarece acesti exempli sunt etichetate se pot calcula acuratețea, precizia, sensibilitatea și alte măsuri și pe baza lor este evaluat clasificatorul.

 23

#### Validarea încricușată

- Există mai multe versiuni de validare încricușată:
1. **Validare încricușată folosind k subseturi - k-fold cross validation.**
- Setul de date D este împărțit în k subseturi disjuncte cu aceeași dimensiune.
- Pentru fiecare subsecție construiește un clasificator folosind acel subsecție ca set de test și reunindu-seturile rămase ca set de antrenare.
- In acest fel se obțin k valori pentru acuratețe (una pentru fiecare clasificator). Media acestor valori este acuratețea finală. Valoarea obținută pentru k este 10.

 24

## Validarea încriușată

- 2. Validare încriușată folosind 2 subseturi - *2-fold cross validation*.**
- Validation.** Pentru  $k = 2$ , metoda de mai sus are avantajul de a folosi seturi mari atât pentru antrenament, cât și pentru testare.

✓ Validarea încriușată este mai bună.

## Validarea încriușată

- 3. Validare încriușată stratificată - *Stratified cross validation*.**
- Este o variație de validare încriușată având  $k$  subseturi. Fiecare subset are aceeași distribuție a etichetelor.
- De exemplu, pentru exemple pozitive și negative, fiecare subset conține aproximativ aceeași proporție de exemple pozitive și aceeași proporție de exemple negative.

26

✓ Validarea încriușată stratificată este mai bună.

## Cuprins

- Ce este învățarea supervizată
- Evaluare clasificatori
- Arbori de decizie: ID3 și C4.5
- Clasificatori bayesieni (Naïve Bayes)
- Mașini cu vectori suport (Support vector machines)
- K-nearest neighbors
- Metode asambliste: Bagging, Boosting, Random Forest

27

✓ Validarea încriușată stratificată este mai bună.

## Ce este un arbore de decizie?

- Un mod comun de a reprezenta un model sau un algoritm de clasificare este un arbore de decizie.
- Având un set de antrenare  $D$  și un set  $A$  de attribute ale exemplelor, fiecare exemplu etichetat din  $D$  are forma:  $(a_1 = v_1, a_2 = v_2, \dots, a_n = v_n)$ .
- Pe baza acestor attribute se poate construi un arbore de decizie astfel:

32

✓ Validarea încriușată stratificată este mai bună.

## Validarea încriușată

- 4. Validație încriușată având subseturi de un element - *Leave one out cross validation*.** Când  $D$  conține doar un număr mic de exemple, o valdare încriușată specială poate fi folosită: fiecare exemplu devine set de test și toate celelalte exemple sunt setul de antrenare.

□ Accuratețea pentru fiecare clasificator este fie 100%, fie 0%.

Media tuturor acestor valori este accuratețea finală.

✓ Validarea încriușată stratificată este mai bună.

## Metoda bootstrap

- Metoda **bootstrap**, face parte din metodele cu eșantionare și constă în obținerea setului de antrenare din setul de date original prin eșantionare cu înlocuire (eșantionul nu este înțăritură din set).
- Cazurile care nu sunt alesă în setul de antrenare sunt utilizate ca set de teste.
- De exemplu, dacă  $D$  are 1000 de exemple etichetate, alegând întărimpare un exemplu de 1000 de la setul de date original, în acesta unele exemple sunt prezente de mai multe ori.

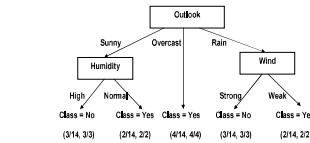
28

✓ Validarea încriușată stratificată este mai bună.

## Ce este un arbore de decizie?

1. Nodurile interne sunt attribute (nicio călărie conține de două ori același atribut).
2. Ramurile sunt lărgiri la valori de deosebit (una sau mai multe) sau intervale pentru aceeași atribut. Uneori pot fi utilizate condiții mai complexe pentru ramificare.
3. Frunzele sunt etichetate cu clase. Pentru fiecare frunză se poate calcula un suport și o probabilitate, suportul este proporția de exemple care au acelăși atribut ca și rădăcina. Probabilitatea este incidența este accuratețea clasificării pentru exemplul care se potrivește cu acea călărie. La trecerea de la arbore de decizie la reguli, fiecare regulă are același suport și aceeași incidență ca frunza din care provine.
4. Orice exemplu se poate verifica cu o singură călărie a arborelui (deci o singură frunză = clasa).

## Exemplu



34

✓ Validarea încriușată stratificată este mai bună.

## De ce 63,2%?

- Din "Data Mining: Concepts and Techniques", să presupunem că avem un set de date cu 14 de exemple. Fiecare exemplu are o probabilitate de 1/14 de a fi selectat, deci probabilitatea de a fi ales este (1/14)<sup>14</sup>.
- Cum alegem de ori, probabilitatea ca un exemplu să nu fie ales în tot acest timp este (1-1/14)<sup>14</sup>.
- Dacă d este mare, probabilitatea se apropiie de  $e^{-d}$  = 0,368. Astfel, 36,8% din exemple nu vor fi selectate pentru setul de antrenare și vor ajunge în setul de test iar restul de 63,2% vor forma setul de antrenare.

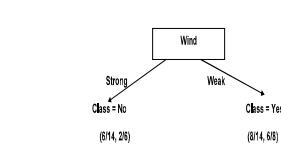
\* J. W. Han, M. Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Morgan Kaufmann, pagina 362.

✓ Validarea încriușată stratificată este mai bună.

## Exemplu

- Numerele de pe ultima linie sunt suportul și incidența asociate fiecărei frunze.
- Pentru același set de date pot fi creați mai mulți arbori de decizie.
- De exemplu un alt arbore de decizie pentru același set de date se arată în figura următoare (cu valori mai mici ale incidenței decât arborele anterior):

## Decision trees



✓ Validarea încriușată stratificată este mai bună.

## Metoda bootstrap

- Statistică, 63,2% din exemplele din  $D$  sunt alesă pentru setul de antrenare iar 36,8% nu. Aceste 36,8% devin setul de test.
- După construirea unui clasificator și urmarea acestuia pe setul de test este determinată accuratețea și astfel poate fi evaluată metoda de construire a clasificatorului în funcție de valoarea obținută.
- Mai multe despre această metodă și alte tehnici de evaluare pot fi găsite în [Sanderson 08].

✓ Validarea încriușată stratificată este mai bună.

## De ce 63,2%?

- Din "Data Mining: Concepts and Techniques", să presupunem că avem un set de date cu 14 de exemple. Fiecare exemplu are o probabilitate de 1/14 de a fi selectat, deci probabilitatea de a fi ales este (1/14)<sup>14</sup>.
- Cum alegem de ori, probabilitatea ca un exemplu să nu fie ales în tot acest timp este (1-1/14)<sup>14</sup>.
- Dacă d este mare, probabilitatea se apropiie de  $e^{-d}$  = 0,368. Astfel, 36,8% din exemple nu vor fi selectate pentru setul de antrenare și vor ajunge în setul de test iar restul de 63,2% vor forma setul de antrenare.

\* J. W. Han, M. Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Morgan Kaufmann, pagina 362.

✓ Validarea încriușată stratificată este mai bună.

## ID3

- ID3 (Iterativ Dichotomiser 3) este un algoritm pentru construirea arborilor de decizie introdus de Ross Quinlan în 1986 (vezi [Quinlan 86]).
- Algoritmul construiește arborele de decizie într-o manieră top-down alegând la fiecare nod atributul „cel mai bun” pentru ramificare.
- Mai întâi este ales un atribut rădăcină, construind o ramură separată pentru fiecare valoare diferență a atributului.

✓ Validarea încriușată stratificată este mai bună.

## ID3

- Setul de antrenare este de asemenea împărțit, fiecare ramură menținând exemplele care se potrivește cu valoarea atributului ramurii respective.
- Procesul se repetă pentru fiecare descendenta până că toate exemplele au aceeași clasă (în acest caz nodul devine o frunză etichetată cu acea clasă) sau toate atributele au fost utilizate (nodul devine, de asemenea, o frunză etichetată cu valoarea modală = clasa majoritară).
- Un atribut nu poate fi ales de două ori pe aceeași călărie; din momentul în care a fost ales pentru un nod, nu va mai fi niciodată testat pentru descendenții aceluia nod.

✓ Validarea încriușată stratificată este mai bună.

✓ Validarea încriușată stratificată este mai bună.

## Exemplu

- Pentru setul de date Play tennis există patru attribute posibile pentru rădăcina arborelui decizional: Aspect, Temperatură, Umiditate și Vânt.
- Entropia intregului set de date și valoare ponderată pentru împărțirea utilizării celor patru attribute sunt:

$$\text{entropy}(D) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0,94$$

✓ Validarea încriușată stratificată este mai bună.

✓ Validarea încriușată stratificată este mai bună.

## Pentru fiecare atribut

$$\text{entropy}(D, \text{Humidity}) = \frac{7}{14} \left( -\log_2 \frac{3}{7} - \log_2 \frac{4}{7} \right) + \frac{7}{14} \left( -\log_2 \frac{6}{7} - \log_2 \frac{6}{7} \right) = 0,79$$

□ La fel:

$$\text{entropy}(D, \text{Wind}) = 0,89$$

$$\text{entropy}(D, \text{Temperature}) = 0,91$$

$$\text{entropy}(D, \text{Outlook}) = 0,69$$

✓ Validarea încriușată stratificată este mai bună.

✓ Validarea încriușată stratificată este mai bună.

## Cel mai bun atribut

- Esența ID3 este modul în care este descoperit atributul „cel mai bun”. Algoritmul folosește teoria informației înserând să crească puritatea seturilor de date de la nodul tată la descendent.
- Să luăm în considerare un set de date:
- $D = \{(e_1, c_1), \dots, e_{14}, c_i\}$
- cu exemple etichetate cu clase din  $C = \{c_1, c_2, \dots, c_n\}$ .
- Atributele exemplelor sunt  $A_1, A_2, \dots, A_p$ .

✓ Validarea încriușată stratificată este mai bună.

## Entropia

- Atunci entropia lui  $D$  poate fi calculată astfel:

$$\text{entropy}(D) = -\sum_{i=1}^n \Pr(c_i) \log_2 \Pr(c_i)$$

- Dacă atributul  $A_i$ , având  $n$  valori distincte este considerat pentru ramificare, acesta va partaja setul în sub集urile disjuncte  $D_1, D_2, \dots, D_n$ .

✓ Validarea încriușată stratificată este mai bună.

✓ Validarea încriușată stratificată este mai bună.

## Cel mai bun atribut: Outlook

- Următorul tabel conține valorile pentru entropie și căstig.
- Cel mai bun atribut pentru nodul rădăcină este Aspect (Outlook), cu căstigul maxim de 0,25.

| Atribut     | entropy | gain |
|-------------|---------|------|
| Humidity    | 0,69    | 0,16 |
| Wind        | 0,89    | 0,05 |
| Temperature | 0,91    | 0,03 |
| Outlook     | 0,69    | 0,25 |

✓ Validarea încriușată stratificată este mai bună.

✓ Validarea încriușată stratificată este mai bună.

## Entropia

- Entropia combinată a acestor subseturi, calculată ca medie ponderată a acestor entropii este

$$\text{entropy}(D, A_k) = \sum_{i=1}^p \frac{\text{count}(D_i)}{\text{count}(D)} * \text{entropy}(D_i)$$

- Toate probabilitățile implicate în ecuații de mai sus sunt determinate prin numărare!

✓ Validarea încriușată stratificată este mai bună.

## Căstigul informational

- Decarece puritatea seturilor de date crește, entropia ( $D$ ) este mai mare decât căstigul ( $D, A_k$ ). Diferența dintre ele se numește căstig informational:

$$\text{Gain}(D, A_k) = \text{entropy}(D) - \text{entropy}(D, A_k)$$

- Cel mai bun atribut este cel având căstigul informational maxim.

✓ Validarea încriușată stratificată este mai bună.

✓ Validarea încriușată stratificată este mai bună.

## Discuție ID3

3. Uneori când doar câteva exemple sunt asociate cu frunze (overfitting) și nu funcționează bine la alte exemple de test.
- Pentru a evita overfitting, arborele poate fi simplificat prin reducere (pruning):
  - Pre-pruning: creșterea este opriță înainte de stărșinul normal. Frunzele nu sunt 100% pure și sunt etichetate cu clasa majoritară (valoarea medie).
  - Post-pruning: după rularea algoritmului, unele subarbore sunt înlătuți pentru exemple din setul de antrenare care se potrivesc. Tehnica post-pruning este mai bună, deoarece în prepruning este greu de estimat când trebuie să se trimită sa ne oprire.

## Discuție ID3

4. Unele attribute (de exemplu  $A$ ) pot fi continue. Valoarea pentru  $A$  pot fi impărtățite în două intervale:
  - Valoarea lui  $A$  poate fi după cum urmează:
    - Se sortează exemplul după  $A$ .
    - Se alege ca valoare candidat media a două valori consecutive pentru care clasa se schimbă.
    - Pentru fiecare valoare candidat de la pasul anterior se calculează căstigul dacă particionează se face folosind acea valoare. Candidatul cu căstig maxim este ales pentru particionare.

### Discuție ID3

- În acest fel, atributul continuu este înlocuit cu unul discret (două valori, una pentru fiecare interval).
- Acest atribut concurează cu atributul rămase pentru „cel mai bun” atribut.
- Procesul se repetă pentru fiecare nod, deoarece valoarea partilionară se poate schimba de la un nod la altul.

50

### Discuție ID3

- Ceiajul următor, unele attribute sunt mai scumpe decât altele (măsurăți pe unitatea în baza).
- Este mai bine ca attributele cu costuri mai mici să fie mai aproape de rădăcina decât cele altibile.
- De exemplu, pentru o unitate de urgență, este mai bine să testăm pușcul și temperatură mai întâi și numai atunci când este necesar să efectuăm o biopsie.
- Acest lucru se poate face prin ponderarea căștișului cu costul:

$$\text{weighted-gain}(A_k) = \frac{\text{gain}(D, A_k)}{\text{cost}(A_k)}$$

### Teorema lui Bayes

- Thomas Bayes (1701 - 1761) a fost un cleric presbiterian și un matematician englez.
- Teorema numită după el poate fi exprimată astfel:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

unde  $P(A|B)$  înseamnă probabilitatea lui A fiind dat B.

51

### C4.5

- C4.5 este versiunea îmbunătățită a ID3 și a fost dezvoltată de același Ross Quinlan. Câteva caracteristici:
- Sunt permise attribute numerice (continuе).
  - Tratează cazul valorilor lipsă.
  - Utilizează post-pruning pentru a face față la date cu zgomot.
  - Cele mai importante îmbunătățiri ale ID3 sunt:
  - Attributele sunt alesă pe baza raportului de câștig (gain ratio) și nu pur și simplu a căștișului informațional.

52

### C4.5

- Folosește post-pruning pentru a reduce dimensiunea arborului. Tărâiera se face numai dacă reduce eroarea estimată. Există două metode de tărîere:
  - Înlocuire sub-arbore: Un sub-arbore este înlocuit cu o frunză, dar fiecare sub-arbore este considerat numai după toți sub-arborii săi. Aceasta este o abordare de jos în sus,
  - Ridicare sub-arbore: Un nod este ridicat și înlocuiește un nod superior. Dar în acest caz, unele exemple trebuie reasignate. Această metodă este considerată mai putin importantă și mai lentă decât prima.

53

### Soluție

$$\begin{aligned} \Pr(\text{Opt}) &= 0.3 \\ \Pr(3 \text{ rânduri} | \text{Opt}) &= 0.4 \\ \Pr(3 \text{ rânduri}) &= 0.2 \end{aligned}$$

Deci:

$$\Pr(3 \text{ rânduri} | \text{Opt}) * \Pr(\text{Opt}) = \frac{\Pr(3 \text{ rânduri})}{\Pr(3 \text{ rânduri})} = 0.4 * 0.3 / 0.2 = 0.6 \text{ sau } 60\%$$

54

### Naive Bayes: Generalități

- Această abordare este una probabilistică.
- Algoritmii bazati pe teorema lui Bayes calculează pentru fiecare exemplu de test nu o singură clasă, ci probabilitatea pentru fiecare clasă din C (setul de clase).
- Dacă setul de date are k attribute,  $A_1, A_2, \dots, A_k$ , obiectivul este să calculem pentru fiecare clasă  $c \in C = \{c_1, c_2, \dots, c_n\}$  probabilitatea ca exemplul de test  $(a_1, a_2, \dots, a_k)$  să aparțină clasei c.

$$\Pr(\text{Class} = c | A_1 = a_1, \dots, A_k = a_k)$$

Dacă este necesară clasificare, clasă cu cea mai mare probabilitate poate fi atribuită acestui exemplu.

55

### Construirea modelului

- Facem următoarea presupunere: „toate attributele sunt condițional independente dându-se clasa  $C = c_j$  atunci:

$$\Pr(A_1 = a_1, \dots, A_k = a_k | C = c_j) = \prod_{i=1}^k \Pr(A_i = a_i | C = c_j)$$

Din cauza acestei presupuneri metoda este numită „naïvă”.

Presupunerea nu este valabilă în toate situațiile.

Practica arată însă că rezultatele obținute folosind aceasta presupunere simplificatoare sunt suficiente de bune în majoritatea cazurilor.

56

### Construirea modelului

- În final, înlocuind expresia anterioră în formula obținem:

$$C = \underset{c_j}{\operatorname{argmax}} \Pr(c_j) * \prod_{i=1}^k \Pr(A_i = a_i | C = c_j)$$

57

### Exemplu

- Fie versiunea simplificată a tabelului PlayTennis:

| Outlook  | Wind   | Play Tennis |
|----------|--------|-------------|
| Overcast | Weak   | Yes         |
| Overcast | Strong | Yes         |
| Overcast | Absent | No          |
| Sunny    | Weak   | Yes         |
| Sunny    | Strong | No          |
| Rain     | Strong | No          |
| Rain     | Weak   | No          |
| Rain     | Absent | Yes         |

58

### Caz special: valori non-categorice sau absente

Valori non-categorice sau valori absente:

- Toate attributele care nu sunt categorice trebuie discretizate (înlocuite cu unele categorice).
- De asemenea, dacă unele attribute au valori lipsă, aceste valori sunt ignorate.

59

### Exemplu

For  $C = \text{Yes}$

$$\Pr(\text{Yes}) + \Pr(\text{Sunny} | \text{Yes}) + \Pr(\text{Absent} | \text{Yes}) = \frac{4}{8} + \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

For  $C = \text{No}$

$$\Pr(\text{No}) + \Pr(\text{Sunny} | \text{No}) + \Pr(\text{Absent} | \text{No}) = \frac{4}{8} + \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Rezultatul este că ambele probabilități au aceeași valoare și clasa poate fi oricără din ele.

60

### SVM: Generalități

- Este prezentată doar ideea generală a metodei de clasificare Support Vector Machines (SVM).

SVM-urile sunt descrise în detaliu în multe documentații și cărți, de exemplu [Liu 11] sau [Han, Kamber 06].

Metoda a fost descoperită în Uniunea Sovietică în anii '70 de Vladimir Vapnik și a fost dezvoltată în SUA după ce Vapnik s-a alăturat AT&T Bell Labs la începutul anilor '90 (a se vedea [Cortes, Vapnik 95]).

61

### Caz special: împărțire la 0

- Expresia modificată este:

$$P(A_i = a_i | C = c_j) = \frac{a_i + s}{b + s + r}$$

unde:

$s = 1 / \text{Număr de exemple din setul de antrenare}$

$r = \text{Numărul valorilor distincte pentru } A_i$

În acest caz toți termenii expresiei sunt mai mari decât zero.

62

### SVM: Model

- Un posibil clasificator este o funcție liniară:

$$f(X) = \langle w \cdot X \rangle + b$$

Așa încât

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

unde:

$w$  este un vector de ponderi (weight vector),

$\langle w \cdot X \rangle$  este produsul scalar între vectorii  $w$  și  $X$ ,

$b$  este un număr real, și

$w$  și  $b$  pot fi scalari, după cum se va vedea.

63

64

### Exemplu

- Studentii din EC004 sunt 70% de la C5 și 30% optionali.
- 20% dintre studenti sunt plasati în primele 3 rânduri, dar pentru optionali acest procent este de 40%.
- Când decanul intră în sală și se așeză undeva în primele 3 rânduri, îngă un student, care este probabilitatea ca acest student să fie din categoria optional?

65

56

66

67

68

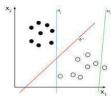
69

70

71

72

Figura 1



Sursa: Wikipedia

[Functii kernel: Ce este și cum se utilizează]

## Cel mai bun hiperplan

- ❑ SVM încercă să găsească „cel mai bun” hiperplan de acea formă.
- ❑ Teoria arată că cel mai bun plan este cel care maximizează aşa-numita “margină” (distanță ortogonală minimă între un punct pozitiv și negativ din setul de antrenament - a se vedea figura următoare pentru un exemplu).

73

## Functii kernel

- ❑ Dar cum putem găsi această funcție de mapare?
- ❑ În soluționarea problemei de optimizare pentru găsirea hiperplanului de separare liniară în nou spațiu  $F$  toți termenii sunt doar de forma  $\phi(X_i) \cdot \phi(X_j)$ .
- ❑ Înlocuind acest produs scalar cu o funcție atât în  $X_i$  cât și în  $X_j$ , nevoia de a găsi  $\phi$  dispără. O astfel de funcție se numește **funcție kernel**  $K(X_i, X_j) = \phi(X_i) \cdot \phi(X_j)$
- ❑ Pentru un hiperplanul de separare în  $F$  trebuie doar să înlocuim toate produsele scalare cu funcția kernel alesă și apoi să continuăm cu problema de optimizare ca în cazul separabil.

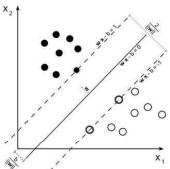
74

## Functii kernel

- ❑ Unele dintre cele mai utilizate funcții kernel sunt:
- ❑ Kernel liniar:  $K(X, Y) = \langle X \cdot Y \rangle + b$
- ❑ Kernel polinomial:  $K(X, Y) = (a * \langle X \cdot Y \rangle + b)^p$
- ❑ Kernel sigmoid:  $K(X, Y) = \tanh(a * \langle X \cdot Y \rangle + b)$

80

Figura 2



Sursa: Wikipedia

## Separarea non-liniară

- ❑ În multe situații nu există un hiperplan care separă exemplile pozitive de cele negative.
- ❑ În astfel de cazuri este posibilă maparea punctelor din setul de antrenare (exemplul din D) într-un alt spațiu dimensional superior.
- ❑ Aici punctele pot fi liniar separabile.
- ❑ Funcția de mapare primește exemple (vectori) din spațul de intrare  $X$  și le mapează în spațiu caracteristic (feature space)  $F$ :

$$\phi: X \rightarrow F$$

75

[Functii kernel: Ce este și cum se utilizează]

## Discuție SVM

- ❑ SVM se folosește când atributele au valori reale
- Când în setul de antrenare există attribute categorice, este necesară o conversie la valori reale.
- ❑ Când sunt necesare mai mult de două clase, SVM poate fi utilizat recursiv.
- Prima utilizare separă clasa 1 de clasa 2 și asta mai departe. Pentru N clase sunt necesare ruje N-1.
- ❑ SVM sunt o metodă foarte bună în clasificarea datelor hiper-dimensionale.

81

## Cuprins

- ❑ Ce este învățarea supervizată
- ❑ Evaluare clasificatori
- ❑ Arbori de decizie: ID3 și C4.5
- ❑ Clasificatori bayesieni (Naive Bayes)
- ❑ Mașini cu vectori suport (Support vector machines)
- ❑ K-nearest neighbors
- ❑ Metode asamblante: Bagging, Boosting, Random Forest

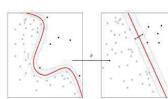
82

Separarea non-liniară

❑ Fiecare punct  $x$  din  $D$  este mapat în  $\phi(X)$ . După maparea tuturor punctelor avem alt set de antrenare conținând vectori din  $F$  și nu din  $X$ , cu  $\dim(F) \geq n = \dim(X)$ :  
 $D = \{(\phi(X_1), y_1), (\phi(X_2), y_2), \dots, (\phi(X_n), y_n)\}$

❑ Pentru un  $\phi$  corespunzător aceste puncte sunt liniar separabile,

Figura 3



76

[Functii kernel: Ce este și cum se utilizează]

## kNN

- ❑ Metoda “Ceai mai apropiat k vecinii” - K-nearest neighbors (kNN) nu produce un model, ci este o metodă simplă pentru determinarea clasei unui exemplu bazat pe etichetele vecinilor săi aparținând setului de antrenare.
- ❑ Pentru rularea algoritmului este necesară o funcție de distanță pentru calcularea distanței de la exemplul de test până la exemplele din setul de antrenare.

83

## kNN

- ❑ O funcție  $f(x, y)$  poate fi utilizată ca funcție de distanță dacă sunt îndeplinite patru condiții:

  1.  $f(x, y) \geq 0$
  2.  $f(x, x) = 0$
  3.  $f(x, y) = f(y, x)$
  4.  $f(x, y) \leq f(x, z) + f(z, y)$ .

84

Algoritm

Întrare:  
❑ Un set de date  $D$  care conține exemple etichetate (setul de antrenare)  
❑ O funcție de distanță  $f$  pentru măsurarea distanței între două exemple  
❑ Un întreg  $k >$  parametrul care spune căi vecini sunt luați în considerare  
❑ Un exemplu de test  $t$

Iesire:

❑ Clasa lui  $t$ 

Metoda:

❑ Se folosește  $f$  pentru a calcula distanța dintre  $t$  și fiecare punct din  $D$   
❑ Se selectează cei mai apropietăți  $k$  vecini  
❑ Se alege punctul cu clasa majoritară din setul de cei mai apropietăți  $k$  vecini.

## Exemplu

- ❑  $K = 3 \rightarrow$  Roșu
- ❑  $K = 5 \rightarrow$  Albastru



- ❑ kNN este foarte sensibil la valoarea parametrului  $k$ .
- ❑ Cea mai bună valoare pentru  $k$  poate fi găsită, de exemplu, prin validare încrușită.

85

[Functii kernel: Ce este și cum se utilizează]

## Bagging

- In ce constă tehnica Bagging:
- ❑ Părind de la setul de date original, construim  $n$  seturi de date de antrenare prin eșantionare cu înlocuire (bootstrap).
  - ❑ Pentru fiecare set de date de antrenare, construim un clasificator folosind același algoritm de învățare (se colță în clasificatorul slab).
  - ❑ Clasificatorul final se obține prin combinarea rezultatelor clasificatorilor slab (prin vot, de exemplu).
  - ❑ Bagging ajută la imbunătățirea preciziei pentru algoritmi instabili de învățare: arbori de decizie, rețele neurale, etc.
  - ❑ Nu ajută în cazul kNN sau Naive Bayes.

86

## Boosting

- ❑ Tehnica Boosting constă în construirea unei secvențe de clasificatori slabii și sădăgăarea lor în strucția clasificatorului final puternic.
- ❑ Clasificatorii slabii sunt ponderați în funcție de acuratețe.
  - ❑ De asemenea, datele sunt reevaluate după ce construim fiecare clasificator slab, astfel încât exemplurile clasificate incorrecții să crească în greutate.
  - ❑ Rezultatul este că următorii clasificatori slabii din secvență se concentrează mai mult pe exemplarele pe care le-au clasificat incorrecționat clasificatorii slabii precedenți.

87

Cuprins

- ❑ Ce este învățarea supervizată
- ❑ Evaluare clasificatori
- ❑ Arbori de decizie: ID3 și C4.5
- ❑ Clasificatori bayesieni (Naive Bayes)
- ❑ Mașini cu vectori suport (Support vector machines)
- ❑ K-nearest neighbors
- ❑ Metode asamblante: Bagging, Boosting, Random Forest

## Metode asamblante

- ❑ Metodele asamblante combină mai mulți clasificatori “slabi” (weak) pentru a obține unul mai bun (mai “puternic” – strong).
- ❑ Clasificatorii combinați sunt similari (utilizează aceeași metodă de învățare), dar setările de antrenare și ponderile exemplilor din ele sunt diferite.

87

[Functii kernel: Ce este și cum se utilizează]

## Random forest

- ❑ Tehnica Random forest este un clasificator asamblist format dintr-un set de arbori de decizie. Clasificatorul final produce valoarea modală a claselor de fiecare arbor.
- ❑ Algoritm este următorul:
1. Alegeri  $T$  – numărul de arbori de construit.
  2. Alegeri  $m$  – numărul de variabile utilizate pentru ramificare la fiecare nod,  $m < M$ , unde  $M$  este numărul de variabile de intrare.

88

## Random forest

1. Construim  $T$  arbori. Pentru fiecare:
- Construim un set de antrenare prin eșantionare cu înlocuire. Din el se crează un arbor de decizie.
- La fiecare nod, selecțiem în variabile la întărirea (attribute) și le folosim pentru a găsi cea mai bună ramificare.
- Crescem arborul la maxim. Nu există tăieri (pruning).
4. Se prezic rezultatele agregând predicțiile arborilor (de ex., medie pentru clasificare, medie pentru regresie).

89

Bagging

❑ Numele Bagging vine de la Bootstrap Aggregating.  
❑ Așa cum am văzut mai devreme metoda bootstrap face parte din metodele de eșantionare și constă în obținerea unui set de antrenare din datele etichetate inițiale prin eșantionare cu înlocuire.

## Exemplu

| Setul original  | a | b | c | d | e | f |
|-----------------|---|---|---|---|---|---|
| Set antrenare 1 | a | b | b | c | e | f |
| Set antrenare 2 | b | b | c | c | d | e |
| Set antrenare 3 | a | b | c | c | d | f |

90

[Functii kernel: Ce este și cum se utilizează]

## Referințe

- Liu [1] Bing Liu, 2011, *Web Data Mining: Exploring Hypertext, Contents, and Usage Data*, Springer, chapter 3.
- Han, Kamber [6] Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques*, Second Edition, Morgan Kaufmann Publishers, 2006.
- Cortes, Vapnik [5] Cortes, Corinna, and Vapnik, Vladimir N., “Support-Vector Networks”, *Machine Learning*, 20(3):273-297, 1995, <http://dx.doi.org/10.1007/BF00050937>
- Wikipedia [4] Wikipedia, free encyclopedia, en.wikipedia.org.
- Sanderson [8] Robert Sanderson, 2008, *Data mining course notes*, Dept. of Computer Science, University of Liverpool 2008, *Classification: Evaluation*, <http://www.cs.liv.ac.uk/~azeroth/courses/current/comps227/lectures/comp227-13.pdf>
- Quinlan [8] Quinlan, J. R. 1986, *Induction of Decision Trees*, *Machine Learn.*, 1, 1 (Mar. 1986), 81–106, <http://www.cs.cmu.edu/~ml/mlcourse/lec15-2008/readings/quinlan.pdf>

91

[Functii kernel: Ce este și cum se utilizează]

## Capitolul 11

### Data mining – clustering

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

1

### Problema

- Dându-se puncte într-un spațiu oarecare – deseori un spațiu cu foarte multe dimensiuni – grupează punctele într-un număr mic de *cluster*, fiecare cluster constând din puncte care sunt "apropiate" într-un anume sens.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

2

### Exemple de aplicatie

- Documentele pot fi percepute ca puncte într-un spațiu multi-dimensional în care fiecare dimensiune corespunde unui cuvânt posibil.
- Positia documentului într-o dimensiune este data de numărul de ori care cuvântul apare în document (sau doar 1 dacă apare, 0 dacă nu).
- Clusterile de documente în acest spațiu corespund deseori cu grupările de documente din același domeniu.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

### Distanta

- Pentru a discuta dacă o mulțime de puncte sunt suficiente de apropiate pentru a fi considerate un cluster avem nevoie de o *măsură a distanței*  $D(x, y)$  care spune cât de departe sunt punctele  $x$  și  $y$ .
- Nu orice funcție poate fi utilizată ca funcție de măsurare a distanței.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

8

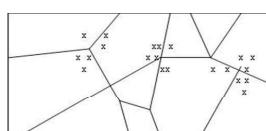
### Exemple de aplicatie

- Cu mulți ani în urmă, în timpul unei izbucniri a holerei în Londra, un medic a marcat localizarea cazurilor pe o hartă, obținând un desen care arăta ca în figura urmatoare:

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

3

### Exemple de aplicatie



F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

4

### Distanta

- Axiomile uzuale pentru o măsură a distanței  $D$  sunt următoarele:
  - $D(x, y) \geq 0$
  - $D(x, x) = 0$ . Un punct este la distanță 0 de el însuși.
  - $D(x, y) = D(y, x)$ . Distanța e simetrică.
  - $D(x, y) \leq D(x, z) + D(z, y)$ .

Inegalitatea triunghiului.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

### Exemple de aplicatie

- Deseori punctele pot fi percepute ca existând într-un spațiu euclidian  $k$ -dimensional și distanța între orice două puncte:
  - $x = [x_1, x_2, \dots, x_k]$  și
  - $y = [y_1, y_2, \dots, y_k]$

este dată într-un din manierele uzuale:

Distanța comună ("normă L<sub>2</sub>")

Distanța Manhattan ("normă L<sub>1</sub>")

Maximul pe o dimensiune ("normă L<sub>\infty</sub>")

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

10

### Exemple de aplicatie

- Vizualizate corespunzător, datele au indicat că apariția cazurilor se grupează în jurul unor intersecții, unde existau puturi infestate, arătând nu numai cauză holerică ci indicând și ce e de făcut pentru rezolvarea problemei.
- Din păcate nu toate problemele de data mining sunt atât de simple, deseori deoarece clusterurile sunt în atât de multe dimensiuni încât vizualizarea este foarte dificilă.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

5

### Exemple de aplicatie

- SkyCat* a grupat în cluster  $2 \times 10^9$  obiecte cerești în stelu, galaxii, quasari, etc.
- Fiecare obiect era un punct într-un spațiu cu 7 dimensiuni, unde fiecare dimensiune reprezintă nivelul radiației într-o bandă a spectrului.
- Proiectul Sloan Sky Survey este o încercare mult mai ambitioasă de a cataloga și grupa înregul univers vizibil.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

6

### Distanta comună

- Distanța comună (sau normă L<sub>2</sub>) este data de formula cunoscută:

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

### Distanta Manhattan

- Distanța Manhattan (sau normă L<sub>1</sub>) este data de formula următoare:

$$\sum_{i=1}^k |x_i - y_i|$$

- Ea poate fi folosită de exemplu si pentru calculul distanței între două puncte pe o placă cu circuite imprimate multistrat.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

12

### Maximul pe o dimensiune

- Este data de formula:

$$\max_{i=1}^k |x_i - y_i|$$

- Aceasta funcție verifică toate cele 4 condiții pentru a fi funcție de distanță.
- Poate fi folosită de exemplu pentru spații euclidiene hiperdimensionale (număr de dimensiuni foarte mare)

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

13

### Alte distante

- Unde nu există un spațiu euclidian în care să plăsăm punctele gruparea devine mult mai dificilă.
- Iată un exemplu în care are sens: o măsură a distanței în lipsa unui spațiu euclidian

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

### Hiperdimensionalitatea

- Cu toate acestea, să presupunem  $k$  foarte mari, să zicem 100.000. Indiferent de folosirea  $L_2$ ,  $L_1$ , sau  $L_\infty$ , suntem că:  
 $D(x, y) = \max_i |x_i - y_i|$  pentru  $x = [x_1, x_2, \dots]$  și  $y = [y_1, y_2, \dots]$ .
- Pentru  $k$  foarte mare, e foarte posibil să existe o dimensiune / astfel încât  $x$  și  $y$  sunt diferențe aproape de maximul posibil, chiar dacă  $x$  și  $y$  sunt foarte apropiate în alte dimensiuni.
- Astfel  $D(x, y)$  va fi foarte apropiată de 1.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

### Hiperdimensionalitatea

- O altă consecință interesantă a hiper-dimensionalității este că toți vectorii,  $x = [x_1, x_2, \dots]$  și  $y = [y_1, y_2, \dots]$  sunt aproape ortogonali.
- Motivul este că dacă proiecțiem  $x$  și  $y$  pe oricare dintre cele  $C_k$  plane formate de două dintre cele  $k$  axe va exista unu în care proiecțiile vectorilor sunt aproape ortogonale (probabilitatea sa existe crește cu numărul de dimensiuni)

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

20

### Alte distante

- Șirurile de caractere, cum sunt secvențele ADN, pot fi similare chiar și dacă există unele inserări și ștergeri precum și modificări ale unor caractere.
- De exemplu, *abcde* și *bcdxye* sunt destul de similară chiar dacă nu au nici o poziție comună și nu au nici chiar aceeași lungime.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

15

### Alte distante

- Astfel, în loc să construim un spațiu euclidian cu câte o dimensiune pentru fiecare poziție, putem defini funcția distanță:  
$$D(x, y) = |x| + |y| - 2|LCS(x, y)|$$
 unde LCS este cea mai lungă subsecvență comună lui  $x$  și  $y$ .
- În exemplul nostru *LCS(abcede, bcdxye)* este *bcd* de lungime 4, deci  $D(abcede, bcdxye) = 5 + 6 - 2 \times 4 = 3$ ; i.e., șirurile sunt destul de apropiate.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

16

### Abordări clustering

- La nivel final, putem împărti algoritmii de grupare în două mari clase:
  - Abordarea tip centroid: "ghicim" centroizii sau punctele centrale pentru fiecare cluster și asignăm punctele la clusterul având cel mai apropiat centroid.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

21

### Abordări clustering

- Abordarea ieherarhică:
  - Începem prin a considera că fiecare punct formează un cluster.
  - Comasăm repetat clusterurile apropiate de două cluste (e.g., distanța dintre centroizii lor), sau pentru căt de bun va exista unu în care proiecțiile vectorilor sunt aproape ortogonale (probabilitatea sa existe crește cu numărul de dimensiuni)

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

22

### Alte distante

- Aceasta funcție de distanță arată cate caractere trebuie sterse sau adăugate unuia dintre șiruri pentru a obține celalalt șir.
- Intr-adevar, pentru a obține de exemplu pe *abcede* din *bcdxye* trebuie să:
  - Adaugam un *a* în prima poziție
  - Ștergem *e* din a doua poziție
  - Ștergem *y* din a treia poziție

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

17

### Hiperdimensionalitatea

- O consecință mai puțin intuitivă a lucrului în spații hiperdimensionale este că aproape toate percherile de puncte sunt la o depărtare aproape egală cu media distanțelor între puncte.
- Exemplu:
  - Să presupunem că aruncăm aleator puncte într-un cub  $k$ -dimensional.
  - Pentru  $k=2$ , ne astăptăm ca punctele să fie răspândite în plan cu unele foarte apropiate între ele și unele foarte apropiate la distanță maximă posibilă.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

18

### Algoritmul k-means

- Acest algoritm este un algoritm popular care *ține datele în memoria centrală*
- Pentru acest algoritm se bazează pe altăl algoritm de clustering (ex.: BFR)
- k*-means alege aleator *k* centroizi de cluster dintre punctele care se grupează și asignează celelalte puncte la această alegând centrul cel mai apropiat de punctul respectiv.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

### Observație

- Dupa asignarea tuturor punctelor se recalculează centroizii clusterelor ca centrul de greutate al punctelor din fiecare cluster.
- El nu este de obicei unul din punctele care formează clusterul ci un punct între ele în spațiu euclidian respectiv.
- Procesul de asignare puncte – recalculare centroizi se repetă pana se îndeplinește o condiție de oprire.

F. Radulescu, Curs: Utilizarea bazelor de date, anul IV CS.

24

## Observatie

- Conceptul de centroid nu este valabil decat in cazul clusteringului in spatiu euclidian.
- Pentru cazurile in care nu avem un spatiu euclidian (punctele nu au coordonate numerice) exista o serie de variante ale algoritmului precum k-medoids sau k-modes.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

25

## Exemplu

- Un exemplu foarte simplu cu cinci puncte in două dimensiuni.
- Consideram pentru k valoarea 2 (k este un parametru de intrare in algoritm).
- Presupunem ca alegem punctele 1, 2 ca centroi initiali.
- Aaignam punctele 3, 4 si 5 la cel mai apropiat centroid.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

26

## Oprire k-means

- Criteriile de oprire - continuare:
- 4. Scădere SSD (suma pătratelor distanțelor) este sub un prag dat.
- SSD măsoară cat de compacte sunt clusterile (ca ansamblu).
- Este suma pătratelor distanțelor de la fiecare punct la centroidul său:

$$SSD = \sum_{i=1}^k \sum_{p \in Cluster_i} Dist(p, Centroid(Cluster_i))^2$$

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

31

## Aplicare k-means

- Dacă nu suntem siguri de valoarea lui k putem incerca valori diferite pentru k.
- Procesul se opreste când găsim cel mai mic astfel încât mărimea lui k nu măsoarează prea mult distanța medie a punctelor față de centroidul lor.
- Exemplul urmator ilustrează acest lucru.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

32

## Exemplu

- Rezultatul este urmatorul:



F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

27

## Exemplu

- Când considerăm punctul 3, presupunem ca acesta este mai apropiat de 1, deci 3 se adaugă clusterului conținând 1.
- Presupunem că atunci când asignăm 4 găsim că 4 este mai aproape de 2 decât de 1, deci 4 se alătură lui 2 în clusterul acestuia.
- În final, 5 este mai aproape de 1 decât de 2, deci el se adaugă la clusterul {1, 3}.
- Recalculăm acum centroizi și gasim noli centroizi C1 și C2.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

28

## Alt exemplu

- Să considerăm datele din figura urmatoare.
- În mod clar k=3 este numărul corect de clustere dar să presupunem că întâi încercăm k=1.
- În acest caz toate punctele sunt într-un singur cluster și distanța medie la centroid va fi mare.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

33

## Alt exemplu - cont

- Presupunem că apoi încercăm k=2.
- Unul dintre cele trei clustere va fi format iar celelalte două vor fi forțate să creze împreună un singur cluster, asa cum arată linia punctată.
- Distanța medie a punctelor la centroid de va măsura astfel considerabil.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

34

## Aplicare k-means

- Repetam acum operația și asignam toate cele 5 puncte la cea mai apropiată centroid (dintre C1 și C2).
- Obținem aceleasi clustere și aleiasem centroizi.
- Este evident că dacă vom continua nu obținem altceva. Procesul se încheie.
- Am obținut clusterele {1, 3, 5} și {2, 4}.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

29

## Oprire k-means

- Criteriile de oprire pot fi:
  1. Clusterele nu se schimbă de la o iteratie la alta (ca în exemplul anterior).
  2. Modificările clusterelor sunt sub un prag fixat (de exemplu, nu mai mult de  $\rho$  puncte schimbă clusterul între două iterări succesoare).
  3. Mișcarea centroizoilor este sub un prag dat (de exemplu, suma distanțelor dintre pozițiile vechi și cele noi pentru centroizi nu depășește între două iterări succesoive un prag fixat).

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

30

## Alt exemplu - cont

- Dacă luăm k=3 atunci fiecare dintre clusterele vizibile va forma un cluster iar distanța medie de la puncte la centroizi se va măsura din nou, aşa cum arată graficul din figura.
- Totuși, dacă mărim k la 4 unul dintre adevaratele clustere va fi particionat artificial în două clustere apropiate, aşa cum arată liniile continui.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

35

## Alt exemplu - cont

- Distanța media la centroid va scădea puțin dar nu mult.
- Acest eșec de a merge mai departe ne arată că valoarea k=3 este corectă chiar dacă datele sunt în atât de multe dimensiuni încât nu putem vizualiza clusterele.
- În acest fel aflăm valoarea corecta a lui k – numărul de clustere

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

36

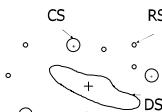
## Algoritmul BFR

- Bazat pe k-means, acest algoritm citește datele o singură dată în transe egale cu memoria centrală disponibilă la fiecare pas.
- Algoritmul lucrează cel mai bine dacă clusterele sunt normal distribuite în jurul unui punct central, eventual cu o deviație standard diferită în fiecare dimensiune.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

37

## Reprezentare clustere în BFR



Cei care au creat acest algoritm și au reprezentat clusterele ca pe niște galaxii.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

38

## Reprezentare - cont

- De notat că aceste trei tipuri de informație, totalizând în cazul în care avem  $k$  dimensiuni  $2k+1$  numere sunt suficiente pentru a calcula statistici importante pentru un cluster sau subcluster.
- Este mai convenabil de menținut pe măsură ce punctele sunt adăugate la cluster decât, să spunem, media și varianța în fiecare dimensiune,

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

43

## Reprezentare - cont

- Cordonata  $\mu_i$  a centroidului clusterului  $i$  în dimensiunea  $N$  este  $SUM_i/N$
- Varianța (dispersia) în dimensiunea  $N$  este:
 
$$\frac{SUMSQ_i}{N} - \left( \frac{SUM_i}{N} \right)^2$$
- iar deviația standard  $\sigma$  este rădăcina pătrată a acesteia.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

44

## Reprezentare clustere în BFR

- Un cluster constă din:
  1. Un nucleu central numit **Discard set - DS**.
  2. Multimea acestor puncte este considerată ca aparținând în mod sigur clusterului.
- Notă: deși numărul punctelor "de aruncat" acestora au de fapt un efect semnificativ pe parcursul executiei algoritmului de vreme ce determină colectiv unde este centrul și care este deviația standard a clusterului în fiecare dimensiune.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

39

## Reprezentare clustere în BFR

- 2. Galaxii înconjurătoare, numite colectiv **Compression set - CS (Multimea comprimată)**.
- Fiecare subcluster din CS constă într-un grup de puncte care sunt suficiente de apropiate unele de altelincă pot fi înlocuite cu statisticile lor, la fel ca și DS.
- Totuși, ele sunt suficiente de departe de orice centroid de cluster încât nu suntem încă siguri de care cluster aparțin.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

40

## Procesare

- La prima încărcare cu date a memoriei centrale, BFR selectează /comprimă/ punctele care vor fi utilizate în algoritm carească lucrările în memoria centrală, c.e., se ia un eşantion de puncte, se optimizează exact clusterul și se aleg centroizi lor ca centroizi inițiali.
- O memorie centrală de puncte este procesată la fel în toate încărcările cu date următoare după cum urmează:

1. Se determină care puncte sunt suficiente de apropiate de un centroid curent astfel încât pot fi luate în DS și statisticile lor ( $N, SUM, SUMSQ$ ) combinate cu statisticile anterioare ale clusterului.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

45

## Procesare

2. În memoria centrală se încearcă gruparea punctelor care nu au fost încă plasate în DS, inclusiv puncte ale RS din pașii precedenți.
- Dacă găsim un cluster de puncte a căror varianță este sub un prag ales, atunci vom privi aceste puncte ca un subcluster, le înlocuim cu statisticile lor și le considerăm parte a CS.
- Toate celelalte puncte vor fi plasate în RS.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

46

## Reprezentare clustere în BFR

- Stăle individuale care nu sunt parte a nici unei galaxii sau subgalaxii, **Multimea refuzată (Retained set - RS)**.
- Aceste puncte nici nu pot fi asignate vreunui cluster și nici grupate în vreun subcluster al CS.
- Ele sunt stocate în memoria centrală ca puncte individuale împreună cu statisticile pentru DS și RS.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

41

## Reprezentare comprimată

- Statisticile utilizate pentru a reprezenta fiecare cluster din DS și fiecare subcluster din CS sunt:
  1. Conținutul numărului de puncte  $N$ .
  2. Vectorul sumelor coordonatelor punctelor în fiecare dimensiune. Vectorul este notat cu  $SUM$  iar componenta pentru dimensiunea  $i$  cu  $SUM_i$ .
  3. Vectorul sumelor pătratelor coordonatelor punctelor în fiecare dimensiune notat cu  $SUMSQ$ . Componenta pentru dimensiunea  $i$  cu  $SUMSQ_i$ .

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

42

## Procesare

- Luăm în considerare unirea unui subcluster nou apărut cu un subcluster anterior din CS.
- Testul pentru a vedea dacă este de dorit să facem asta este că multimea combinată de puncte să albă o varianță sub un anumit prag.
- De notat că statisticile tijunte pentru subclusterurile din CS sunt suficiente pentru a calcula varianța mulțimii combinante.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

47

## Procesare

- Dacă este ultimul pas, i.e. nu mai sunt date, atunci vom asigna subclusterelor din CS și punctele din RS la cel mai apropiat cluster de ele chiar dacă ele vor fi destul de departe de orice centroid de cluster.
- În felul acesta obținem distrelere finale produse de algoritmul.

F. Radulescu, Curs: Utilizarea bazeelor de date, anul IV CS.

48

## Scalarea multidimensională

- În multe cazuri nu avem un spațiu euclidian ci doar o mulțime de puncte și distanța între oricare două dintre acestea.
- Exemplu: un graf în care cunoaștem lungimea fiecarui arc. Din aceste lungimi putem afla distanța între oricare două noduri ca fiind lungimea drumului minim între ele

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

49

## Scalarea multidimensională

- Se poate demonstra că având  $N$  puncte și distanțele între oricare 2 dintre ele putem crea un spațiu cu  $N-1$  dimensiuni
- În acest spațiu punctele sunt plasate exact (distanța calculată din coordonatele după plasare este aceeași cu distanța de la care s-a plecat pentru orice percheie de puncte)

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

50

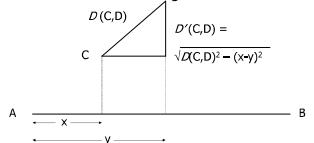
## Fastmap

1. Se aleg două puncte aflate la distanță cat mai mare,  $a$  și  $b$ . Acestea devin o axă de coordonate (cu originea în  $a$ ).
  2. Pentru orice punct  $c$  din cele  $N$  se calculează coordonata pe această axă conform formulei anterioare:
- $$x = (D^2(a, c) + D^2(a, b) - D^2(b, c)) / (2 * D(a, b))$$

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

55

## Fastmap



56

## Scalarea multidimensională

- Problema este că în cazul unui număr mare de puncte rezultă un număr mare de dimensiuni (spațiu hiperdimensional).
- Ideal ar fi să plasăm cat mai exact cele  $N$  puncte într-un spațiu având  $K$  dimensiuni unde  $K << N$ .
- Acum procesul se numește scalare multidimensională.
- Plasarea celor  $N$  puncte nu este 100% exactă (distanțele calculate din coordonatele rezultante nu sunt total exacte cu cele de la care s-a pornit)

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

51

## Scalarea multidimensională

- Formula de bază folosită este cea prin care având două puncte putem afla distanțele proiecției unui al treilea punct pe segmentul format de primele două puncte,
- Formula este obținută din teorema lui Pitagora generalizată (teorema cosinusului)

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

52

## Fastmap

3. Pentru următoarele axe se vor folosi nu distanțele initiale dintre puncte ci distanțe reportate în modul următor:
$$D'^2 = D^2 - (x - y)^2$$
4. Procesul se sfârșește după calculul numărului dorit de coordonate sau când nu mai pot fi alese noi axe de coordonate.

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

57

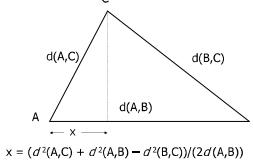
## Probleme în cazuri reale

- În cazurile reale (matricea nu este euclidiană) se poate întâmpla că patratul lui  $D'$  calculat cu formula anterioară să dea un număr negativ.
- În astfel de cazuri pentru a putea continua se poate lua  $D' = 0$ .
- Aceasta alegeră duce însă la erori care se propagă.
- Iată un exemplu de rulare pentru algoritm în cazul în care sunt considerate 2000 de noduri,

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

58

## Proiecția lui $C$ pe $AB$



53

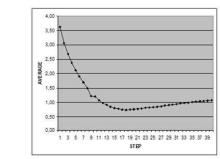
## Fastmap

- Algoritmul Fastmap este unul dintre algoritmii de scalare multidimensionale.
- Aceasta este un algoritm prin care se calculează succesiv coordonatele punctelor, ceea ce corespondă (o dimensiune) la fiecare pas.
- Pasul algoritmului este următorul:

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

54

## Probleme în cazuri reale



F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

59

## Probleme în cazuri reale

- Figura arată media diferențială între dreal și dcălculat unde:
- Dreala este distanța între puncte de la care s-a pornit (cunoscută prin ipoteza problemei)
- Dcălculat este distanța dintre puncte calculată pe baza coordonatelor obținute pana la pasul respectiv.
- Se observă că există un minim după 18 pași (spațiu optim are deci pentru acest exemplu 18 dimensiuni,  $k=18$ )

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

60

## Probleme în cazuri reale

- În mod normal graficul ar trebui să tindă asimptotic către 0.
- Faptul că nu se întâmplă asa e datorat erorilor induse de considerarea lui  $D' = 0$  în cazul în care patratul este negativ.
- Erorile acumulate fac ca după al 18-lea pas graficul să înceapă să crească.

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

61

## Bibliografie

- J.D.Ullman - CS345 --- Lecture Notes, Clustering I, II  
<http://info.stanford.edu/~ullman/cs345notes.html>

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

62

## Sfârșitul capitoului 10

F.Radulescu, Curs: Ullman's notes de date, anii IV-CS.

63

## Capitolul 12

### Data mining – căutare pe web

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

1

## Căutarea pe web

- ❑ Puncte importante :
  1. **Rangul paginii**, pentru descoperirea celor mai "importeante" pagini de web, utilizat de Google.
  2. **Indeci și autorități**, o evaluare mai detaliată a importanței paginilor web utilizând o varianta a calculului de valori proprii utilizată pentru rangul paginii.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

2

## Exemplul 1

- ❑ Fie  $[n, m, a]$  vectorul importanței pentru cele trei pagini : Netscape, Microsoft respectiv Amazon.
- ❑ Atunci ecuația care descrie valorile asymptotice ale acestor trei variabile este :

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

3

## Rangul paginii

- ❑ Intuitiv rezolvăm problema definiției "importeantei" recursiv : o pagină este importantă dacă pagini importante conțin legături către ea.
- ❑ Creăm o matrice stochastică a Internetului astfel :
  - > Fiecare pagină / corespunde liniei și coloanei / a matricii.
  - > Dacă pagina  $j$  are  $n$  succesiuni (legături), atunci elementul  $j, j$  al matricii este  $1/n$  dacă pagina / este unul dintre acești succesiuri ai paginii  $j$  și 0 altfel.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

4

## Rangul paginii

- ❑ Intuiția care stă în spatele acestor matrici este :
- ❑ Să ne imaginăm că inițial fiecare pagină are o unitate de importanță.
- ❑ La fiecare pas fiecare pagină își împarte importanța între succesișorii săi și primește noi fracțiuni de importanță de la predecesorii săi.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

5

## Exemplul 1

- ❑ Primele patru iterări dă următoarele estimări :

$$\begin{aligned} n &= \begin{bmatrix} 1 & 1 & 5/4 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} \\ m &= \begin{bmatrix} 1 & 1/2 & 3/4 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} \\ a &= \begin{bmatrix} 1 & 3/2 & 1 \\ 0 & 1/8 & 17/16 \\ 1/2 & 0 & 0 \end{bmatrix} \end{aligned}$$

- ❑ La limită, soluția este  $n = a = 6/5$ ;  $m = 3/5$ .

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

6

## Observații

- ❑ De notat că nu putem să obținem niciodată valorile absolute ale lui  $n$ ,  $m$  și  $a$  ci **doar raportul lor**, de vreme ce asertjunea inițială că fiecare a pornit de la 1 a fost arbitrară.
- ❑ Deoarece matricea este stochastică (suma pe fiecare coloană este 1), procesul de **relaxare** de mai sus converge către **vectorul principal de valori proprii** al matricii.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

10

## Rangul paginii

- ❑ După mai multe iterări, importanța fiecărui pagină atinge o limită care este compoziția corespunzătoare ei din vectorul principal de valori proprii al matricii.
- ❑ Această importanță este de asemenea probabilitatea ca un navigator pe web, pornind de la o pagină aleatorie și urmând legături aleator alese din fiecare pagină, să ajungă la pagina în discuție după o lungă serie de legături.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

5

## Exemplul 1

- ❑ În 1839 Internetul constă din doar trei pagini : Netscape, Amazon și Microsoft. Legăturile între aceste trei pagini erau ca în figura următoare:



F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

6

## Probleme cu grafuri reale

- ❑ 2 tipuri de probleme:
  1. **Dead end** : o pagină care nu are succesiuri nu are către cine să-și trimită importanță. În final totă importanța "se va scurge" din Internet.
  2. **Capcane** : un grup de una sau mai multe pagini care nu au legături către pagini din afara grupului vor acumula la final totă importanță din Internet.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

11

## Exemplul 2: Dead end

- ❑ Să presupunem că Microsoft încercă să profite că este un monopol înăsturând toate legăturile din situl său. Noul Internet este ca în figura următoare:



F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

12

## Exemplul 2

- ❑ Ecuatia matricială este în acest caz:

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 0 \end{bmatrix} * \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

- ❑ Se observă că suma pe coloane nu mai este intotdeauna 1 (coloane cu suma nula)

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

13

## Exemplul 2

- ❑ Primii patru pași ai soluției iterative sunt :

$$\begin{aligned} n &= \begin{bmatrix} 1 & 1 & 3/4 \\ 0 & 1 & 1/4 \\ 1/2 & 0 & 0 \end{bmatrix} \\ m &= \begin{bmatrix} 1 & 1/2 & 1/4 \\ 0 & 1 & 1/4 \\ 1/2 & 0 & 0 \end{bmatrix} \\ a &= \begin{bmatrix} 1 & 1/2 & 1/2 \\ 0 & 1/8 & 3/16 \\ 1/2 & 0 & 0 \end{bmatrix} \end{aligned}$$

- ❑ În acest caz, fiecare dintre  $n$ ,  $m$  și  $a$  ține catre 0; i.e. toată importanță se scurge afară.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

14

## Prevenire dead end și capcane

- ❑ Ecuatia cu taxare:

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = 0.8 \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix} + \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}$$

- ❑ Soluția acestei ecuații este  $n = 7/11$ ;  $m = 21/11$ ;  $a = 5/11$ .

- ❑ De notat că suma celor trei valori nu este 3 dar obținem o distribuție mult mai rezonabilă a importanței decât în Exemplul 3.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

19

## Spam

- ❑ "Spamming" este în acest context încercarea multor situri web de a părea că sunt despre un subiect care atrage vizitatorii fără că într-adevăr să fie deosebit de interesant.
- ❑ Google, ca și alte motoare de căutare, încercă să potrivescă cuvintele din cererile de căutare cu cuvinte din pagini web.

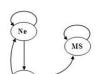
- ❑ Cu toate acestea, Google, spre deosebire de alte motoare de căutare, tinde să creată o pagină web facând mai greu pentru aceasta să pară că fiind despre ceva ce nu este.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

20

## Exemplul 3

- ❑ Microsoft decide să nu folosească decât legături către el însuși de acum încolo. Acum, Microsoft a devenit un capcană. Noul Internet este în figura următoare:



15

## Exemplul 3

- ❑ Ecuatia matricială este în acest caz:

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} * \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

- ❑ Suma pe coloane este 1 dar apar valori de 1 pe diagonala principala a matricii.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

16

## Spam

- ❑ Utilizarea rangului paginii pentru a măsura importanța în locul unei măsuri mult mai naivă ca "numărul de legături către acea pagină" protejează de asemenea împotriva spamului.

- ❑ Măsura naivă poate fi insăzată de un spammer care creează 1000 de pagini care se referă între ele în timp ce rangul paginii recunoaște că nici una dintre acestea nu are importanță reală.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

21

## Indeci și autorități

- ❑ Intuitiv, definim "index" și "autoritate" într-un mod mutual recursiv: un index conține legături către multe autorități iar o autoritate este referită de mulți indeci.

- ❑ Autoritățile pot fi pagini care oferă informații despre un subiect.

- ❑ Indecii sunt pagini care nu furnizează informații ci spun unde se găsesc informații.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

22

## Exemplul 3

- ❑ Primii pași ai algoritmului produc valori:

$$\begin{aligned} n &= \begin{bmatrix} 1 & 1 & 3/4 \\ 0 & 1 & 1/4 \\ 1/2 & 0 & 0 \end{bmatrix} \\ m &= \begin{bmatrix} 1 & 3/2 & 7/4 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} \\ a &= \begin{bmatrix} 1 & 1/2 & 1/2 \\ 0 & 3/8 & 5/16 \\ 1/2 & 0 & 0 \end{bmatrix} \end{aligned}$$

- ❑ Se observă acumularea pe linia m.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

17

## Prevenire dead end și capcane

- ❑ În loc de a aplica matricea direct, "taxăm" fiecare pagină cu o fracțiune din importanță sa curentă și distribuim importanța taxată în mod egal tuturor paginilor.
- ❑ Dacă folosim o taxă de 20% ecuația din exemplul 3 devine cea de pe transparentul urmator.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

18

## Indeci și autorități

- ❑ Utilizează o formalizare matricială similară cu cea de la rangul paginii dar fără restricția stochastică. Numărăm fiecare legătură ca 1, indiferent de cătă succesorii sau predecesorii are o pagină.
- ❑ Aplicarea repetată a matricii duce la divergență, dar putem introduce un factor de scalare pentru a ţine valourile calculate pentru gradul de "autoritate" sau de "indexare" pentru fiecare pagină între limite finite.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

23

## Indeci și autorități

- ❑ Definim matricea  $A$  ale cărei linii și coloane corespund paginilor web având elementul  $A_{ij} = 1$  dacă pagina  $i$  referă pagina  $j$  și 0 altfel.
- ❑ De notat că  $A^T$ , transpusa lui  $A$ , arată că matricea utilizată pentru calculul rangului paginilor din  $A^T$  are 1 acolo unde matricea pentru rang are fracții.

F.Rădulescu, Curs Utilizarea bazei de date, anul IV CS.

24

## Indecsi si autoritati

- Fie  $a$  si  $h$  doi vectori iar componenta lor corespunde gradului de autoritate respectiv indexare a paginii. Fie  $\lambda$  si  $\mu$  factorii de scalare corespunzători care vor fi calculați mai târziu. Atunci putem afirma că:
- $h = \lambda A a$ . Adică gradul de indexare al fiecărei pagini este suma gradelor de autoritate ale tuturor paginilor referente, scalată cu  $\lambda$ .
- $a = \mu A^T h$ . Adică gradul de autoritate al fiecărei pagini este suma gradelor de indexare ale tuturor paginilor care o referă, scalată cu  $\mu$ .

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

25

## Indecsi si autoritati

- Din ecuațiile (1) și (2) putem deduce folosind substituția, două ecuații care leagă vectorii  $a$  și  $h$  doar de ei însăși:
- $$a = \lambda A^T A a$$
- $$h = \lambda \mu A A^T h$$
- Ca urmare, putem calcula  $h$  și  $a$  prin relaxare, obținând vectorul principal de valori proprii al matricilor  $A A^T$  și respectiv  $A^T A$

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

26

## Extragerea de cunoștințe din web

- Puncte importante:
- 1. **Numărarea dinamică a mulțimilor de articole:** Căutarea de mulțimi interesante de articole într-un spațiu mult prea mare pentru a se putea lăua în considerare fiecare pereche de articole.
- 2. **"Carti și autori":** Intrigantul experiment al lui Sergey Brin de extragere de date relaționale din web.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

31

## Numarare dinamica

- Problema este de a găsi mulțimi de cuvinte care apar "neobișnuit de des" împreună pe web, e.g. "New" și "York" sau "Ducesa, de, York".
- "Neobișnuit de des" poate fi definit în diverse moduri pentru a încorpora ideea că numărul de documente web conținând mulțimea de cuvinte este mult mai mare decât cel așteptat în cazul în care cuvintele ar fi fost alese la întâmplare, fiecare cuvânt cu probabilitatea sa de apariție într-un document.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

32

## Exemplul 4

- Fie graful următor:



F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

27

## Exemplul 4

- Matricele sunt:

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \quad A^T = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad AA^T = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix}, \quad A^T A = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 2 \end{bmatrix}$$

- Dacă utilizăm  $\lambda = \mu = 1$  și considerăm că vectorii  $h$  =  $[h_n, h_m, h_a]$  și  $a$  =  $[a_n, a_m, a_a]$  sunt inițial fiecare  $[1, 1, 1]$ , primele trei iterări ale ecuațiilor pentru  $a$  și  $h$  sunt:

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

28

## Numarare dinamica

- Un mod adecvat este **entropa per cuvânt din mulțime**. Formal, **interesul** unei mulțimi de cuvinte  $S$  este:

$$\log \left( \frac{\prod_{w \in S} prob(w)}{|S|} \right)$$

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

33

## Numarare dinamica

- De notat că împărțirea dimensiunii lui  $S$  pentru a evita "efectul Bonferroni", în care sunt atât de multe mulțimi de o dimensiune dată încât unele, din motive probabilistice, par a fi corelate,

- Exemplu: Daca  $a$ ,  $b$  și  $c$  (cuvinte) apar fiecare în 1% din toate documentele și  $S = \{a, b, c\}$  apar în 0,1% din documente, interesul lui  $S$  este  $(\log(0,001/0,01 \times 0,01 \times 0,01))/3 = \log(2)(1000/3)$  adică aproximativ 3,3.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

34

## Exemplul 4

- Secesiunea de valori pentru a este:

|       |   |   |   |    |     |
|-------|---|---|---|----|-----|
| $a_n$ | = | 1 | 5 | 24 | 114 |
| $a_m$ | = | 1 | 5 | 24 | 114 |
| $a_a$ | = | 1 | 4 | 18 | 84  |

- iar pentru  $h$ :

|       |   |   |   |    |     |
|-------|---|---|---|----|-----|
| $h_n$ | = | 1 | 6 | 28 | 132 |
| $h_m$ | = | 1 | 2 | 8  | 36  |
| $h_a$ | = | 1 | 4 | 20 | 96  |

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

29

## Exemplul 4

- Vectorul  $a$ , **scădat corespunzător**, va converge către un vector în care:

➤  $a_n = a_m$  și

➤ fiecare dintre aceste numere este mai mare ca  $a_a$  în raportul  $1 + \sqrt{3}/2$  sau aproximativ 1,36

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

30

## Numarare dinamica

- Problema tehnică: interesul nu este monoton sau "închis în jos" în modul de la produse cu larg suport.

- Astăzintă putem avea mulțimi  $S$  cu o valoare mare a interesului și totuși unele sau chiar toate submulțimile sale stricte să nu fie interesante.

- Prin contrast, dacă  $S$  are suport larg, atunci toate submulțimile sale au cel puțin același suport.

- Observație: Cu mai mult de  $10^6$  cuvinte diferențiate apărând pe web nu este posibil nici măcar să considerăm toate perechile de cuvinte,

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

35

## DICE

- DICE (dynamic itemset counting engine) vizitează repetat paginile web într-un mod de tip "round-robin" ("zborul prigorului/prigoriei").



36

## DICE

- De fiecare dată numărării aparițiile anumitor mulțimi de cuvinte și ale fiecărui cuvânt din aceste mulțimi.
- Numarul de mulțimi numărate este suficient de mic încât contorii lor încap în memoria centrală.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

37

## DICE

- Din cînd în cînd, să spunem la fiecare 5000 de pagini, DICE își reconsideră mulțimile pentru care numără. Înălță acele mulțimi care au cel mai mic interes și le înlocuiește cu alte mulțimi.

- Alegera noilor mulțimi se bazează pe proprietatea numită **heavy edge** care este o observație justificată experimental că acele cuvinte care apar în mulțimi cu interes ridicat au probabilitatea mai mare să apară în alte mulțimi cu interes ridicat.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

38

## Cum lucreaza

1. Se pornește de la un eşantion al tuplurilor (liniilor) care se doresc găsite.

2. În exemplul discutat în lucrarea lui Brin au fost folosite cinci exemple de perechi cu titluri de cărți și autori acestora.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

43

## Cum lucreaza

2. Pe baza exemplelor cunoscute, se caută pagini unde apar aceste date pe web.

3. Dacă se găsește un şablon care identifică un număr de tupluri cunoscute și este suficient de specific încât și puțin probabil să identifice prea mult, atunci se acceptă acest şablon.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

44

## DICE

- Astfel, când selectează noi mulțimi pentru a începe numărarea, DICE este direcțional în favoarea cuvintelor care apar deja în mulțimi cu interes ridicat.
- Totuși, nu se bazează pe aceste cuvinte altfel nu ar putea niciodată să găsească mulțimi cu interes ridicat compuse din multele cuvinte pe care nu le-a considerat niciodată.
- Unele (dar nu toate) din construcțiile pe care le utilizează DICE pentru crearea noilor mulțimi sunt:

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

39

## DICE: noi mulțimi

1. Două cuvinte aleatorii. Aceasta este singura regulă independentă de assertiunea "heavy edge" și ajuta noi cuvinte să ajungă în mulțime.
2. Un cuvânt dintr-o mulțime interesantă și un cuvânt aleator.
3. Două cuvinte din două mulțimi interesante diferite.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

40

## Cum lucreaza

3. Fiind dată o mulțime de şabloane acceptate, se caută date care satisfac aceste şabloane și se adaugă la mulțimea datelor cunoscute.

4. Se repetă pașii (2) și (3) de un număr de ori. În exemplul citat au fost utilizate patru citări care au dus la 15,000 de tupluri; aprox. 95% au fost perechi adevarate titlu-autor.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

45

## Ce este un şablon

- Are 5 componente:
  1. **Ordinea**, i.e., dacă titlul apare în text înaintea autorului sau vice-versa. Într-un caz general, în care tuplurile au mai mult de 2 componente, ordinea va fi dată de permutea componentelor.
  2. **Prefixul adresei web (URL)**.
  3. **Prefixul/textul**, care apare înaintea primului dintre titlu și autor

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

46

## Exemplu

- Un şablon posibil poate consta din următoarele:
    1. **Ordinea**: titlu și apoi autor.
    2. **Prefixul URL**: www.stanford.edu/dass
    3. **Prefixul/textul**, care urmează după al doilea dintr-o serie de date. Atât prefixul căt și suffixul au fost limitate la 10 caractere.
- Aici `<L>><titlu></>` de **autor** `</>`
- Titlul este orice apără între `prefix` și  `mijloc` ; autorul este orice apără între `mijloc` și `suffix`.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

47

## DICE: noi mulțimi

4. Reuniunea a două mulțimi interesante a căror intersecție are dimensiunea 2 sau mai mult.
5.  $\{a, b, c\}$  dacă toate mulțimiile  $\{a, b\}$ ,  $\{a, c\}$  și  $\{b, c\}$  sunt găsite ca fiind interesante.
- Bineînțeles, în general sunt mult prea multe opțiuni de a aplica cele de mai sus în toate modulele posibile astfel încât se utilizează o selecție aleatoare a opțiunilor dând o anumită sensă fiecăreia dintre ele.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

41

## Carti si autori

- Ideea principală este de a căuta pe web faptele de un anumit tip, de genul celor care ar putea forma o relație de genul **Cartă(titlu, autor)**. Procesarea este sugerată de figura următoare:



F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

42

## Ce este un sablon

- Are 5 componente:
  1. **Mijlocul** text care apare între cele două elemente de date,
  2. **Suffixul**/textul care urmează după al doilea dintr-o serie de date. Atât prefixul căt și suffixul au fost limitate la 10 caractere.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

47

## DICE: noi mulțimi

4. Reuniunea a două mulțimi interesante a căror intersecție are dimensiunea 2 sau mai mult.
5.  $\{a, b, c\}$  dacă toate mulțimiile  $\{a, b\}$ ,  $\{a, c\}$  și  $\{b, c\}$  sunt găsite ca fiind interesante.
- Bineînțeles, în general sunt mult prea multe opțiuni de a aplica cele de mai sus în toate modulele posibile astfel încât se utilizează o selecție aleatoare a opțiunilor dând o anumită sensă fiecăreia dintre ele.

F. Radulescu, Curs Utilizarea webului de date, anul IV CS.

43

## Tuning sablon

- Definim **specificitatea** unui şablon ca fiind produsul lungimilor prefixului, mijlocului, sufixului și prefixului URL.
- În mare, specificitatea măsoară cât de posibil este să găsim date care corespund şablonului; cu cât specificitatea este mai mare, cu atât ne aşteptăm la mai puține aparitii ale acestuia în date.

F.Rădulescu, Curs Ulliman, baza de date, anul IV CS

49

## Tuning sablon

- Şablonul trebuie să îndeplinească două condiții pentru a fi acceptat:
  1. Trebuie să fie cel puțin 2 elemente de date cunoscute care apar conform aceluia şablon.
  2. Produsul specificității şablonului cu numărul de aparitii de date conform acestuia trebuie să depășească un anumit prag T (nespecificat în articolul original).

F.Rădulescu, Curs Ulliman, baza de date, anul IV CS

50

## Constructia sabloanelor

- Prefixul comun este [www.stanford.edu/class/cs](http://www.stanford.edu/class/cs)
- Dacă trebuie să spargem grupul, atunci următorul caracter, 3 sau 1, sparge grupul în două, cu acele aparitii ale datelor din prima pagină (pot fi multe astfel de aparitii) mergând într-un prim grup și aparitiile din celelalte două pagini în ceeaலălt:
  - >[www.stanford.edu/class/cs345/index.html](http://www.stanford.edu/class/cs345/index.html)
  - >[www.stanford.edu/class/cs340/readings.html](http://www.stanford.edu/class/cs340/readings.html)
- Să...
  - >[www.stanford.edu/class/cs145/index.html](http://www.stanford.edu/class/cs145/index.html)

F.Rădulescu, Curs Ulliman, baza de date, anul IV CS

55

## Gasirea aparitiilor din sabloane

1. Se găsesc toate URL-urile care se potrivesc cu prefixul URL al cel puțin unui şablon.
2. Pentru fiecare astfel de pagină se parcurge textul folosind o expresie regulată construită din prefixul, mijlocul și sufixul şablonului.
3. Se extrage din fiecare potrivire titlul și autorul, după ordinea specificată în şablon.

F.Rădulescu, Curs Ulliman, baza de date, anul IV CS

56

## Pasii executiei

1. Găsirea aparitiilor pornind de la datele cunoscute
2. Construcția sabloanelor din aparitiile de date
3. Găsirea aparitiilor pornind de la sabloane

F.Rădulescu, Curs Ulliman, baza de date, anul IV CS

51

## Aparitie

- O aparitie a unui tuplu (existent in tabela) constă în:
  1. Un anumit titlu și autor.
  2. Adresa Internet completa (URL) și nu doar prefixul ca în cazul şablonului,
  3. Ordinea, prefixul, mijlocul și sufixul şablonului după care au apărut titlul și autorul respectiv.

F.Rădulescu, Curs Ulliman, baza de date, anul IV CS

52

## Bibliografie

- J.D.Ullman - CS345 --- Lecture Notes: PageRank, Hubs-and-Authorities , Web Mining <http://infolab.stanford.edu/~ullman/cs345-notes.html>

F.Rădulescu, Curs Ulliman, baza de date, anul IV CS

57

## Sfârșitul capitolului 12

F.Rădulescu, Curs Ulliman, baza de date, anul IV CS

58

## Constructia sabloanelor

1. Se grupează aparitiile de date după ordinea și mijlocul lor. De exemplu, un grup din acest "group-by" poate corespunde ordinii "titlu-apoi-autor" și mijlocului "</>" de ".
2. Pentru fiecare grup se găsește cel mai lung prefix, sufix și prefix URL comun.
3. Dacă testul de specificitate pentru acest şablon este îndeplinit, se acceptă şablonul.

F.Rădulescu, Curs Ulliman, baza de date, anul IV CS

53

## Constructia sabloanelor

- Dacă testul de specificitate nu este îndeplinit, se încearcă spargerea grupului în două prin extinderea lungimii prefixului URL cu un caracter și apoi se repetă pasul (2). Dacă este imposibil să spargem grupul (pentru că există doar un URL) atunci am eşuat în a produce un nou şablon din acel grup.
- Exemplu:** Să presupunem că grupul conține trei URL-uri:
  - >[www.stanford.edu/class/cs345/index.html](http://www.stanford.edu/class/cs345/index.html)
  - >[www.stanford.edu/class/cs145/index.html](http://www.stanford.edu/class/cs145/index.html)
  - >[www.stanford.edu/class/cs340/readings.html](http://www.stanford.edu/class/cs340/readings.html)

F.Rădulescu, Curs Ulliman, baza de date, anul IV CS

54

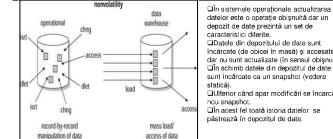
## Depozite de date și Modelarea dimensională

### Câteva definiții

▪ Wikipedia:

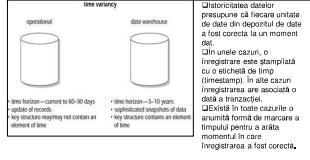
- Un depozit de date este locul de stocare al datelor electronice ale unei organizații. Depozitele de date sunt concepute pentru a facilita raportarea și analiza (din carteau lui Inmon spun că [1]).
- Un depozit de date găzduiește o formă standardizată, consistentă, curată și integrată a datelor proveniente din diverse sisteme operaționale utilizate în aceea organizație, structurată pentru a aborda în mod specific cerințele de raportare și de analiză.

### Colecție nevolatilă



2

### Date istorice (time variant)



8

### Câteva definiții

▪ R. Kimball (vezi [2, 3]):

- Un depozit de date este o copie a datelor tranzacționale structurate special pentru interogare și analiză.
- Conform acestei definiții:
    - Forma datelor stocate (SGDBR, fișier) nu este legată de definiția unui depozit de date.
    - Depozitarea datelor nu este legată exclusiv de „factorii de decizie” sau utilizată doar în procesul de luare a decizilor.

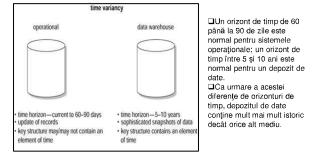
### Câteva definiții

▪ Inmon (vezi [1, 3]):

- Un depozit de date este o colecție de date pentru asistență decizilor factorilor de conducere care este:
  - orientată pe subiecte,
  - integrată,
  - nevolatilă,
  - istorică (time variant)

4

### Date istorice (time variant)

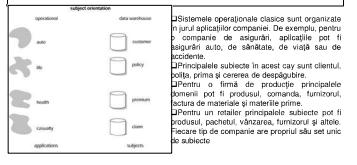


### Date normalize vs. Modelare dimensională

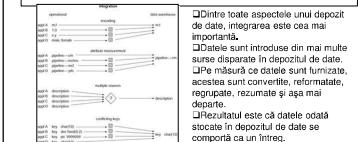
- Există două abordări principale pentru stocarea datelor într-un depozit de date:
  - 1. Abordarea normalizată (Inmon - [1])
  - 2. Abordarea dimensională (Kimball - [2])
- Aceste abordări nu se exclud reciproc și există și alte abordări. Abordările dimensionale pot implica normalizarea datelor într-un anumit grad.
- Cursul de azi se bazează pe abordarea dimensională și cartea lui Kimball & Ross - vezi [2].

10

### Colecție orientată pe subiecte



### Colecție integrată



6

### Date normalize vs. Modelare dimensională

- În abordarea normalizată, datele din depozitul de date sunt stocate urmărind, într-o anumită măsură, regulile cunoșute privind normalizarea din domeniul bazelor de date.
- Tabelele sunt grupate pe subiecte care reflectă categoriile generale de date (de exemplu date despre client, producător, finanțe etc.).
- Principialul avantaj al acestei abordări este că este simplu să adaugi noi informații în baza de date.

1

### Date normalize vs. Modelare dimensională

- Un dezavantaj al acestei abordări este faptul că, din cauza numărului mare de tabele implicate, poate fi dificil pentru utilizatorii să reușească prin join date din diferite surse pentru a obține informații semnificative și să acceseze informații fără o înțelegere precisă a surselor de date și a structurii depozitului de date.

12

### Date normalize vs. Modelare dimensională

- În abordare dimensională, datele tranzacționale sunt împărțite în fapte sau măsuri (date numerice ale tranzacțiilor) și dimensiuni (informații referitoare care conferă contextul faptelor).
- De exemplu, o tranzacție de vânzare poate fi defășurată în lăptea preluării unei produse, comanda și prețul plătit pentru produse și în dimensiuni preluare, data comenzi, numele clientului, codul produsului, locația expedierii, locația de facturare și angajul de vânzări responsabil de primirea comenzii.

### Date normalize vs. Modelare dimensională

- Un avantaj esențial al unei abordări dimensionale este că depozitul de date este mai ușor de înțeles și de folosit. De asemenea regăsirea datelor în depozitul de date poate să funcționeze foarte rapidă.
- Principalele dezavantaje ale abordării dimensionale sunt:
  1. Perțină la integrarea tabelelor de fapte și dimensiuni într-o singură depozit de date cu date din diferite sisteme operaționale este mai complicată și
  2. Este mai dificil de modificat structura depozitului de date dacă organizația care adoptă abordarea dimensională schimbă modul în care își desfășoară activitatea.

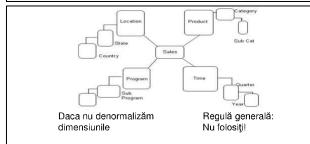
14

### Scheme stea



9

### Schema Fulg de nea (snowflake)



20

### Modelare dimensională

- Este o modalitate de structurare pentru un depozit de date
- Bazat pe:
  - Tabele de fapte (sau măsuri)
  - Tabele de dimensiuni

### Tabelă de fapte (măsuri)

- Reprezintă un proces de business
- Contine măsurătorile sau valourile sau faptele proceselor de business
- De exemplu „cantitate” și „preț total” în cazul unei comenzi de achiziție
- Majoritatea atributelor (coloanele) sunt additive (vânzări în această lună), unele sunt semi-additive (sold la data de ieșire), atele nu sunt additive (cod produs)
- Nivelul de detail se numește „granularitatea” tabelului – ce reprezintă o linie din tabela de fapte respectivă.
- Conține chei strânsă pentru fiecare tabelă de dimensiuni cu care se relatează.

16

### Cei patru pași ai modelării dimensionale

- Obiectiv: proiectarea unei baze de date dimensionale, luând în considerare patru pași în această ordine:
1. Selectarea procesul de business de modelat.
  2. Declarația gradului de detaliere (declare the grain).
  3. Stabilirea dimensiunilor pentru tabelele de fapte.
  4. Identificarea atributelor tabelelor de fapte.

(Referință: Ralph Kimball, Margy Ross - The Data Warehouse Toolkit, Second Edition, Wiley & Sons, 2002, pp.26-25)

1

### 1. Selectare proces de business

- Un proces este o activitate naturală de business desfășurată într-o organizație
- De obicei este susținut de un sistem de colectare a datelor.
- Exemple de procese de afaceri includ:
  - Aprovisionare,
  - Comenzi,
  - Livrări,
  - Facturare,
  - Inventar (stocuri),
  - Contabilitate

22

### Tabelele de dimensiuni

- Reprezintă cine, ce, unde, când și cum pentru o măsură
- Reprezintă entități din lumea reală și nu procese de business
- Dau contextul unei măsuri (subject)
- De exemplu pentru tabelul de fapte Vanzări, caracteristicile măsurii „cantitate” pot fi locație (unde), timp (când), produs (ce).
- Atributele dimensiunilor sunt coloanele din respectivul tabel.

### Tabelele de dimensiuni

- În dimensiunea Locație, atributele pot fi codul locației, județul, țara, codul poștal.
- În general, atributele dimensiunilor sunt utilizate în etichetele raportărilor și în condiții cum ar fi Tara = "USA".
- Înainte de a crea o depozit de date trebuie decis ce conține acesta. Să zicem că se doresc un depozit de date care să conțină vânzările pe diverse locații ale magazinelor, pe timp (data) și pe produse. Atunci dimensiunile vor fi: Locație, Dată și Produs.

18

### 2. Declarația gradului de detaliere

- Declarația gradului de detaliere înseamnă specificarea exactă a ceea ce reprezintă o linie din tabelul de fapte.
- Aceasta oferă răspunsul la întrebarea „Cum descrieți o singură linie din tabelul de fapte?”

3

### 2. Declarația gradului de detaliere

- Exemple - ce este o linie din tabelul de fapte:
- O linie de pe bonul de casă al unui client, generat de POS
  - O linie de pe factură primită de la o firmă
  - Un permis de îmbărcare pentru a urca într-un zbor
  - Valoarea zilnică a stocului pentru fiecare produs dintr-un depozit
  - Soldul la sfârșit de lună pentru un cont bancar

24



### Atribute dimensiuni: Produs

- Dimensiunea Produs descrie fiecare SKU din magazinul alimentar.
  - Un magazin obisnuit din lanț poate stoca 60.000 de SKU-uri, dar atunci când avem în vedere diferite scheme de merchandising din lanț și produse istorice care nu mai sunt disponibile, dimensiunea produsului ar avea cel puțin 150.000 de linii sau mai mult.

F. Fabiano - UCB - Novembre

### Atribute dimensiuni: Produs

- o Product Key (PK)
  - o Product Description
  - o SKU Number (Natural Key)
  - o Brand Description
  - o Category Description
  - o Department Description
  - o Package Type Description
  - o Package Size
  - o Fat Content
  - o Diet Type
  - o Weight
  - o Weight Units of Measure
  - o Storage Type
  - o Shelf Life Type
  - o Shelf Width
  - o Shelf Height
  - o Shelf Depth

50

### Atribute dimensiuni: Magazin

- Store Name
  - Store Number (Natural Key)
  - Store Street Address
  - Store City
  - Store County
  - Store State
  - Store Zip Code
  - Store Manager
  - Store District
  - Store Region
  - Floor Plan Type
  - Photo Processing Type
  - Financial Service Type
  - Selling Square Footage
  - Total Square Footage
  - First Open Date
  - Last Remodel Date

1

### Atribute dimensiuni: Promoție

- Dimensiunea **Promotie** descrie condițiile de promovare în care a fost vândut un produs.
  - Condițiile de promovare includ reducere temporare de preț, plasare preferențială, anunțuri în ziare și cupoane.
  - Această dimensiune este adesea numită dimensiune cauzală - causal dimension (spre deosebire de o dimensiune obiectivă - *casus dimension*), deoarece descrie factorii gândiți să provoace o schimbare a cifrelor de vânzări.

56

### Atribute dimensiuni: Produs

1

### Atribute dimensiuni: Produs

- ❑ Un tabel rezonabil de tip **Produs** are 50 sau mai multe atribute descriptive.
  - ❑ Fiecare atribut este o sursă bogată pentru construirea rapoartelor.

52

### Atribute dimensiuni: Promoție

- Promotion Key (PK)
  - Promotion Name
  - Price Reduction Type
  - Promotion Media Type
  - Ad Type
  - Display Type
  - Coupon Type
  - Ad Media Name
  - Display Provider
  - Promotion Cost
  - Promotion Begin Date
  - Promotion End Date .....

1

### **Număr tranzacție: Dimensiuni degenerate**

- Numerele de control operațional, cum ar fi numerele de ordine, numerele facturii și numerele de facturare, de obicei, dă naștere la dimensiuni goale și sunt reprezentate ca dimensiuni degenerate (adică avem chei strânsă în tabela de fapte fără tabele de dimensiune corespunzătoare).

58

## Atribute dimensiuni: Magazin

- Dimensiunea **Magazin** descrie fiecare magazin din lantul nostru.
  - Dimensiunea **Magazin** este dimensiunea geografică principală în studiu nostru de cauză.
  - Fiecare magazin poate fi găsită căruia o locație. Din această cauză, putem asocia un magazin cu orice atribut geografic, cum ar fi codul poștal, județul și statul din Statele Unite.
  - De obicei, magazinele se plasează în districte și regiuni.

1

#### Atribute dimensiuni: Magazin

- ❑ De obicei, magazinele se plasează în districte și regiuni.
  - ❑ Aceste două ierarhii diferențe sunt ambele reprezentate cu urșinjă în dimensiunea magazinului, deoarece atât ierarhile geografice cât și cele regionale ale magazinului sunt bine definite pentru o singură linie din dimensiunea Magazin.
  - ❑ Nu este neobișnuit să reprezintăm mai multe ierarhii într-un tabel cu dimensiuni. În mod ideal, numele și valorile atributului ar trebui să fie unice pentru ierarhile multiple.

1

## Bibliografie

1. W.H. Inmon - Building The Data Warehouse, Third Edition, Wiley & Sons, 2002
  2. Ralph Kimball, Margy Ross - The Data Warehouse Toolkit, Second Edition, Wiley & Sons, 2002
  3. Dimitra Vlitas, CS 680 Course notes
  4. Wikipedia – Pages on Data Warehouse, etc.
  5. <http://searchsqlserver.techtarget.com>
  6. Vincent Raineri, Building a Data Warehouse with Examples in SQL Server, Springer, 2008

1

|                                                                                                                                                                                     |                                                                                                                                                                                                                            |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Accesul userilor comuni la o baza de date Oracle se poate face daca instanta este in starea:                                                                                        | OPEN                                                                                                                                                                                                                       |
| Adaugarea de noi membri intr-un grup de Redo Log se face cu comanda:                                                                                                                | ALTER DATABASE ADD LOGFILE MEMBER                                                                                                                                                                                          |
| Adaugarea unei constrangeri de integritate pe o tabela se face cu o comanda Oracle de tipul:                                                                                        | ALTER TABLE table_name ADD CONSTRAINT constraint_name ...                                                                                                                                                                  |
| Algoritmi de clasificare sunt de tipul:                                                                                                                                             | Predictiv                                                                                                                                                                                                                  |
| Algoritmi de clustering sunt de tipul:                                                                                                                                              | Descriptiv                                                                                                                                                                                                                 |
| Alocarea SGA se face la:                                                                                                                                                            | Pornirea instantei                                                                                                                                                                                                         |
| Binning": daca avem o cutie ("bin") continand valorile 4, 8, 9, 15, prin netezirea folosind capetele intervalului obtinem:                                                          | 4, 4, 4, 15                                                                                                                                                                                                                |
| Blocurile de date sunt scrise pe disc in memorie de:                                                                                                                                | Procesul DBWR                                                                                                                                                                                                              |
| Bonurile de casa stocate pe hard discul serverului de magazin sunt:                                                                                                                 |                                                                                                                                                                                                                            |
| Buffer cache este parte a:                                                                                                                                                          | SGA                                                                                                                                                                                                                        |
| Ca dată despre o persoană, genul (masculin, feminin) este un atribut de tip:                                                                                                        | Binar simetric                                                                                                                                                                                                             |
| Ca efect al comenzii ALTER SYSTEM KILL SESSION tranzactia curenta:                                                                                                                  | Este anulata                                                                                                                                                                                                               |
| Cand cota unui user este modificata la valoarea 0 (aceasta fiind o valoare mai mica decat spatiul ocupat la acel moment de acel user in acel tablespace) atunci in acel tablespace: | Nu se mai pot crea noi obiecte                                                                                                                                                                                             |
| Cand cota unui user intr-un tablespace e trecuta pe o valoare mai mica decat spatiul ocupat:                                                                                        | Nu se mai pot crea obiecte noi iar cele existente nu pot creste                                                                                                                                                            |
| Când începem cu un set de articole etichetate (etichete dintr-un set de clase) avem:                                                                                                | Învățarea supervizată                                                                                                                                                                                                      |
| Cand o tranzactie este comisa (COMMIT) se scriu pe disc:                                                                                                                            | Inregistrari din Redo Log                                                                                                                                                                                                  |
| Cand se creeaza un user Oracle trebuie specificat obligatoriu in comanda de creare :                                                                                                | Parola de conexiune la baza de date                                                                                                                                                                                        |
| Care afirmatie este adevarata:                                                                                                                                                      | Putem alege schema fulg-de-nea din schema star prin denormalizare                                                                                                                                                          |
| Care afirmatie este adevarata:                                                                                                                                                      | O tabela de facts poate fi asociata la mai multe dimensiuni                                                                                                                                                                |
| Care afirmatie este adevarata:                                                                                                                                                      | Un segment e format din extensii                                                                                                                                                                                           |
| Care afirmatie este adevarata:                                                                                                                                                      | a si b sunt false (a. un fisier apartine unei singure extensii, b. un tablespace e stocat intr-un singur fisier)                                                                                                           |
| Care afirmatie este adevarata:                                                                                                                                                      | Clausa FORCE LOGGING din CREATE TABLESPACE are prioritate fata de un NOLOGGING de la crearea unei tabele in acel                                                                                                           |
| Care afirmatie este adevarata:                                                                                                                                                      | Clausa LOGGING din CREATE TABLE are prioritate fata de un NOLOGGING al tablespace-ului in care se pune tabela.                                                                                                             |
| Care afirmatie este adevarata:                                                                                                                                                      | Un segment se poate intinde pe mai multe fisiere de date                                                                                                                                                                   |
| Care afirmatie este adevarata:                                                                                                                                                      | Tabelele de fapte si cele de dimensiuni contin cate o cheie strana pentru fiecare tabelă de celalalt tip de care sunt legate intr-o schema stea ?                                                                          |
| Care afirmatie este corecta :                                                                                                                                                       | Cand apare un log-switch se face si un checkpoint                                                                                                                                                                          |
| Care afirmatie este corecta:                                                                                                                                                        | Există un tablespace numit SYSAUX                                                                                                                                                                                          |
| Care afirmatie este corecta:                                                                                                                                                        | Există un tablespace numit SYSTEM                                                                                                                                                                                          |
| Care afirmatie este falsa privitor la tablespace-ul SYSAUX:                                                                                                                         | Nu necesita privilegii sporite pentru operare                                                                                                                                                                              |
| Care afirmatie este falsa privitor la tablespace-ul SYSTEM:                                                                                                                         | Nu necesita privilegii sporite pentru operare                                                                                                                                                                              |
| Care afirmatie este falsa:                                                                                                                                                          | O extensie se poate intinde pe mai multe fisiere de date ??? (Un segment se poate intinde pe mai multe fisiere de date) pare corect                                                                                        |
| Care comanda este corecta pentru vizualizarea tablespace-urilor create intr-o baza de date Oracle:                                                                                  | SELECT tablespace_name FROM dba tablespaces;                                                                                                                                                                               |
| Care comanda se poate folosi pentru a muta un index intr-un alt tablespace decat cel in care a fost creat:                                                                          | ALTER INDEX index_name REBUILD new_tablespace_name;                                                                                                                                                                        |
| Care dintre afirmatiile urmatoare sunt adevarate pentru un tablespace in starea READ ONLY dintr-o baza de date Oracle:                                                              | Ambele afirmatii prezente la a si b sunt adevarate (Sunt permise numai operatii de citire din tablele stocate in acel tablespace / Sunt permise operatii de stergere din dictiunari a tablelor stocate in acel tablespace) |
| Care dintre urmatoarele optiuni permit unui user Oracle care primeste un privilegiu de modificarile de date pe o tabela, sa-l transmita unui alt user:                              | Ambele variante specificate la a si b creeaza permisiunea (GRANT ANY OBJECT PRIVILEGE / GRANT ... WITH ADMIN OPTION)                                                                                                       |
| Care dintre urmatoarele privilegii Oracle fac parte din categoria privilegiilor SYSDBA:                                                                                             | CREATE DATABASE si RESTRICTED SESSION                                                                                                                                                                                      |
| Care dintre urmatoarele privilegii pe obiecte se pot acorda unui user Oracle:                                                                                                       | Toate privilegiile specificate la a si b (Modificare structura tabela si executie procedura stocata / Interrogare pe un view si o secenta)                                                                                 |
| Care element este o succesiune contigua de blocuri?                                                                                                                                 | Extensie                                                                                                                                                                                                                   |
| Care este afirmatia corecta :                                                                                                                                                       | DBA=rol, SYSDBA=privilegiu                                                                                                                                                                                                 |
| Care este ordinea corecta a etapelor de pornire BD:                                                                                                                                 | Pornire instanta, Montare, Deschidere                                                                                                                                                                                      |
| Care mod de oprire necesita apoi operatia de recuperare:                                                                                                                            | Abort                                                                                                                                                                                                                      |
| Care operatie nu poate fi efectuata cu SYSOPER doar:                                                                                                                                | CREATE DATABASE                                                                                                                                                                                                            |
| Care zona de memorie este folosita circular:                                                                                                                                        | Redo Log Buffer                                                                                                                                                                                                            |
| Cate coloane are vederea VSSGA?                                                                                                                                                     | 2                                                                                                                                                                                                                          |
| Cate zone de tip SGA si PGA pot fi alocate simultan in memorie pentru o instanta Oracle:                                                                                            | O zona SGA si potential mai multe zone PGA ?                                                                                                                                                                               |
| Ce poate fi stocat in mai mult de un fisier de date:                                                                                                                                | Segment si Tablespace                                                                                                                                                                                                      |
| Cererea GRANT ... ON ... adauga privilegii:                                                                                                                                         | Obiect                                                                                                                                                                                                                     |
| Cererea GRANT ... TO ... adauga privilegii:                                                                                                                                         | Sistem                                                                                                                                                                                                                     |
| Clausa AUTOEXTEND permite cresterea dimensiunii :                                                                                                                                   | Fisierelor de date                                                                                                                                                                                                         |
| Clausa DEFAULT STORAGE are efect in dimensionarea:                                                                                                                                  | Extensiilor                                                                                                                                                                                                                |
| Copiera unei table se poate face cu:                                                                                                                                                | CREATE TABLE                                                                                                                                                                                                               |
| Crearea unei baze de date Oracle se poate face folosind credentialele userilor:                                                                                                     | SYS                                                                                                                                                                                                                        |
| Crearea unei baze de date se face :                                                                                                                                                 | Cu instanta pornita                                                                                                                                                                                                        |
| Crearea unei salvari pentru fisierelor de control se face cu comanda:                                                                                                               | ALTER DATABASE BACKUP CONTROLFILE                                                                                                                                                                                          |
| Crearea unui nou fisier de control se face:                                                                                                                                         | Dupa pornirea instantei                                                                                                                                                                                                    |
| Crearea unui nou fisier de control se face:                                                                                                                                         | Cu instanta pornita ?                                                                                                                                                                                                      |
| Daca A este o multime frecventa si s este pragul de suport atunci:                                                                                                                  | A ca multime este in cel putin s% tranzactii                                                                                                                                                                               |
| Daca a, b și c (cuvinte) apar fiecare in 2% din toate documentele si S={a, b, c} apar in 0.2% din documente, interesul lui S este aproximativ:                                      | 2,7                                                                                                                                                                                                                        |
| Daca a, b și c (cuvinte) apar fiecare in 3% din toate documentele si S={a, b, c} apar in 0.3% din documente, interesul lui S este aproximativ:                                      | 2,3                                                                                                                                                                                                                        |

|                                                                                                                                                                                                                                                 |                                                                                                                            |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|
| Daca a, b și c (cuvinte) apar fiecare în 3% din toate documentele și S={a, b, c} apar în 0.3% din documente, interesul lui S este aproximativ:                                                                                                  | 2,3                                                                                                                        |
| Daca a, b și c (cuvinte) apar fiecare în 4% din toate documentele și S={a, b, c} apar în 0.4% din documente, interesul lui S este aproximativ:                                                                                                  | 2                                                                                                                          |
| Daca am modifi ca experimental 'Carti si Autori' pentru a obtine 'Carti si Autori si Edituri', un sablon ar avea:                                                                                                                               | 7 parti componente                                                                                                         |
| Daca avem 10000 de cosuri a cate 10 articole fiecare iar pragul de suport e 1000 putem avea maxim                                                                                                                                               | 100 articole frecvente                                                                                                     |
| Daca avem 10000 de cosuri a cate 100 articole fiecare iar pragul de suport e 1000, atunci putem avea maxim                                                                                                                                      | 1000 articole frecvente                                                                                                    |
| Daca avem 10000 de cosuri a cate 100 articole fiecare, iar pragul de suport e 1000, atunci putem avea maxim:                                                                                                                                    | 1000 articole frecvente                                                                                                    |
| Daca este atinsa o limita a profilului la nivel de operatie (call) atunci:                                                                                                                                                                      | Doar operatia curenta este anulata                                                                                         |
| Daca extensiile sunt definite prin: INITIAL=100K, NEXT=200K, PCTINCREASE=100, a treia extensie va avea:                                                                                                                                         | 400K ?                                                                                                                     |
| Daca in matricea de calcul a similaritatii liniile si coloanele sunt studenti si (I, J) = 1 daca studentul I are medie egală cu studentul J atunci daca doua coloane A si B sunt 100% similară atunci:                                          | A si B au medii egale                                                                                                      |
| Daca in matricea de calcul a similaritatii liniile si coloanele sunt studenti si (I, J) = 1 daca studentul I are medie mai mare decat studentul J atunci daca doua coloane A si B sunt 100% similară atunci:                                    | A si B au medii egale                                                                                                      |
| Daca in matricea de calcul a similaritatii liniile si coloanele sunt studenti si (I, J) = 1 daca studentul I are medie mai mica decat studentul J atunci daca doua coloane A si B sunt 100% similară atunci:                                    | A si B au medii egale                                                                                                      |
| Daca in momentul blocarii unui cont de user Oracle acesta avea o tranzacție în desfășurare, aceasta:                                                                                                                                            | Este revocată (ROLLBACK)                                                                                                   |
| Daca INITIAL este 10M, NEXT este 20M, PCTINCREASE este 0 atunci extensia 3 va avea:                                                                                                                                                             | 20M                                                                                                                        |
| Daca INITIAL este 10M, NEXT este 20M, PCTINCREASE este 50 atunci extensia 3 va avea:                                                                                                                                                            | 30M                                                                                                                        |
| Daca INITIAL este 50M, NEXT este 10M, PCTINCREASE este 10 atunci extensia 3 va avea:                                                                                                                                                            | 11M                                                                                                                        |
| Daca la instalarea Oracle database nu s-a creat nicio baza de date, se poate crea ulterior folosind:                                                                                                                                            | Utilitarul Database Configuration Assistant                                                                                |
| Daca matricea pentru calcularea rangului unei pagini are linile: (1/2, 0, 1/2), (1/2, 0, 0) si (0, 0, 1/2) atunci:                                                                                                                              | Este o situatie tip Dead End                                                                                               |
| Daca matricea pentru calcularea rangului unei pagini are linile: (1/2, 1/3, 0), (0, 1/3, 0) si (1/2, 1/3, 1) atunci:                                                                                                                            | Este o situatie tip capcana (spider trap)                                                                                  |
| Daca matricea pentru calcularea rangului unei pagini are linile: (1/2, 1/3, 0), (0, 1/3, 0) si (1/2, 1/3, 1/2) atunci:                                                                                                                          | Matricea este gresit calculata                                                                                             |
| Daca matricea pentru calcularea rangului unei pagini are linile: (1/3, 1/2, 1), (1/3, 0, 0) si (1/3, 1/2, 0) atunci:                                                                                                                            | Este o situatie tip capcana (spider trap)                                                                                  |
| Daca o baza de date are asignat la startare un singur fisier de control si acesta va fi alterat fizic, se poate folosi o copie a lui pentru a restara baza de date ?                                                                            | Da, dar fisierul copie trebuie sa aiba aceeasi cale si nume ca cel alterat                                                 |
| Daca o multime de 3 articole e frecventa (pragul de suport fiind 1000) atunci:                                                                                                                                                                  | Fiecare submultime a sa de doua articole apare in cel putin 1000 de cosuri                                                 |
| Daca o tabela a fost creata cu optiunea NOPARALLEL atunci:                                                                                                                                                                                      | Tabela poate fi parcursa in paralel folosind comentarii de tip 'hint'                                                      |
| Dacă pentru 200 de articole 50 nu sunt clasificate corect, acuratețea este:                                                                                                                                                                     | 75%                                                                                                                        |
| Dacă pentru 200 de articole 50 nu sunt clasificate corect, atunci rata de eroare este:                                                                                                                                                          | 25%                                                                                                                        |
| Dacă pentru 200 de articole, toate pozitive, 50 nu sunt clasificate corect, atunci rechemarea este:                                                                                                                                             | 75%                                                                                                                        |
| Daca pentru 200 de articole, toate pozitive, 50 nu sunt clasificate corect, atunci scorul F1 este:                                                                                                                                              | 85%                                                                                                                        |
| Dacă pentru 2000 de articole 300 nu sunt clasificate corect, acuratețea este in intervalul:                                                                                                                                                     | [0,8, 1]?                                                                                                                  |
| Dacă pentru 2000 de articole 300 nu sunt clasificate corect, rata de eroare este in intervalul:                                                                                                                                                 | [0, 0,4]                                                                                                                   |
| Dacă pentru 2000 de articole din care 1000 sunt positive si restul negative, 300 de pozitive si 300 de negative nu sunt clasificate corect atunci precizia este in intervalul:                                                                  | [0,5, 0,7]                                                                                                                 |
| Dacă pentru 2000 de articole din care 1000 sunt positive si restul negativ, 300 de pozitive si 300 de negative nu sunt clasificate corect atunci scorul Z (Z-score) este in intervalul:                                                         | // Z_score: (v - v_mean) / dev standard                                                                                    |
| Dacă pentru 2000 de articole din care 1000 sunt positive si restul negativ, 300 de pozitive si 300 de negative nu sunt clasificate corect atunci senzitivitatea (recall) este in intervalul:                                                    | [0,5, 0,7]                                                                                                                 |
| Dacă pentru 2000 de articole din care 1000 sunt positive si restul negativ, 300 de pozitive si 300 de negative nu sunt clasificate corect atunci specificitatea este in intervalul:                                                             | [0,5, 0,7]?                                                                                                                |
| Dacă pentru datele unei retele de magazine o linie a tabelei de fapte reprezintă un intreg bon de cumpărături al unui client, atunci care dimensiune nu poate fi atășată schemei sale:                                                          | Produsul?                                                                                                                  |
| Daca precizia si senzitivitatea (recall) sunt egale, atunci scorul Z este:                                                                                                                                                                      | Egal cu ele?                                                                                                               |
| Daca prin modificarea unei linii a unei tabele Oracle, spatiul liber al blocului de date specificat de parametrul PCTFREE este insuficient:                                                                                                     | Linia respectiva este mutata automat in alt bloc de date unde este spatiu suficient                                        |
| Daca se inseraza o linie intr-o tabela si se doreste o lista cu userul curent, adresa sesiunii, adresa transactiei, numele si numarul segmentului de rollback folosit in sesiunea curenta a unei baze de date Oracle, se pot folosi view-urile: | V\$SESSION , V\$TRANSACTION si V\$ROLLNAME                                                                                 |
| Daca spatiul liber dintr-un bloc de date Oracle este sub procentul specificat de parametrul PCTFREE, in blocul respectiv:                                                                                                                       | Se pot face numai modificari de date                                                                                       |
| Daca un user Oracle a primit un privilegiu de insert pe o coloana a unei tabele, acel privilegiu poate fi vazut in dictionar in view-ul:                                                                                                        | DBA_COL_PRIVS                                                                                                              |
| Database Writer (DBWn) este un proces:                                                                                                                                                                                                          | de background                                                                                                              |
| Datele stocate intr-o tabela temporara Oracle pot fi accesate printr-o cerere de interogare:                                                                                                                                                    | Po toata durata sesiunii curente, in functie de parametrul specificat la creare                                            |
| DBWRITER este parte                                                                                                                                                                                                                             | Instantei Oracle                                                                                                           |
| Debloarea contului unui user Oracle se face cu comanda:                                                                                                                                                                                         | ALTER USER                                                                                                                 |
| Debloarea contului unui user Oracle se face cu comanda:                                                                                                                                                                                         | ALTER USER                                                                                                                 |
| Declarația grain-ului înseamnă a specifica                                                                                                                                                                                                      | o linie din tabela de fapte                                                                                                |
| Deschiderea fisierelor de control se face:                                                                                                                                                                                                      | la montarea BD                                                                                                             |
| Deschiderea fisierelor de date se face la momentul:                                                                                                                                                                                             | Deschiderii BD                                                                                                             |
| Determinarea valorilor de index si autoritate pentru pagini se face:                                                                                                                                                                            | Se pot determina separat                                                                                                   |
| Deviația (abaterea) standard a mulțimii {1, 1, 1, 1, 1} este:                                                                                                                                                                                   | 0                                                                                                                          |
| Deviația (abaterea) standard a mulțimii {20, 0, 20, 0, 20, 0} este:                                                                                                                                                                             | 10                                                                                                                         |
| Dictionarul de date al sistemului se memoreaza in:                                                                                                                                                                                              | Tablespace-ul SYSTEM                                                                                                       |
| Dimensiunea blocului de date din Database Buffer Cache este dat de:                                                                                                                                                                             | DB_BLOCK_SIZE                                                                                                              |
| Dimensiunea membrilor unui grup de fisiere redo log:                                                                                                                                                                                            | Trebuie sa fie aceeasi                                                                                                     |
| Dimensiunea standard a unui bloc de date dintr-o baza de date Oracle se poate seta:                                                                                                                                                             | Cand se creeaza baza de date                                                                                               |
| Dimensiunea unui bloc de date dintr-o baza de date Oracle se poate seta                                                                                                                                                                         | Cand se creeaza baza de date                                                                                               |
| Dimensiunea unui bloc de date dintr-o baza de date Oracle se poate seta cu comanda:                                                                                                                                                             | CREATE DATABASE ...                                                                                                        |
| Dimensiunea unui bloc de date dintr-o baza de date Oracle se poate seta cu comanda:                                                                                                                                                             | Nicuna dintre comenzile specificate la a si b (ALTER SYSTEM SET BLOCK_SIZE / ALTER DBA_TABLESPACES ... SET BLOCK_SIZE=...) |
| Dimensiunea zonei de memorie Redo Log Buffer e data de:                                                                                                                                                                                         | LOG_BUFFER                                                                                                                 |
| Din principiul apriori rezulta si ca daca avem doua multimi frecvente de articole nedisjuncte A si B atunci si multimea urmatoare (nevida) e frecventa:                                                                                         | A - B                                                                                                                      |

|                                                                                                                                                                                              |                                                                                                               |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|
| Din principiu apriori rezulta si ca daca avem doua multimi frecvente de articole nedisjuncte A si B atunci si multimea urmatoare (nevida) e frecventa:                                       | A - (A - B)                                                                                                   |
| Discretizarea face parte din etapa de:                                                                                                                                                       | Preprocesarea datelor                                                                                         |
| Dispersia (varianța) multimi $\{1, 1, 1, 1, 1\}$ este:                                                                                                                                       | 0                                                                                                             |
| Distanța edit intre sirurile 'Eu sunt acasă' si 'Eu nu sunt acasă' este:                                                                                                                     | 3                                                                                                             |
| Distanța Manhattan intre punctele A(1, 2) si B(59, -29) este:                                                                                                                                | 89                                                                                                            |
| DMT inseamnă:                                                                                                                                                                                | Dictionary managed tablespace                                                                                 |
| Dupa crearea unui user Oracle nou, pentru conectarea la baza de date trebuie sa i se aloce:                                                                                                  | O parola de acces si anumite privilegii de system                                                             |
| Eliminarea unui grup de REDO LOG se face cu o cerere:                                                                                                                                        | Alter database                                                                                                |
| Entropie (D) = $-\sum_j [Pr(c_j) * \log_2 Pr(c_j)]$ . Dacă D are 50% exemple pozitive si 50% negative, entropia lui D este:                                                                  | 1                                                                                                             |
| Entropie (D) = $-\sum_j [Pr(c_j) * \log_2 Pr(c_j)]$ . Dacă D contine 100% exemple negative, entropia lui D este:                                                                             | 0                                                                                                             |
| Entropie (D) = $-\sum_j [Pr(c_j) * \log_2 Pr(c_j)]$ . Dacă D contine 100% exemple pozitive, entropia lui D este:                                                                             | 0                                                                                                             |
| Entropie (D) = $-\sum_j [Pr(c_j) * \log_2 Pr(c_j)]$ . Dacă D contine in propoții egale exemple din 4 clase diferite, entropia lui D este:                                                    | 2                                                                                                             |
| Este Data mining:                                                                                                                                                                            | Folosirea unui arbore de decizie                                                                              |
| Etapile procesarii unei cereri sunt, in ordine:                                                                                                                                              | Parse Execute Fetch                                                                                           |
| Există vederi dinamice care pot fi consultate cu instantă opriță?                                                                                                                            | Nu                                                                                                            |
| Extinderea spațiului alocat unui index creat pe o tabela Oracle se poate face prin:                                                                                                          |                                                                                                               |
| Faptul că datele dintr-un depozit de date sunt non-volatilă înseamnă că:                                                                                                                     | O data încărcată în depozit date nu mai sunt modificate sau șterse.                                           |
| Faptul că un depozit de date este orientat (axat) pe subiecte înseamnă că:                                                                                                                   | Datele sunt organizate considerând categoriile de informații stocate. ?                                       |
| Fie coloana C = [1, 0, 0, 1, 1, 0, 1, 1, 1]. O signație calculată cu dispersia de tip 3-min poate fi:                                                                                        | 356                                                                                                           |
| Fie coloana C = [1, 0, 1, 0, 0, 1, 0, 0, 1]. O signație calculată cu dispersia de tip 3-min poate fi:                                                                                        | 789                                                                                                           |
| Fie coloana C = [1, 0, 1, 0, 0, 0, 1, 0, 1]. O signație calculată cu dispersia de tip Min poate fi si:                                                                                       | 765                                                                                                           |
| Fie coloana C = [1, 0, 1, 0, 1, 0, 0, 0]. O signație calculată cu dispersia de tip Min poate fi si:                                                                                          | 567                                                                                                           |
| Fie coloana C = [1, 0, 1, 0, 1, 0, 1, 0]. O signație calculată cu dispersia de tip 3-Min poate fi si:                                                                                        | 678                                                                                                           |
| Fie coloana C = [1, 0, 1, 0, 1, 0, 1, 0]. O signație calculată cu dispersia de tip Min poate fi si:                                                                                          | 555                                                                                                           |
| Fie coloana C = [1, 0, 1, 1, 1, 1, 1, 1, 1]. O signație calculată cu dispersia de tip 3-min poate fi:                                                                                        | 123                                                                                                           |
| Fie coloanele C1 = [1, 0, 0, 1, 1, 0, 1, 1] si C2 = [1, 0, 1, 0, 1, 0, 1, 0]. Atunci ele sunt:                                                                                               | 50% similară                                                                                                  |
| Fie coloanele C1 = [1, 1, 1, 1, 0, 1, 1, 1] si C2 = [1, 0, 1, 0, 0, 1, 0, 0]. Atunci ele sunt:                                                                                               | 33% similară                                                                                                  |
| Fie coloanele C1 = [1, 1, 1, 1, 0, 1, 1, 1] si C2 = [1, 0, 1, 0, 1, 0, 1, 1]. Atunci ele sunt:                                                                                               | 66% similară                                                                                                  |
| Fie coloanele C1 = [1, 1, 1, 1, 1, 1, 1, 1] si C2 = [1, 0, 1, 0, 1, 0, 0, 0]. Atunci ele sunt:                                                                                               | 40% similară                                                                                                  |
| Fie in plan punctele A, B and C si distanta dintre ele $d(A, B) = 3$ , $d(A, C) = 5$ , $d(B, C) = 4$ (numere pitagoreice). Folosind FastMap, coordonata lui A pe axa BC (origine in B) este: | 0                                                                                                             |
| Fie in plan punctele A, B and C si distanta dintre ele $d(A, B) = 3$ , $d(A, C) = 5$ , $d(B, C) = 4$ (numere pitagoreice). Folosind FastMap, coordonata lui C pe axa AB (origine in A) este: | 3                                                                                                             |
| Fie in plan punctele A, B and C si distanta dintre ele $d(A, B) = 3$ , $d(A, C) = 5$ , $d(B, C) = 4$ (numere pitagoreice). Folosind FastMap, coordonata lui C pe axa AB (origine in A) este: | 3                                                                                                             |
| Fie in plan punctele A, B and C si distanta dintre ele $d(A, B) = 3$ , $d(A, C) = 5$ , $d(B, C) = 4$ (numere pitagoreice). Folosind FastMap, coordonatalui A pe axa BC (origine in B) este:  | 0                                                                                                             |
| Fie multimea de cosuri: $\{(1, 2, 3, 4), (2, 3, 4, 5), (3, 4, 5, 6)\}$ si pragul de suport 50%. Atunci avem:                                                                                 | 5 perechi frecvente                                                                                           |
| Fie multimea de cosuri: $\{(1, 2, 3, 4), (2, 3, 4, 5), (3, 4, 5, 6)\}$ si pragul de suport 50%. Atunci avem:                                                                                 | 4 articole frecvente                                                                                          |
| Fie multimea de cosuri: $\{(1, 2, 3, 4), (2, 3, 4, 5), (3, 4, 5, 6)\}$ si pragul de suport 50%. Atunci avem:                                                                                 | 2 articole frecvente                                                                                          |
| Fie multimea de cosuri: $\{(1, 2, 3, 4), (2, 4, 5), (1, 3, 4)\}$ si pragul de suport 50%. Atunci avem:                                                                                       | 4 perechi frecvente                                                                                           |
| Fie multimea de cosuri: $\{(1, 3, 4), (2, 3, 5), (1, 2, 3, 5), (2, 5), (1, 2, 3, 5)\}$ si pragul de suport 80%. Atunci numarul de multimi frecvente maximale este:                           | 3                                                                                                             |
| Fie multimea de cosuri: $\{(1, 3, 4), (2, 3, 5), (1, 2, 3, 5), (2, 5), (1, 2, 3, 5)\}$ si pragul de suport 80%. Atunci regula 2 → 5 are incredere:                                           | 80%                                                                                                           |
| Fie multimea de cosuri: $\{(1, 3, 4), (2, 3, 5), (1, 2, 3, 5), (2, 5), (1, 2, 3, 5)\}$ si pragul de suport 80%. Atunci regula 3 → 1 are incredere:                                           | 75% ?                                                                                                         |
| Fie punctele $\{1, 2, 3, 4, 5, 6\}$ . Daca aplicam K-Means pentru K = 3 si centroizi initiali 1, 2 si 3 obtinem clusterlele:                                                                 | (1), (2, 3), (4, 5, 6)                                                                                        |
| Fie punctele $\{1, 2, 5, 6\}$ in 1D. Daca aplicam K-Means pentru K = 2 si centroizi initiali 1 si 2 obtinem clusterlele:                                                                     | (1, 2) and (5, 6)                                                                                             |
| Fisierele de control care se creeaza intr-o baza de date Oracle pot fi specificate folosind comenzi                                                                                          | Ambele comenzi specificate la a si b pot fi folosite, in functie de context (CREATE DATABASE / ALTER SYSTEM)  |
| Fisierele de control se creeaza automat:                                                                                                                                                     | La crearea BD                                                                                                 |
| Fisierele de control sunt de tipul :                                                                                                                                                         | Binar                                                                                                         |
| Fisierele de control sunt scrisă:                                                                                                                                                            | In paralel (multiplexat)                                                                                      |
| Fisierele de date contin printre altele:                                                                                                                                                     | Dictionarul de date si obiectele userului                                                                     |
| Fisierele de log care se creeaza intr-o baza de date Oracle pot fi specificate folosind comenzi:                                                                                             | Ambele comenzi specificate la a si b pot fi folosite, in functie de context (CREATE DATABASE/ ALTER DATABASE) |
| Fisierele de redo-log sunt scrise de:                                                                                                                                                        | Nici a (Procesul server) nici b (Procesul client) nu sunt corecte                                             |
| Fisierele de redo-log sunt scrise de:                                                                                                                                                        | Un proces de background                                                                                       |
| Fisierele de tip Redo Log se folosesc pentru:                                                                                                                                                | Recuperarea dupa incident                                                                                     |
| Fisierele de tip REDO LOG sunt folosite:                                                                                                                                                     | Pentru recuperarea datelor (recovery)                                                                         |
| Fisierul de parametri INIT.ORA este de tip:                                                                                                                                                  | Text                                                                                                          |
| Fisierele de control sunt de tipul                                                                                                                                                           | Binar                                                                                                         |
| Fortarea unui checkpoint se face cu comanda:                                                                                                                                                 | ALTER SYSTEM CHECKPOINT                                                                                       |
| Fortarea unui log switch se face cu comanda:                                                                                                                                                 | ALTER SYSTEM SWITCH LOGFILE                                                                                   |
| Gradele militare sunt valori de tip:                                                                                                                                                         | Ordinal                                                                                                       |
| High Water Mark poate fi mutat in sus printr-o cerere SQL de tip:                                                                                                                            | ALTER TABLE                                                                                                   |
| HWM poate fi mutat in jos:                                                                                                                                                                   | Ca efect al unor comenzi DDL                                                                                  |

|                                                                                                                                                                                                                      |                                                                                                            |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|
| In HWM poate fi mutat în sus:                                                                                                                                                                                        | Ca efect al unor comenzi DML                                                                               |
| In 1D, funcția $D(x, a) = x + \sqrt{y}$ poate fi folosită ca funcție de distanță?                                                                                                                                    | Nu                                                                                                         |
| In 1D, funcția $D(x, y) =  x - y $ poate fi folosită ca funcție de distanță?                                                                                                                                         | Da                                                                                                         |
| In 1D, funcția $D(x, y) = \sqrt{ x - y }$ poate fi folosită ca funcție de distanță?                                                                                                                                  | Nu                                                                                                         |
| In 1D, funcția $D(x, y) = \sqrt{x} + \sqrt{y}$ poate fi folosită ca funcție de distanță?                                                                                                                             | Nu                                                                                                         |
| In 3D, distanța cea mai mare dintre două puncte este data în general de:                                                                                                                                             | Distanța Manhattan (norma L1)                                                                              |
| In 3D, distanța cea mai mică dintre două puncte este data în general de:                                                                                                                                             | Maximul pe o dimensiune                                                                                    |
| In algoritmului BFR pentru un spațiu având $k$ dimensiuni, pentru fiecare Discard Set se tin:                                                                                                                        | $2k + 1$ numere                                                                                            |
| In cadrul algoritmului k-means, se procedea succesiv la:                                                                                                                                                             | Asocierea punctelor la centoizi                                                                            |
| In cadrul cursului nostru DICE și abrevierea pentru:                                                                                                                                                                 | Dynamic Itemset Counting Engine                                                                            |
| In care dintre următoarele operații se folosesc segmentele de tip Undo într-o baza de date Oracle                                                                                                                    | Modificare și stergere de date în baza de date                                                             |
| In cazul algoritmului BFR aplicat pe multimea particulelor de praf dintr-o incapere, pentru fiecare grup de tip Discard Set având 1000 de particule se retin statistici formate din numere unde:                     |                                                                                                            |
| In cazul algoritmului BFR aplicat pe multimea tantarilor dintr-o incapere, pentru fiecare grup de tip Compression Set având 10 de tantari se retin statistici formate din numere unde:                               |                                                                                                            |
| In cazul calculului rangului paginii, Dead End semnifica:                                                                                                                                                            | O pagină care nu are succesor                                                                              |
| In cazul calculului rangului paginii, o Capcana semnifica:                                                                                                                                                           | Un grup de una sau mai multe pagini care nu au legături către pagini din afara grupului                    |
| In cazul carei opriri nu sunt închise fisierile:                                                                                                                                                                     | Abort                                                                                                      |
| In cazul carei opriri se executa ROLLBACK pentru tranzacțiile active:                                                                                                                                                | Imediat și Abort                                                                                           |
| În cazul căutării mulțimilor frecvente în datele de vânzări ale unui supermarket, reducerea dimensionalității ("Dimensionality reduction") trebuie să păstreze:                                                      | Produsul                                                                                                   |
| In cazul cererii CREATE TABLE, suma dintre PCTFREE și PCTUSED trebuie sa fie                                                                                                                                         | Intre 1 și 100                                                                                             |
| In cazul cererii CREATE TABLE, valoarea lui PCTFREE poate fi:                                                                                                                                                        | Intre 1 și 99                                                                                              |
| In cazul comenzii ALTER USER ... ACCOUNT LOCK:                                                                                                                                                                       | Sesiunea curentă nu este afectată                                                                          |
| In cazul creării unei tabele, clauza CACHE semnifica:                                                                                                                                                                | La o parcurgere completă, tabela va fi plasată în memorie în zona celor mai recent utilizate blocuri       |
| In cazul creării unei tabele, clauza NOCACHE semnifica:                                                                                                                                                              | La o parcurgere completă, tabela va fi plasată în memorie în zona celor mai puțin recent utilizate blocuri |
| In cazul experimentului 'Carti si Autori', un sâmbon are:                                                                                                                                                            | 5 parti componente                                                                                         |
| In cazul FastMap, dacă avem punctele A, B și C având distanțele $D(A, B) = D(A, C) = D(B, C) = 1$ , dacă primul pas se alege ca axa BC, atunci distanța reportată pentru pasul al doilea între punctele A și B este: | $\sqrt{3}/2$                                                                                               |
| In cazul implementării tranzacțiilor AF prin automate finite, procesul de obținere a acestora presupune introducerea unor obiecte administrative, printre care și:                                                   | Functie de transmisie                                                                                      |
| In cazul în care la crearea unei tablespaces se specifică SEGMENT SPACE MANAGEMENT AUTO, spațul liber din segment este gestionat:                                                                                    | Cu ajutorul unor bitmapuri                                                                                 |
| In cazul în care la crearea unei tablespaces se specifică SEGMENT SPACE MANAGEMENT MANUAL, spațul liber din segment este gestionat:                                                                                  | Cu ajutorul unor liste                                                                                     |
| In cazul în care se crează un cluster STUD cu 2 tabele stud1 și stud2 având fiecare câte 10 linii, cererea SELECT * FROM STUD returnează:                                                                            | O eroare                                                                                                   |
| In cazul lansării unei operații de oprire a bazei de date, sunt permise noi conexiuni în cazul opririi:                                                                                                              | Nu se mai permit noi conexiuni                                                                             |
| In cazul normalizării de tip Scalare zecimală, mulțimea {1, 2, 3, 6} devine:                                                                                                                                         | {0,1, 0,2, 0,3, 0,6}                                                                                       |
| In cazul normalizării de tip z-score pentru mulțimea {1, 2, 2, 3} obținem:                                                                                                                                           | Mai multe valori de 0                                                                                      |
| In cazul normalizării de tip z-score pentru mulțimea {1, 2, 2, 3} obținem:                                                                                                                                           | O singură valoare strict pozitivă                                                                          |
| In cazul normalizării de tip z-score pentru mulțimea {1, 2, 3, 10} obținem:                                                                                                                                          | Nici o valoare de 0                                                                                        |
| In cazul normalizării de tip z-score pentru mulțimea {1, 2, 3, 6} obținem un set cu:                                                                                                                                 | O singură valoare strict pozitivă                                                                          |
| In cazul normalizării Min-Max, mulțimea {1, 2, 3, 6} devine:                                                                                                                                                         | {0,0, 0,2, 0,4, 1}                                                                                         |
| In cazul Random Forest, numărul atributelor luate în considerare la fiecare ramificare este:                                                                                                                         | Mai mic decât numărul de atribute ale setului de antrenament                                               |
| In cazul separării SVM / neliniare, asa-numitul spațiu caracteristic - Feature Space - are de obicei:                                                                                                                | Mai multe dimensiuni                                                                                       |
| In cazul SVM, o funcție kernel este de tipul:                                                                                                                                                                        | $K(X_i, X_j) = f_1(X_i) * f_1(X_j)$                                                                        |
| In cazul tabelelor de partitie:                                                                                                                                                                                      | Liniile sunt împărțite în mai multe grupuri                                                                |
| In cazul tabelelor de tip cluster (clustered tables):                                                                                                                                                                | Acestea sunt parte a unui grup de tabele având în comun anumite coloane                                    |
| In cazul tabelelor de tip cluster (clustered tables):                                                                                                                                                                | Acestea sunt parte a unui grup de tabele având în comun anumite coloane                                    |
| In cazul tabelelor organizate ca index (index organized), acestea se pot partiona?                                                                                                                                   | Da, dacă atributul după care se face partionarea este o submultime a cheii primare.                        |
| In cazul tabelelor organizate ca index (index organized):                                                                                                                                                            | Înregistrările sunt organizate ca arbore B                                                                 |
| In cazul tabelelor partionate, partitii pot să fie stocate în tablespaces-uri diferite?                                                                                                                              | Da                                                                                                         |
| In cazul tabelelor partionate:                                                                                                                                                                                       | Liniile sunt împărțite în mai multe grupuri                                                                |
| In cazul tabelelor organizate ca index (index organized), acestea se pot partiona?                                                                                                                                   | Da, dacă atributul după care se face partionarea este o submultime a cheii primare.                        |
| In cazul unei reguli de asociere $X \rightarrow Y$ , X și Y sunt:                                                                                                                                                    | Mulțimi de articole                                                                                        |
| In cazul unei tabele create cu CREATE GLOBAL TEMPORARY TABLE, datele inserate sunt vizibile:                                                                                                                         | Doar sesiunii care le inserează                                                                            |
| In cazul unei tabele create cu CREATE GLOBAL TEMPORARY TABLE, datele inserate:                                                                                                                                       | Sistemul le sterge la terminarea sesiunii care le-a inserat?                                               |
| In cazul unui user identificat prin parola, care dintre parolele de mai jos este corectă:                                                                                                                            | PA_S#ROLAS#_USER12345678901234                                                                             |
| In cazul unui user identificat prin parola, care dintre parolele de mai jos este gresita:                                                                                                                            | PAROLA_USER_1_1234%67890123                                                                                |
| În cursul nostru ID3 înseamnă:                                                                                                                                                                                       | Interactive Dichotomizer 3                                                                                 |
| În cursul nostru SCN înseamnă:                                                                                                                                                                                       | System Change Number                                                                                       |
| In formule de calcul pentru gradul de autoritate și index pentru pagini, $\lambda$ și $\mu$ sunt:                                                                                                                    | Numere reale                                                                                               |
| In funcție de gestiunea extensiilor, un tablespace poate fi de tipul:                                                                                                                                                | LMT sau DMT                                                                                                |
| In matricea stochastică pentru determinarea rangului paginii numarul de succesiuni ai unei pagini este dat de:                                                                                                       | Numarul de valori non zero de pe coloana corespunzătoare paginii                                           |
| In matricea stochastică pentru determinarea rangului paginii suma este 1 pe:                                                                                                                                         | Coloane                                                                                                    |
| În metodele asambliste:                                                                                                                                                                                              | Mai mulți clasificatorii sunt agregați într-un singur                                                      |

|                                                                                                                                                                                                                                                                                                          |                                                                                                                                    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|
| In momentul blocarii unui cont (LOCK), daca userul e logat la acel moment:                                                                                                                                                                                                                               | Sesiunea nu va fi afectata de schimbare                                                                                            |
| In Oracle SYS este:                                                                                                                                                                                                                                                                                      | Un user                                                                                                                            |
| In plan, distanta cea mai mare dintre doua puncte este data in general de:                                                                                                                                                                                                                               | Distanta Manhattan (norma L1)                                                                                                      |
| In plan, distanta cea mai mica dintre doua puncte este data in general de:                                                                                                                                                                                                                               | Maximul pe o dimensiune                                                                                                            |
| In schema unui user exista printre altele obiecte de tipurile:                                                                                                                                                                                                                                           | Indescribabile si tipuri                                                                                                           |
| In segmentele stocate intr-un tablespace permanent dintr-o baza de date Oracle, trecut in starea READ ONLY, se pot face operatii de:                                                                                                                                                                     | Ambele tipuri de operatii specificate la a si b (Citire de date / Scriere de date pana la terminarea tranzactiilor in desfasurare) |
| Increderea unei reguli de asociere X → Y este data de formula:                                                                                                                                                                                                                                           | Suport (X ∪ Y) / Suport (X)                                                                                                        |
| Intr-o baza de date Oracle segmentele pot fi:                                                                                                                                                                                                                                                            | Toate obiectele specificate la a si b (Tabele permanente si tabele temporare / Indescribabile si view-uri)                         |
| Intr-un spatiu ID avem 2 puncte rosii (1 și 3) si 2 puncte albastre (10 și 11). Folosind kNN pentru k = 3, punctul 5 este albastru?                                                                                                                                                                      | Nu                                                                                                                                 |
| Intr-un spatiu ID avem 2 puncte rosii (1 și 3) si 2 puncte albastre (10 și 11). Folosind kNN pentru k = 3, punctul 7 este albastru?                                                                                                                                                                      | Da                                                                                                                                 |
| IQR pentru multimea {20, 0, 20, 0, 20, 0} este:                                                                                                                                                                                                                                                          | 20                                                                                                                                 |
| KNN                                                                                                                                                                                                                                                                                                      | Este doar o metoda de clasificare care nu produce un clasificator                                                                  |
| La crearea unui user Oracle trebuie specificat obligatoriu in comanda de creare:                                                                                                                                                                                                                         | Parola de conexiune la baza de date                                                                                                |
| La determinarea rangului unei pagini (fara taxare) daca in final se ajunge la [0,0, ...]                                                                                                                                                                                                                 | Există cel putin un Dead End                                                                                                       |
| La determinarea rangului unei pagini (fara taxare) daca in final se ajunge la [0,0, ...]                                                                                                                                                                                                                 | Există pagini fară succesor                                                                                                        |
| La determinarea rangului unei pagini (fara taxare) daca in final se ajunge la un vector cu o singura componenta nula:                                                                                                                                                                                    | Există doar o pagină autoreferită                                                                                                  |
| La determinarea rangului unei pagini: o pagina este importantă daca:                                                                                                                                                                                                                                     | E referita de pagină importantă                                                                                                    |
| La execuția comenzii ALTER SYSTEM ENABLE RESTRICTED SESSION sesiunile deja existente:                                                                                                                                                                                                                    | Rămân active până la terminarea lor                                                                                                |
| La inchiderea bazei de date blocurile de date modificate din memorie sunt scrise pe disc de:                                                                                                                                                                                                             | Procesul Database Writer (DBWn)                                                                                                    |
| La startarea unei baze de date Oracle fisierele de control sunt asignate la baza de date din :                                                                                                                                                                                                           | Parametrul CONTROL_FILES ?                                                                                                         |
| La startarea unei instante Oracle parametrii de initializare pot fi prelauți din fisierul:                                                                                                                                                                                                               | SPFILE                                                                                                                             |
| La startarea unei instante Oracle parametrii de initializare pot fi prelauți din:                                                                                                                                                                                                                        | Fisierul persistent SPFILE                                                                                                         |
| La trecerea unui tablespace OFFLINE se pot folosi parametri:                                                                                                                                                                                                                                             | NORMAL si IMMEDIATE                                                                                                                |
| LCS('Ionescu Vasile', 'Popescu Ion') are valoarea: (Obs: LCS=cea mai lungă secvență comună)                                                                                                                                                                                                              | 6 ("oescu ")                                                                                                                       |
| Lista cu fisierelor de control se poate obține cu comanda:                                                                                                                                                                                                                                               | SELECT VALUE FROM V\$PARAMETER WHERE NAME = 'control_files';                                                                       |
| Lista cu fisierelor de date se poate obține cu comanda:                                                                                                                                                                                                                                                  | SELECT NAME FROM V\$DATAFILE;                                                                                                      |
| Lista cu fisierelor de redo log se poate obține cu comanda:                                                                                                                                                                                                                                              | SELECT MEMBER FROM V\$LOGFILE;                                                                                                     |
| Lista cu fisierelor de redo log se poate obține cu comanda:                                                                                                                                                                                                                                              | SELECT MEMBER FROM V\$LOGFILE                                                                                                      |
| LMT inseamnă:                                                                                                                                                                                                                                                                                            | Locally Managed tablespace                                                                                                         |
| Localizarea și deschiderea fisierelor de control se face la:                                                                                                                                                                                                                                             | Montarea BD                                                                                                                        |
| Locația fisierelor de control este afișata de sistem din:                                                                                                                                                                                                                                                | Fisierul de initializare de (tip INIT.ORA)                                                                                         |
| Log Writer (LGWR) este proces:                                                                                                                                                                                                                                                                           | de background                                                                                                                      |
| Luat în considerare o populație cu 40% bărbiți și 60% femei. Fetele poartă pantaloni scurți sau fuste în proporție egală, iar bărbiții doar pantaloni scurți. Un observator vede de la mare distanță o persoană purtând pantaloni scurți. Care este probabilitatea ca persoana respectivă să fie femeie? | aprox. 43%                                                                                                                         |
| Luat în considerare o populație cu 50% bărbiți și 50% femei. 1/4 dintre fete poartă pantaloni scurți, 3/4 poartă fuste și bărbiții doar pantaloni scurți. Un observator vede de la mare distanță o persoană purtând pantaloni scurți. Care este probabilitatea ca persoana respectivă să fie femeie?     | 20%                                                                                                                                |
| Luat în considerare o populație cu 50% bărbiți și 50% femei. Fetele poartă pantaloni scurți sau fuste în proporție egală, iar bărbiții doar pantaloni scurți. Un observator vede de la mare distanță o persoană purtând pantaloni scurți. Care este probabilitatea ca persoana respectivă să fie femeie? | aprox. 33%                                                                                                                         |
| Membrii unui grup de fisier redoblog sunt scrisi:                                                                                                                                                                                                                                                        | In paralel (multiplexat)                                                                                                           |
| Modificarea parametrilor unei baze de date Oracle se poate face cu comanda:                                                                                                                                                                                                                              | ALTER SYSTEM SET parameter_name = parameter_value                                                                                  |
| Montarea bazei de date se face:                                                                                                                                                                                                                                                                          | Dupa pornirea instantei                                                                                                            |
| Multiplexarea fisierelor de control intr-o baza de date Oracle se face cu comanda                                                                                                                                                                                                                        | ALTER SYSTEM SET control_files ='\$HOME/file1.ctl','\$HOME/file2.ctl' SCOPE=SPFILE                                                 |
| Numarul blocurilor alocate unui index si procentul utilizat din spatiul alocat, se poate face printre-o interogare pe view-ul:                                                                                                                                                                           |                                                                                                                                    |
| Numarul de blocuri de date din Database Buffer Cache este dat de:                                                                                                                                                                                                                                        | DB_BLOCK_BUFFERS                                                                                                                   |
| Numarul maxim de tablespace-uri nu poate depasi numarul fisierelor de:                                                                                                                                                                                                                                   | Date                                                                                                                               |
| Numărul minim de fisier de tip Fisier de Control este:                                                                                                                                                                                                                                                   | Unul                                                                                                                               |
| Numărul minim de fisier de tip Fisier de Date este:                                                                                                                                                                                                                                                      | Unul                                                                                                                               |
| Numărul minim de fisier de tip Fisier de Redo Log este:                                                                                                                                                                                                                                                  | Două                                                                                                                               |
| Numele fisierelor de control folosite de o baza de date Oracle pot fi afilate din:                                                                                                                                                                                                                       | VSCONTROLFILE                                                                                                                      |
| Numele fisierelor de log folosite de o baza de date Oracle pot fi afilate din                                                                                                                                                                                                                            | VSLOGFILE                                                                                                                          |
| Numele segmentelor de undo utilizate in sesiunea curenta se pot vizualiza folosind view-ul:                                                                                                                                                                                                              | VSROLLNAME                                                                                                                         |
| Numele si calea fisierelor de control create intr-o baza de date Oracle pot fi afilate din următoarele view-uri:                                                                                                                                                                                         | Ambele view-uri specificate la a si b contin aceasta informatie (V\$PARAMETER / VSCONTROLFILE)                                     |
| Numele si calea fisierelor de control folosite de o baza de date Oracle pot fi afilate folosind:                                                                                                                                                                                                         | Interogare pe V\$PARAMETER/VSCONTROLFILE sau comanda SHOW PARAMETER                                                                |
| Numele si calea fisierelor de log create intr-o baza de date Oracle pot fi afilate din următoarele view-uri:                                                                                                                                                                                             | V\$LOGFILE                                                                                                                         |
| Numele si calea fisierelor de log folosite de o baza de date Oracle pot fi afilate folosind:                                                                                                                                                                                                             | Interogare pe V\$LOGFILE                                                                                                           |
| Numele si calea fisierelor temporare folosite de o baza de date Oracle pot fi afilate folosind:                                                                                                                                                                                                          | Interogare pe V\$TEMPFILE                                                                                                          |
| Numele view-urilor dinamice din dicționarul bazei de date Oracle pot fi vizualizate printre-o interogare pe:                                                                                                                                                                                             | VSFIXED_TABLE                                                                                                                      |
| O baza de date contine un numar de tablespace:                                                                                                                                                                                                                                                           | &gt;1                                                                                                                              |
| O dimensiune degenerată reprezintă:                                                                                                                                                                                                                                                                      | Numeri operaționale de control (număr factură, număr comandă, etc.)                                                                |
| O funcție la distanță trebuie să îndeplinească și condiția:                                                                                                                                                                                                                                              | $f(x, y) \&lt;= f(x, z) + f(z, y)$                                                                                                 |
| O instanta Oracle se compune din:                                                                                                                                                                                                                                                                        | O zonă de memorie și procese de background                                                                                         |
| O instanță Oracle este compusă din:                                                                                                                                                                                                                                                                      | O zonă de memorie numită SGA și procese de background                                                                              |

|                                                                                                                                                                                                                                                                            |                                                                                                                                                        |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| O instanta Oracle poate fi folosita pentru a deschide simultan doua baze de date Oracle?                                                                                                                                                                                   | Nu                                                                                                                                                     |
| O sesiune Oracle reprezinta:                                                                                                                                                                                                                                               | Conexiunea unui user la baza de date                                                                                                                   |
| O tabelă de dimensiuni reprezintă:                                                                                                                                                                                                                                         | Entități ale lumii reale și nu procese de afaceri                                                                                                      |
| O tabelă de fapte (măsuri) reprezintă:                                                                                                                                                                                                                                     | Un proces de afaceri                                                                                                                                   |
| O tabela Oracle aflată în relație cu o altă tabela pe care s-a creat o constrângere de tip FOREIGN KEY poate fi stearsa din dicționar daca:                                                                                                                                | Se dezactivează mai întâi constrângerea de integritate                                                                                                 |
| O tabela Oracle care este relaționată cu altă tabela (pe care s-a creat o constrângere FOREIGN KEY), poate fi stearsa din dicționar daca:                                                                                                                                  | Se sterg mai întâi datele relate din tabela cu care este în relație (pe care s-a creat constrângerea)                                                  |
| O tabela Oracle care este relaționată cu altă tabela (pe care s-a creat o constrângere FOREIGN KEY), poate fi stearsa din dicționar daca:                                                                                                                                  | Se sterg mai întâi datele relate din tabela cu care este în relație (pe care s-a creat constrângerea)                                                  |
| O tabela Oracle care este relaționată cu altă tabela (pe care s-a creat o constrângere FOREIGN KEY), poate fi stearsa din dicționar daca:                                                                                                                                  | // după cat^                                                                                                                                           |
| Obiectele din dicționarul bazei de date Oracle sunt create în tablespaci-uri de tip                                                                                                                                                                                        | SYSTEM                                                                                                                                                 |
| Oprirea unei baze de date Oracle în mod tranzacțional implica:                                                                                                                                                                                                             | Finalizarea tranzacțiilor curente și deconectarea imediata a userilor                                                                                  |
| Opțiunea FLASHBACK ON de la crearea unui tablespace este utilizata pentru:                                                                                                                                                                                                 | A reduse baza de date la o stare anterioara                                                                                                            |
| Parametrul LOG_CHECKPOINT_INTERVAL se masoara in :                                                                                                                                                                                                                         | Blocuri                                                                                                                                                |
| Parametrul LOG_CHECKPOINT_TIMEOUT se masoara in :                                                                                                                                                                                                                          | Secunde                                                                                                                                                |
| Parametrul PCTFREE al unui bloc de date Oracle reprezinta:                                                                                                                                                                                                                 | Procentul rezervat pentru creșterea spațiului de stocare alocat liniilor din blocul respectiv în urma operațiilor de update pe acelle lini             |
| Parametrul PCTUSED al unui bloc de date Oracle reprezinta:                                                                                                                                                                                                                 | Procentul de spațiu ocupat sub care blocul este trecut în lista de blocuri disponibile pentru inserare de date                                         |
| Partiile create pe o tabela Oracle pot fi stocate in:                                                                                                                                                                                                                      | Ambele cazuri specificate la a și b (Același tablespace ca cel al tabeliei / Tablespace diferit decât cel al tabeliei)                                 |
| Pe tabelele stocate intr-un tablespace permanent dintr-o baza de date Oracle, aflat în starea OFF LINE, se pot face:                                                                                                                                                       | Nu se poate face nicio operatie deoarece se genereaza eroare                                                                                           |
| Pentru 10000 tranzactii cu 10 articole fiecare si s = 10000 putem avea cel mult:                                                                                                                                                                                           | 10 articole frecvente                                                                                                                                  |
| Pentru a afla care este grupul curent al fisierelor de log dintr-o baza de date Oracle se poate face o interogare pe:                                                                                                                                                      | V\$THREAD                                                                                                                                              |
| Pentru a afla care este membrul curent al unui grup din fisierele de log dintr-o baza de date Oracle, se poate face o interogare pe:                                                                                                                                       | V\$LOG                                                                                                                                                 |
| Pentru a afla care membri ai unui grup din fisierele de log dintr-o baza de date Oracle sunt inactivi, se poate face o interogare pe:                                                                                                                                      | V\$LOG                                                                                                                                                 |
| Pentru a afla dimensiunea în blocuri a fisierelor de control dintr-o baza de date Oracle poate fi folosit view-ul dinamic                                                                                                                                                  | V\$CONTROLFILE                                                                                                                                         |
| Pentru a afla din dicționarul bazei de date Oracle care este spațiu liber, ca număr de blocuri, în tablespace-ul permanent aferent utilizatorului curent, se poate face o interogare pe:                                                                                   | DBA_FREE_SPACE și USER_USERS                                                                                                                           |
| Pentru a afla din dicționarul bazei de date Oracle care sunt îndecsi creati pe tabelele din utilizator curent, în ce tablespace sunt creati și care blocuri le sunt alocate, se poate face o interogare pe:                                                                | DBA_SEGMENTS și DBA_USERS                                                                                                                              |
| Pentru a afla din dicționarul bazei de date Oracle care sunt indecsi creati pe tabelele din utilizator curent, în ce tablespace sunt creati și care blocuri le sunt alocate, se poate face o interogare pe:                                                                | DBA_USERS și DBA_FREE_SPACE                                                                                                                            |
| Pentru a afla din dicționarul bazei de date Oracle marimea în blocuri a tablespace-ului permanent asignat utilizatorului curent și în ce fisier este creat spațiu respectiv, se poate face o interogare pe view-urile:                                                     | DBA_TABLESPACES și USER_USERS ???                                                                                                                      |
| Pentru a afla proprietarul și data la care a fost creată o tabela în baza de date se poate face o interogare pe:                                                                                                                                                           | DBA_OBJECTS                                                                                                                                            |
| Pentru a afla structura unei tabele din baza de date:                                                                                                                                                                                                                      | Informația se poate obține prin ambele metode specificate la a și b (Se poate face o interogare pe DBA_TAB_COLUMNS Se poate folosi comanda DESCRIBE)   |
| Pentru a crea o constrângere de integritate pe o tabela Oracle se poate folosi comanda:                                                                                                                                                                                    | ALTER TABLE table_name ADD CONSTRAINT constraint_name ...                                                                                              |
| Pentru a face o extensie unei tabele Oracle, într-un fisier nou creat asignat la baza de date, se pot folosi comenzi:                                                                                                                                                      | ALTER TABLESPACE și ALTER TABLE                                                                                                                        |
| Pentru a face o extensie unei tabele Oracle, într-un fisier nou creat asignat la baza de date, se pot folosi comenzi :                                                                                                                                                     | ALTER TABLESPACE și ALTER TABLE                                                                                                                        |
| Pentru a face o lista cu adresa tranzacției, ID-ul segmentului de rollback pe care îl folosește, nr. de blocuri generate și numele fisierului de rollback utilizat din sesiunea curentă a unei baze de date Oracle, se pot folosi view-urile:                              | V\$TRANSACTION și DBA_DATA_FILES                                                                                                                       |
| Pentru a face o lista cu fisierele temporare create intr-o baza de date Oracle și starea lor, se poate face o interogare pe:                                                                                                                                               | V\$TEMPFILE                                                                                                                                            |
| Pentru a face o lista cu numele fisierelor de log dintr-o baza de date Oracle, fisierile membru și dimensiunea lor se execută o cerere SELECT pe:                                                                                                                          | V\$LOGFILE și V\$LOG                                                                                                                                   |
| Pentru a face o lista cu numele instantei curente și grupurile fisierelor de log afisate în starea OPEN intr-o baza de date Oracle, se execută o cerere SELECT pe:                                                                                                         | V\$THREAD                                                                                                                                              |
| Pentru a face o lista cu numele instantei curente, numele tablespace-ului permanent asignat utilizatorului curent, numele tabelelor create de utilizator curent și numarul maxim de extensii permise pentru fiecare tabela Oracle, se pot folosi view-urile:               | V\$INSTANCE și USER_TABLES                                                                                                                             |
| Pentru a face o lista cu numele tablespace-ului alocat pentru segmentele temporare de sortare din sesiunea curentă a unei baze de date Oracle, numărul de extensii și blocuri libere, precum și fisierul alocat, se pot folosi view-urile:                                 | V\$SORT_SEGMENT și DBA_TEMP_FILES                                                                                                                      |
| Pentru a face o lista cu numele, marimea în bytes și starea segmentelor undo din sesiunea curentă a unei baze de date Oracle, se pot folosi view-urile:                                                                                                                    | V\$ROLLNAME și V\$ROLLSTAT                                                                                                                             |
| Pentru a face o lista cu numele, tipul și starea tablespace-ului alocat pentru segmentele temporare de sortare din sesiunea curentă a unei baze de date Oracle, precum și numărul maxim de blocuri de sortare alocate fiecarui segment temporar, se pot folosi view-urile: | V\$SORT_SEGMENT și DBA_TABLESPACES                                                                                                                     |
| Pentru a face o lista cu spațiu alocat utilizatorilor dintr-o baza de date Oracle în tablespace-urile permanente de sistem, se poate face o interogare pe:                                                                                                                 | DBA_TS_QUOTAS                                                                                                                                          |
| Pentru a face o lista cu utilizator, ID-ul sesiunii curente, starea ei și tablespace-ului permanent asociat utilizatorului din sesiunea curentă a unei baze de date Oracle, se pot folosi view-urile:                                                                      | V\$SESSION și USER_USERS                                                                                                                               |
| Pentru a mari cu o anumita dimensiune spațială de stocare intr-un fisier asignat unui tablespace Oracle, se poate folosi comanda:                                                                                                                                          | ALTER DATABASE                                                                                                                                         |
| Pentru a mari cu o anumita dimensiune spațială de stocare într-un tablespace Oracle cu opțiunea AUTOEXTEND ON, se poate folosi comanda:                                                                                                                                    | Ambele comenzi specificate la a și b pot fi folosite (ALTER DATABASE / ALTER TABLESPACE)                                                               |
| Pentru a putea executa comanda DROP DATABASE trebuie să ai privilegiul:                                                                                                                                                                                                    | SYSDBA                                                                                                                                                 |
| Pentru a specifica o extensie la spațiu de stocare al unui tablespace creat cu opțiunea EXTENT LOCAL MANAGEMENT intr-o baza de date Oracle, se poate folosi comanda:                                                                                                       | CREATE TABLESPACE                                                                                                                                      |
| Pentru a sterge fizic un fisier de date asignat la un tablespace permanent:                                                                                                                                                                                                | Se stergă tablespace-ul din dicționar cu opțiune de stergere a fisierului asignat ???                                                                  |
| Pentru a vedea dacă o baza de date Oracle a fost startata în modul arhiva a fisierelor de log, se execută o cerere SELECT pe:                                                                                                                                              | V\$DATABASE ??                                                                                                                                         |
| Pentru a vedea dacă un tablespace Oracle este autoextensibil se poate face o interogare pe:                                                                                                                                                                                | DBA_DATA_FILES                                                                                                                                         |
| Pentru a vedea din dicționar care este spațiu alocat unui utilizator Oracle, se poate face o interogare pe:                                                                                                                                                                | DBA_USERS sau DBA_TS_QUOTAS sau Ambele view-uri specificate la a și b contin această informație                                                        |
| Pentru a vedea din dicționar care este spațiu utilizat de către utilizator Oracle, se poate face o interogare pe:                                                                                                                                                          | DBA_TS_QUOTAS                                                                                                                                          |
| Pentru a vedea parola unui utilizator creat pe baza de date se poate face o interogare pe:                                                                                                                                                                                 | Informația nu se poate obține din baza de date                                                                                                         |
| Pentru a vedea structura tabelelor unei tabele Oracle se poate folosi următorul view din dicționarul bazei de date:                                                                                                                                                        | USER_TAB_COLUMNS                                                                                                                                       |
| Pentru administrarea automată a segmentelor de tip undo dintr-o baza de date Oracle, este necesară:                                                                                                                                                                        | Ambele condiții specificate la a și b sunt necesare (Existența unui tablespace de tip undo / Alocarea unui tablespace de tip undo la instanta curentă) |
| Pentru multimea de 2 tranzactii {(1, 2, 3, 4), (3, 4, 5)}, regula 1 → 3 are o incredere de:                                                                                                                                                                                | 50%                                                                                                                                                    |
| Pentru multimea de 2 tranzactii {(1, 2, 3, 4), (3, 4, 5)}, regula 3 → 1 are o incredere de:                                                                                                                                                                                | 50%                                                                                                                                                    |
| Pentru multimea de 2 tranzactii {(1, 2, 3, 4), (3, 4, 5)}, regula 3 → 4 are o incredere de:                                                                                                                                                                                | 100%                                                                                                                                                   |
| Pentru recunoașterea punctelor izolate (outliers) se folosesc valoarea $1.5 * IQR$ . În acest caz, cât ar trebui să fie X din multimea {0, 1, 2, X} pentru a fi considerat punct izolat:                                                                                   | 5,5                                                                                                                                                    |
| Pentru recunoașterea punctelor izolate (outliers) se folosesc valoarea $1.5 * IQR$ . În acest caz, cât ar trebui să fie X din multimea {0, 1, 2, X} pentru a fi considerat punct izolat:                                                                                   | 5,5                                                                                                                                                    |
| Pentru schimbarea setului de caractere al bazei de date trebuie să ai privilegiul:                                                                                                                                                                                         | SYSDBA                                                                                                                                                 |

|                                                                                                                                           |                                                                                                                                                                                                           |
|-------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Pentru setarea zonei de memorie utilizata pentru sortare in sesiunea curenta a unei baze de date Oracle, se foloseste comanda:            | ALTER SYSTEM                                                                                                                                                                                              |
| Pentru setarea zonei de memorie utilizata pentru sortare in sesiunea curenta a unei baze de date Oracle, se foloseste comanda:            | ALTER SYSTEM                                                                                                                                                                                              |
| Pentru unificarea spatiilor contigui dintr-un tablespace(defragmentare) al unei baze de date Oracle se foloseste comanda:                 | ALTER TABLESPACE                                                                                                                                                                                          |
| Pentru unificarea spatiilor contigui dintr-un tablespace(defragmentare) al unei baze de date Oracle se foloseste comanda:                 | ALTER SYSTEM                                                                                                                                                                                              |
| Pentru vizualizarea segmentelor de undo utilizate in sesiunea curenta se poate folosi view-ul:                                            | VSROLLNAME                                                                                                                                                                                                |
| Pot crea o baza de date daca esti:                                                                                                        | SYSDBA                                                                                                                                                                                                    |
| Prin comanda ALTER PROFILE se poate schimba:                                                                                              | Celelalte doua variante sunt false                                                                                                                                                                        |
| Prin intermediu profilului se pot restrictiona printre altele:                                                                            | Resurse referitoare la sistem si la parola                                                                                                                                                                |
| Prin profiluri se limiteaza printre altele:                                                                                               | Memoria folosita pentru acel user                                                                                                                                                                         |
| Printre tipurile de organizare pentru tabele se numara:                                                                                   | Tabele partitionate                                                                                                                                                                                       |
| Privilegiile sunt de doua tipuri:                                                                                                         | Obiect sau sistem                                                                                                                                                                                         |
| Privilegiul CREATE TABLE este un:                                                                                                         | Privilegiu sistem                                                                                                                                                                                         |
| Privilegiul RESTRICTED SESSION este inclus:                                                                                               | Si in SYSDBA si in SYSOPER                                                                                                                                                                                |
| Procesul Archiver (ARCN):                                                                                                                 | Copiera fisierelor Redo Log in arhiva de pe disc                                                                                                                                                          |
| Procesul Checkpoint (CKPT):                                                                                                               | Scrie periodic pe disc blocurile de date modificate                                                                                                                                                       |
| Procesul de background DBWR este folosit de o instanta Oracle pentru:                                                                     | Ambele variante prezentate la a si b sunt corecte (a. cteira si incarcarea in Database Buffer Cache a blocurilor de date din fisiere, b. Scrierea blocurilor de date din Database Buffer Cache in fisiere |
| Procesul de background LGWR este folosit de o instanta Oracle pentru:                                                                     | Scrierea datelor din Redo Log Buffer in fisierile de log                                                                                                                                                  |
| Procesul Log Writer scrie pe disc:                                                                                                        | Inaintea ca DBWRITER sa scrie pe disc                                                                                                                                                                     |
| Procesul prin care o linie este spartă în mai multe bucăți care sunt stocate în mai multe blocuri, se numește:                            | Înlantuirea liniilor (row chaining)                                                                                                                                                                       |
| Procesul prin care Oracle ia o linie dintr-un bloc și o mută în alt bloc, lasând în locul ei un pointer se numește:                       | Migrarea liniilor                                                                                                                                                                                         |
| Procesul server este:                                                                                                                     | Un proces care deservește unul sau mai multe procese client                                                                                                                                               |
| Procesul System Monitor (SMON):                                                                                                           | Este folosit pentru recuperarea după incident                                                                                                                                                             |
| Pot să vedea dacă o modul curent de lucru este RESTRICTED consultând:                                                                     | V\$INSTANCE                                                                                                                                                                                               |
| Rangul unei pagini reflectă importanța paginii ca:                                                                                        | Valoare relativă                                                                                                                                                                                          |
| Recuperarea care poate fi necesată la pornirea după un incident presupune:                                                                | Actualizarea fisierelor de Redo Log pe baza modificărilor din fisierile de date                                                                                                                           |
| Recuperarea după incident este necesată cand oprirea s-a facut in modul:                                                                  | Abort                                                                                                                                                                                                     |
| Redenumirea unui fisier de Redo Log presupune si executia cererii:                                                                        | ALTER DATABASE RENAME FILE                                                                                                                                                                                |
| Redenumirea fisierelor de date în care se stochează baza de date se poate face:                                                           | Cu baza de date montată                                                                                                                                                                                   |
| Redenumirea unui fisier de Redo Log presupune si executia cererii:                                                                        | ALTER DATABASE RENAME FILE                                                                                                                                                                                |
| Regula de clasificare a fluxurilor în FD și FDC spune că un flux este FDC și în cazul în care:                                            | Iese dintr-un element al multimii de stop                                                                                                                                                                 |
| Relatia dintre Data Warehouse (DW) și Data Mart (DM) este urmatoarea:                                                                     | DW include DM                                                                                                                                                                                             |
| Rezultatele intorcatorie de o cerere SELECT pe o tabela Oracle pot fi ordonate după ROWID:                                                | Da, cu clauza ORDER BY                                                                                                                                                                                    |
| Rolul Oracle DBA nu contine:                                                                                                              | Nici SYSDBA si nici SYSOPER<br>(A=sutece, B = bere)                                                                                                                                                       |
| S-a constatat că cei care cumpără A cumpără mai mult decât media B, unde:                                                                 | a si b sunt adevarate (a. online redo log files, b. archived redo log files)                                                                                                                              |
| Scrierile înregistrărilor de tip Redo Log se face în fisierele:                                                                           | Comenzi pentru crearea tabelelor si view-urilor din dicționarul bazei de date                                                                                                                             |
| Scriptul catalog.sql folosit la crearea unei baze de date Oracle conține:                                                                 | Nu                                                                                                                                                                                                        |
| Se poate asocia cote pentru utilizatori pe tablespace-urile temporare?                                                                    | Nu                                                                                                                                                                                                        |
| Se recomanda ca fisierile de control (daca sunt mai multe) sa fie:                                                                        | Nu există recomandări legate de plasarea acestor fisiere, CRED                                                                                                                                            |
| Se recomanda ca fisierile de Redo Log sa fie plasate:                                                                                     | Pe dispozitive diferite?                                                                                                                                                                                  |
| Se recomanda ca membrii unui grup de REDO LOG sa fie:                                                                                     | Pe dispozitive (discuri) diferite                                                                                                                                                                         |
| Segmentele continând tabele se pot plasa în tablespace-uri de tip:                                                                        | In orice tip de tablespace                                                                                                                                                                                |
| Segmentele de tip Undo dintr-o baza de date Oracle, sunt folosite cand se executa:                                                        | Operatiile de modificare și stergere de date în baza de date                                                                                                                                              |
| Segmentele temporale ale utilizatorilor sunt stocate toate:                                                                               | Pot fi în tablespace-uri diferite pentru utilizatori diferiți?                                                                                                                                            |
| Segmentele temporale se pot stoca:                                                                                                        | In orice tablespace                                                                                                                                                                                       |
| Spatialul alocat pentru zona de memorie Database Buffers din sesiunea curentă, se poate vedea printr-o interogare pe:                     | VSSGA                                                                                                                                                                                                     |
| Specificarea unei extensiuni spațiale de stocare al unui tablespace permanent creat într-o baza de date Oracle, se poate face cu comanda: | Ambele comenzi specificate la a si b pot fi utilizate (ALTER TABLESPACE ... ADD DATAFILE ... / ALTER DATABASE                                                                                             |
| Specificarea unei extensiuni spațiale de stocare al unui tablespace permanent creat într-o baza de date Oracle, se poate face cu comanda: | Ambele comenzi specificate la a si b pot fi utilizate (ALTER TABLESPACE ... ADD DATAFILE ... / ALTER DATABASE                                                                                             |
| SQL*Plus va genera:                                                                                                                       | Un proces user                                                                                                                                                                                            |
| Stergerea unui fisier de control în Oracle se face:                                                                                       | Cu instantă opritura                                                                                                                                                                                      |
| Stergerea unui fisier de control se face :                                                                                                | Cu instantă opritura                                                                                                                                                                                      |
| Stergerea unui segment se face cu o cerere de tip:                                                                                        | Nici unul din celelalte două răspunsuri nu este corect (CLEAR SEGMENT SI ALTER DATABASE)                                                                                                                  |
| Stergerea unui tablespace se face cu o cerere de tip:                                                                                     | DROP TABLESPACE                                                                                                                                                                                           |
| Stergerea unui user Oracle din dicționar cu opțiunea CASCADE , are ca efect:                                                              | USER_TAB_COLUMNS                                                                                                                                                                                          |
| Structura tabelelor a unei tabele se poate vedea în dicționarul bazei de date Oracle în :                                                 | Minimul și maximul                                                                                                                                                                                        |
| Sumarul de 5 numeri (Five number summary) conține printre altele:                                                                         | Supor (X U Y)                                                                                                                                                                                             |
| Suporul unei reguli de asociere X → Y este dat de formula:                                                                                | Permanent                                                                                                                                                                                                 |
| Tabelele create de un user într-o baza de date Oracle pot fi create într-un tablespace:                                                   | Unelele Oracle                                                                                                                                                                                            |
| Tablespace-ul SYSAUX conține date utilizate de:                                                                                           | Bootstrap Aggregating                                                                                                                                                                                     |
| Termenul Bagging provine de la:                                                                                                           |                                                                                                                                                                                                           |

|                                                                                                                          |                                                                                                                                           |
|--------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| Tranzactii: {(2, 3, 5), (2, 3, 6), (1, 4, 6)} ; s = 50%. Atunci articolele frecvente sunt:                               | {2, 3, 6}                                                                                                                                 |
| Tranzactii: {(1, 2, 3), (2, 3, 4), (3, 4, 5)} ; s = 50%. Atunci numarul de articole frecvente este:                      | 3                                                                                                                                         |
| Tranzactii: {(1, 2, 3), (2, 3, 4), (3, 4, 5)} ; s = 50%. Atunci numarul de perechi frecvente este:                       | 3                                                                                                                                         |
| Tranzactii: {(1, 2, 3), (2, 3, 4), (3, 4, 5)} ; s = 50%. Atunci numarul de perechi frecvente este:                       | 3                                                                                                                                         |
| Tranzactii: {(1, 2, 3, 5), (2, 3, 4, 5), (3, 4, 5)} ; s = 50%. Suportul regulii {3} -> {5} este:                         | 100%                                                                                                                                      |
| Tranzactii: {(1, 2, 3, 5), (2, 3, 4, 5), (3, 4, 5)}. Suportul regulii {3} → {5} este:                                    | 100%                                                                                                                                      |
| Tranzactii: {(1, 2, 3, 5), (2, 3, 5), (1, 4, 6)} ; s = 50%. Atunci articolele frecvente sunt:                            | {1, 2, 3, 5}                                                                                                                              |
| Tranzactii: {(2, 3, 5), (2, 3, 6), (1, 4, 6)} ; s = 50%. Atunci articolele frecvente sunt:                               | {2, 3, 6}                                                                                                                                 |
| Trunchierea unei tabele Oracle are ca efect:                                                                             | Stergerea datelor si a indecsilor creati pe tabela                                                                                        |
| Un arbore de decizie poate fi convertit in:                                                                              | Un set de reguli                                                                                                                          |
| Un bloc de date Oracle poate stoca:                                                                                      | Mai multe liniile dintr-o tabela                                                                                                          |
| Un bloc din Database Buffer poate contine la un moment dat:                                                              | Unul sau mai multe blocuri ale bazei de date ?                                                                                            |
| Un clasificator este construit pe baza unui:                                                                             | Set de antrenament                                                                                                                        |
| Un fisier de date poate sa apartina                                                                                      | Unul singur tablespace                                                                                                                    |
| Un hint introdus intr-o cerere SQL are forma:                                                                            | /*+ ... */                                                                                                                                |
| Un index de tip arbore creat pe o coloana a unei tabele Oracle:                                                          | Creeaza o tabela de index si optional se poate crea unicite pe coloana specificata in index                                               |
| Un nou fisier de date se poate adauga folosind comanda:                                                                  | ALTER TABLESPACE ...                                                                                                                      |
| Un obiect (tabela, index, etc) care are ca si corespondent:                                                              | Un segment                                                                                                                                |
| Un privilegiu de inserare pe o coloana a unei tabele, primiti de catre un user, poate fi vazut in dictionar in view-ul : | DBA_COL_PRIVS                                                                                                                             |
| Un proces server are la dispozitie o zonă de memorie exclusiva numita:                                                   | PGA                                                                                                                                       |
| Un proces server lucrează impreună cu:                                                                                   | Şi cu procesul user și cu serverul Oracle                                                                                                 |
| Un proces server se conecteaza intotdeauna cu:                                                                           | Unul sau mai multe procese user                                                                                                           |
| Un rol este:                                                                                                             | O colectie de privilegii                                                                                                                  |
| Un server Oracle este compus din:                                                                                        | O instanta Oracle si o baza de date                                                                                                       |
| Un tablespace creat intr-o baza de date Oracle cu optiunea AUTOALLOCATE poate fi extins de catre:                        | Sistemul de gestiune a bazei de date                                                                                                      |
| Un tablespace de tip UNDO dintr-o baza de date Oracle poate contine:                                                     | Segmente de tip undo                                                                                                                      |
| Un tablespace de tip UNDO dintr-o baza de date Oracle poate contine:                                                     | Segmente de tip undo                                                                                                                      |
| Un tablespace de tip UNDO dintr-o baza de date Oracle poate contine:                                                     | Segmente de tip undo                                                                                                                      |
| Un tablespace de tip UNDO dintr-o baza de date Oracle poate contine:                                                     | Segmente de tip undo ???                                                                                                                  |
| Un tablespace de tip unde se poate crea intr-o baza de date Oracle cu comanda:                                           | Ambele comenzi specificate la a si b pot fi folosite, in functie de context (CREATE DATABASE / CREATE UNDO TABLESPACE)                    |
| Un tablespace permanent dintr-o baza de date Oracle, se poate seta cu autoextensie cu comanda:                           | Ambele comenzi specificate la a si b pot fi folosite (ALTER DATABASE / ALTER TABLESPACE)                                                  |
| Un tablespace permanent dintr-o baza de date Oracle, se poate seta cu autoextensie cu comanda:                           | Ambele comenzi specificate la a si b pot fi folosite (ALTER DATABASE / ALTER TABLESPACE)                                                  |
| Un tablespace permanent dintr-o baza de date Oracle, se poate seta cu autoextensie:                                      | Ambele variante specificate la a si b pot fi folosite (Cand se creeaza baza de date sau tablespace-ul / Dupa ce se creeaza tablespace-ul) |
| Un tablespace poate deveni OFFLINE:                                                                                      | Atat la crearea cat si dupa aceea                                                                                                         |
| Un tablespace poate deveni READ ONLY:                                                                                    | Dupa ce a fost creat READ-WRITE                                                                                                           |
| Un tablespace poate fi de unul din tipurile:                                                                             | Permanent, Temporar, Undo                                                                                                                 |
| Un tablespace temporar dintr-o baza de date Oracle poate stoca:                                                          | Segmente temporare                                                                                                                        |
| Un user Oracle interacționează direct cu:                                                                                | Un proces user                                                                                                                            |
| Un user Oracle se poate conecta la baza de date folosind:                                                                | Ambele metode (Parola de acces specificata la crearea userului / Parola de sistem de operare specificata ca optiune la crearea            |
| Un user poate avea in același timp:                                                                                      | Mai multe sesiuni deschise cu același server Oracle                                                                                       |
| Un utilizator poate avea deschise simultan doua sau mai multe ferestre SQL*Plus pe aceeasi masina?                       | Da                                                                                                                                        |
| Userul SYS este:                                                                                                         | Proprietarul tabelelor din dictionarul bazei de date                                                                                      |
| Userul SYSEM este:                                                                                                       | Proprietarul altor tabele decat cele din dictionarul de date                                                                              |
| Valoarea curenta (FALSE) a lui RESOURCE_LIMIT se poate modifica prin:                                                    | ALTER SYSTEM SET RESOURCE_LIMIT=TRUE                                                                                                      |
| Valoarea de autoritate a unei pagini reflecta aceasta proprietate ca:                                                    | Valoare relativă                                                                                                                          |
| Valoarea de index a unei pagini reflecta aceasta proprietate ca:                                                         | Valoare relativă                                                                                                                          |
| Valoarea implicita a lui PCTFREE este:                                                                                   | 10                                                                                                                                        |
| Valoarea implicita a lui PCTINCREASE este:                                                                               | 50                                                                                                                                        |
| Valoarea implicita a lui PCTUSED este:                                                                                   | 40                                                                                                                                        |
| Valoarea implicita pentru INITRANS in cazul indecsilor este:                                                             | 2                                                                                                                                         |
| Valoarea implicita pentru INITRANS in cazul tabelelor este:                                                              | 1                                                                                                                                         |
| Valoarea implicita pentru MAXTRANS in cazul tabelelor este:                                                              | 255                                                                                                                                       |
| Varianta (Dispersia) este:                                                                                               | Pătratul abaterii standard sigma                                                                                                          |
| Varianta (Dispersia) multimi {0, 1, 2, 3, 4} este:                                                                       | 2                                                                                                                                         |
| Varianta (Dispersia) multimi {20, 0, 20, 0, 20, 0} este:                                                                 | 100?                                                                                                                                      |
| Vedere DBA_FREE_SPACE are o cheie primara:                                                                               | Numerica                                                                                                                                  |
| Vedere VSCONTROLFILE poate fi folosita dupa pasul:                                                                       | Montare BD?                                                                                                                               |
| Vedere V\$LOGFILE contine:                                                                                               | Informatii despre fisierile curente de tip REDO LOG                                                                                       |
| Vedere V\$PARAMETER este disponibila dupa:                                                                               | Pornirea instantei                                                                                                                        |

Vedereea V\$PARAMETER este populata cu informatii citite din:

Vedereea V\$TABLESPACE are o cheie primara:

Vederile dinamice pentru performante se pot consulta:

Vizualizarea grupurilor si membrilor fisierelor de log dintr-o baza de date Oracle se face cu comanda:

Vizualizarea grupurilor si membrilor fisierelor de log dintr-o baza de date Oracle se face cu comanda:

Vizualizarea tablespace-urilor create intr-o baza de date Oracle se face cu comanda:

Fisierele de control (de tip SPFILE)

Numerica

Unele dintre ele dupa pornirea instantei

SELECT group#, member FROM v\$logfile;

SELECT group#, member FROM v\$logfile;

SELECT tablespace\_name FROM dba\_tablespaces;