**Ce este analiza componentelor principale.Evidentiati 5 categorii de pb care pot fi solutionate cu ajutorul tehnicilor de analiza a componentelor principale.**

ACP repr.procedura algoritmica de analiza multidimensionala ce consta in determinarea unor noi variabile, numite componente principale, definite sub forma combinatiilor liniare de variabile initiale cu varianta maxima.5 categorii de probleme sol.cu ACP:a)reducerea complexitatii datelor.b)evidentierea si fixarea patternului asocierilor dintre variabile.c)det.variab.latente care se afla in spatele variab.masurate.

**Interpretati logica analizei componentelor principale(inclusiv d.p.d.v geometric).**

D.p.d.v geometric,variabilele numite componente principale definesc un nou spatiu al obiectelor. Logica analizei componentelor principale se bazeaza pe ideea fundamentala ca se pot face anumite transformari asupra observatiilor initiale,care sa determine maximizarea variantei individuale pt.anumite variabile si minimizarea variantei pt.alte variabile.

**Definiti componentele principale si mentionati proprietatile acestora:**

Componentele principale sunt variabile vectoriale abstracte,definite sub forma unor combinatii liniare de variabile initiale si care au urmatoarele 2 proprietati fundamentale: a)sunt necorelate 2 cate 2 si suma patratelor coeficientilor care definesc combinatia liniara ce corespunde unei component principale este egala cu unitatea;

**Formulati modelul matematic al ACP,definiti si interpretati marimile definitorii ale acestora.**

Opt ф (x,w)

AεMn\*k , unde criteriul de optim poate fi de maxim sau minim, in functie de natura functiei ф

SR: w=At \* t

Daca functia ф este de tip distanta,atunci criteriul de optim va fi reprezentat de minimizarea functiei ф. Daca functia ф este o masura a cantitatii de informative adusa de noua modelitate de repr. a obiectelor,criteriul de optim va fi repr. de max. functiei ф.

**Ilustrati modul de deducere a componentelor principale.**

Determinarea coeficientilor combinatiei liniare ce defineste comp.principala w,in conditiile maximizarii variantei acestei comp.principale,este echivalenta cu a alege dintre cele n valori proprii ale mattricii de covarianta ∑ pe cea mai mare si a determina componentele vectorului de ponderi α ce defineste respective comp.principala prin calculul vectorului propriu al matricii ∑ asociat cu acea valoare proprie.

**Definiti si justificati 3 dintre componentele principale**

Componentele principale sunt vectori ortogonali care preiau cât mai mult din varianta variabilelor vector originale astfel: prima componenta principala preia maximul posibil din varianta variabilelor originale, a doua componenta principala preia maximul din varianta ramasa dupa ce este eliminata prima componenta

a)sunt coordinate obiectilor(generate prin proiectii pe axele noului spatiu)sunt denumite scoruri principale si au proprietatea ca pastreaza configuratia initiala a acestora.

b) componentele principale conserva variant totala a variab initiale.

c) dependent de unitatile de masura.

**Interpretati vectorii si valorile proprii ale matricii de covarianta**

Vectorii proprii ai matricei de covarianta pot fi utilizati pentru construirea unei transformari unitare. In urma utilizarii acestei transformari, in noul spatiu de trasaturi rezultat, matricea de covarianta va fi una diagonala. Valorile proprii exprima varianta explicata prin fiecare noua componenta principala si corelatia dintre variabilele vechi.

**Ce sunt scorurile principale si cum se determina?**

Coordonatele obiectelor în noul spatiu, adica proiectiile obiectelor pe axele acestuia, sunt evaluarile obiectelor

în raport cu noile variabile si se numesc *scoruri principale .* Daca vom presupune ca au fost retinute p componente principale si daca vom nota cu U matricea de dimensiunea n\*p , ale carei coloane sunt cei p vectori proprii care defines cele p component principale, atunci matricea scorurilor poate fi determinat astfel :W(p\*t)=A(p\*n)\* x(n\*t).Liniile matricii W reprezinta scorurile corespunzatoare noilor variabile sau observatiile celor p componente principale.

**Ce este matricea factor ?**

matricea factor arată corelația dintre variabilele inițiale și componentele principale. Pe baza acestei matrici și a reprezentărilor sale grafice se poate determina informația cuprinsă de fiecare componentă principala. Matricea factor este f.importanta in acp deoarece permite identificarea unor submultimi ale variabilelor originale care se grupeaza in jurul unei component principale permintand identificarea unui component real pt aceste constructii abstracte.

**Detaliati modul in care pot fi interpretate componentele principale.**

Numarul de componente principale este egal cu numarul grupelor de variabile care sunt foarte puternic corelate între ele.

Analiza componentelor principale este o metoda de *reexprimare* a variabilelor originale sub forma unui numar mai mic de noi variabile, numite componente principale, care sunt combinatii liniare de varianta maxima ale variabilelor originale.

Cu ajutorul componentelor principale se poate defini o structura a dependentei dintre variabilele originale mai simpla si mai clara, deci mai usor de interpretat.

Fiecare component principala ofera informatii despre marimea cantatii de informatie preluata din variabilele originale.

**Criterii de alegere a nr de componente principale:**

1) Criteriul empiric:se determina matricea de covarianta, in matricea de covarianta se identifica grupuri de var. puternic corelate intre ele si ale caror componente sunt slab corelate cu componentele altor grupuri.

2) Criteriul procesului de acoperire: potrivit acestui criteriu, nr de comp princ este dat de un procent mai mare de 70-75%):

3)Criteriul povarnisurilor/grohotisurilor: nr de componente principale e obtinut prin reprezentarea grafica a valorilor proprii atasate matricii de covarianta initiale.

4)Criteriul lui Kaiser: nr de componente principale e dat de valorile porprii mai mari sau egale cu 1.

5)Criteriul statistic: este cel mai riguros.

**Analiza factoriala (definitii+tipuri de probleme ce pot fi rezolvate cu aceasta):**

Def: O analiza de tip multidimensional are drept obiectiv evidentierea corelatiilor existente la nivelul unei multimi de indicatori cu ajutorul unui nr redus de factori necorelati si ordonati, in fct de variabilitatea impusa de ei. Tipuri de probleme: detectarea structurii in relatiile dintre variabile; reducerea numarului variabilelor; rezolvarea multicoliniaritatii in regresia multipla; validarea constructiei unei scale compuse din mai multi itemi;

**Structura generala a modelului de analiza factoriala:**

Modelul analizei factoriale are la baza doua ipoteze fundamentale: prima ipoteza se refera la presupunerea ca nivelul sau valorile unui ansamblu de variabile aleatoare X1,X2,...,Xn se formeaza ca rezultat exclusiv al influentei a 3 categorii de factori:

O multime formata din p factori comuni, f1,f2,...,fp, a caror influenta se considera a se exercita asupra fiecareia dintre cele n variabile considerate

O multime formata din n factori unici, U1,U2,...,Un, a caror influenta se considera a se exercita in mod individual.

**Definiti si interpretati descompunerea variabilitatii in contextul analizei factoriale:**

analiza factoriala isi propune sa reexprime variabilitatea continuta in spatiul initial intr-o maniera diferentiata in functie de rolul pe care il au in formarea acesteia factorii comuni, pe deoparte factorii unici pe de alta parte factorii Prin utilizarea tehnicilor de analiza multidimensionala variabilitatea spatiului cauzal n dimensional determinat de multimea variabilelor indicator x1,x2,…xn este conservata intr-o proportie mai mare sau mai mica prin intermediul variabilitatii induse de un numar mai redus de factori abstracti f1,f2…fp , p<n care sunt factori comuni. Impreuna cu factorul unic acesti factori determina un spatiu p+1 dimensional numit spatiu test ‘u’.

**Ce sunt scorurile factor, cum se calculeaza si cum se interpreteaza acestea:**

O anumita observatie, corespunzatoare unui factor dat, este determinata sub forma unui scor corespunzator respectivului factor, scor format pe baza contributiei variabilelor originale. urmatoarea relatie:

fi=bi1x1+bi2x2+...+binxn

unde bij reprezinta coeficientii scorurilor Sub forma matriciala, aceasta relatie poate fi scrisa astfel: f=F’x

**Metode de estimare a modelului factorial**

AF=analiza de tip multidimen.: AF exploratorie-are drept obiectiv identif unei struct a dependentelor cauzale existente la niv unei mult. de var.indicatori si generarea fact.comuni. AF confirmatorie- modalit de utilizare a AF in care se urmareste verif unor ip. formulate in legatura cu o struct cauzala deja identificata.

**Recun formelor +exemple.**

TRF reprez ansamblul met,teh, reg de analiza si decizie care au ca scop identif apartenentei obiectelor (entitati informationale) in cadrul unor clase cu individualitate bine definite. TRF este un domeniu multidisciplinar: th sistemelor, th deciz, th.probabilit

**Principalele concept ale recun formelor**

Ipoteza fundamental care se face in cadrul TRF este aceea ca la niv spatiului formelor exista un anumit grad de eterogenitate ceea ce implica posibilitatea existentei claselor. Insa la nivelul aceleasi clase se considera ca obiectele sunt caracterizate printr-un anumit grad de omogenitate. Intre clasele identificate avem astfel urmatorele caract: similaritate si disimilaritate. In TRF, urmatorele notiuni sunt echivalente: clasa=grupa=categ=cluster.

**Prob gen a clasificarii.**

Consta in det. apartenentei unui obiect din Ω ={O1,..On} pop de obiecte, la una din clasele latente existente in pop initiala.1)caract. obiectelor incluse in proc. de clasif. pot fi exprimate sub forma unor var cantitative sau calitative, masurate pe una din cele 4 scale de masurare cunoscute:nominala,ordinala,de interval,de raport.

**Sistemele de recun controlata si necontrolata**

1.SRC=acele sist in cadrul carora nr de clase este cun iar formele analizate sunt repartizate intr-una din aceste clase. Setul de forme a caror apartenenta la clasa este cun se mai numeste set de referinta/set de prototipuri/set de formare si reprez un esantion extras din pop analizata format din ob a caror apartenenta la clase este cun si care sunt fol in proc de clasif pt det reg de decizie ce det alocarea ob pe clase.

**Ce este analiza cluster, care sunt conceptele fundamentale ale acesteia si care sunt domeniile utilizarii ei**

***Analiza cluster*** poate fi definita ca reprezentand o multime de principii, metode si algoritmi de clasificare, avand ca scop organizarea datelor sub forma unor structuri informationale relevante.

efectuarea unei clasificari cu ajutorul analizei cluster consta în obtinerea unor *solutii cluster* sau a unor *partitii*, reprezentate de o multime de clase sau clustere, care verifica proprietatile mentionate anterior.

utilizarea cea mai frecventa marketing, în investigatiile psihosociale sau în evaluarile economico-sociale la nivel teritorial.

**Def scopurile analizei cluster si descrieti tipul informatiilor utilizate in analiza cluster**

scopul general al analizelor de tip cluster este acela de creare a asa-numitelor *taxonomii* sau *tipologii*.

Constructia tipologiilor este bazata pe analiza *asemanarilor s*i *deosebirilor* existente între obiectele unei multimi date.

Identificarea pe o mare cantitate de informatii brute a unor categorii, clase sau grupe informationale reprezinta deasemenea unul dintre scopurile generale si, în acelasi timp, principale ale oricarei analize cluster.

În general, masurarea gradului de proximitate dintre obiecte se face cu ajutorul a doua grupe de indicatori, cunoscute sub numele de *indicatori de similaritate s*i *indicatori de disimilaritate.*

**Definiti analiza cluster si aratati cum se clasifica metodele de analiza cluster**

***Analiza cluster*** o multime de principii, metode si algoritmi de clasificare, avand ca scop organizarea datelor sub forma unor structuri informationale semnificative, relevante.

*metode de tip ierarhic s*i *metode de tip iterativ* sau *de partitionare*.

***Analiza cluster de tip ierarhic* sau *arborescent***este o metoda de clasificare bazata pe gruparea obiectelor pe baza de *agregare succesiva* în clase din ce în ce mai largi de obiecte sau de *dezagregare succesiva* în clase din ce în ce mai mici.

***Algoritmi de partitionare***

Algoritmii de partitionare includ o serie de metode de analiza cluster, cu mult mai performante decât metodele de clasificare ierarhica. Dintre cei mai importanti algoritmi de partitionare, mentinam: algoritmul celor *K*-medii si algoritmul celor *K*-medoizi.

**Definiti conceptual de distanta si descrieti cateva modalitati de evaluare a distantelor dintre forme**

Distanta dintre doua clustere este, de fapt, o distanta dintre doua *multimi* de puncte, adica o distanta mai dificil de evaluat.

*Metoda centroidului* evalueaza distanta dintre doua clustere ca *distanta între centroizii* celor doua clustere.

*Metoda lui Ward* evalueaza distanta dintre doua clustere *suma totaa| a patratelor abaterilor* la nivelul

configuratiei cluster rezultate din comasarea celor doua clustere pentru care se evalueaza distanta.

**Formulati criteriul general al clasificarii si aratati cum se evalueaza variab inter si intra clasa (Cazul uni-demns)**

**Criteriu general de clasificare**:Clasificarea obiectelor în clase se face în asa fel încât sa se asigure o *variabilitate minima în interiorul claselor s*i o *variabilitate maxima între clase.*

Metoda Ward este metoda de evaluare a distantelor dintre clustere care urmareste maximizarea gradului de omogenitate din interiorul clusterelor, ceea ce e echivalent cu minimizarea variabilitatii intraclusterelor. Minimizarea variabilitatii intracluster conduce automat la maximinizarea variabilitatii intercluster.

O problema dificila care apare în analiza cluster, este legata de *necesitatea evaluarii distantelor dintre clase* sau *clustere*. Dificultatea acestei probleme este data de faptul ca distantele dintre clase sau clustere sunt, de fapt, distante între *multimi de obiecte* sau distante între *multimi de variabile*.

**Formulati criteriul general al clasificarii si aratati cum se evalueaza variab inter si intra clasa (Cazul n-dimens)**

**Criteriu general de clasificare**: Clasificarea obiectelor în clase se face în asa fel încât sa se asigure o variabilitateminima în interiorul claselor si o variabilitate maxima între clase.

**Analiza cluster de tip ierarhic**

Analiza cluster de tip ierarhic este o metoda de clasificare bazata pe gruparea obiectelor pe baza de

agregare succesiva în clase din ce în ce mai largi de obiecte sau de dezagregare succesiva în clase din ce în ce mai mici.

Ipoteza de baza a clasificarii ierarhice:

Rezultatul utilizarii analizei cluster de tip ierarhic îl reprezinta o multime de structuri particulare de clustere, numita arbore ierarhic.

Algoritmii de clasificare • algoritmi de clasificare prin agregare, amalgamare sau combinare;

• algoritmi de clasificare prin dezagregare sau divizare.

**Metoda agregarii simple de analiza cluster**

Reprezinta metoda de clasificare ierarhica de tip ascendant care in fiecare etapa a procesului de clasificare comaseaza 2 dintre clusterele pentru care distanta intre cei mai apropiati vecini e minima in comparative cu alte perechi de clustere.

**Metoda agregarii complete de analiza cluster**

Este de tip ascendant, care in fiecare etapa a clasificarii comaseaza 2 clustere pentru care distanta dintre cei mai indepartati vecini din acele clustere. Clusterizarea de acest tip se mai numeste si analiza cluster de distanta maxima sau analiza cluster de tip MAX.

**Metoda centroidului de analiza cluster**

**Defini**t**ie**: Metoda centroidului evalueaza distanta dintre doua clustere ca distanta între centroizii celor doua clustere.

Evaluarea distantei dintre doua clustere cu ajutorul metodei centroidului se face calculând mai întâi centroizii celor doua clustere, dupa care se evalueaza distanta dintre clustere ca distanta între acesti centroizi.

**Metoda lui Ward**

Este o metoda de evaluare a distantei dintre doua clustere, care se bazeaza pe maximizarea gradului de omogenitate a clusterelor. De regula, gradul de omogenitate a unui cluster se considera a fi cu atât mai mare, cu cât suma totala a patratelor abaterilor intracluster este mai mica.

Elementul caracteristic al metodei lui Ward este reprezentat de faptul ca prin comasarea a doua clustere se urmareste obtinerea unei omogenitati maxime la nivelul tuturor clusterelor care apartin unei configuratii date a obiectelor pe clustere

**K-means**

Algoritmul K-means este o metodă de determinare a clusterelor pe care le formează mai multe pattern-uri. Procedura este una de instruire nesupervizată. Se presupune cunoscut numărul K al clusterelor. Fiecare cluster are un centroid. Algoritmul lucrează cu K clustere, deci K dintre punctele folosite la instruire vor fi centriozii celor K clustere. Întrucât initializarea centroizilor se face aleator, există posibilitatea ca mai multe rulări ale algoritmului să conducă la rezultate diferite.

**Dendrograma** Ca rezultat al algoritmului se obţine arborele de clasificare (dendrograma).

Prin secţionarea orizontală a dendrogramei se obţine o partiţie a mulţimii elementelor clasificate. Componentele partiţiei sunt clasele căutate.

Pe axa orizontală sunt elementele iniţiale (ordinea este cea care permite desenarea arborelui). Pe axa verticală sunt distanţele dintre obiecte, de exemplu, între obiectele 4 şi 6 este o distanţă egală cu 4.

**Cum se alege numarul de clustere in cazul clasificarilor de tip ierarhic?**

*Algoritmii* sau *metodele de tip ierarhic* au ca scop producerea *mai multor* soluţii cluster, soluţii numite *ierarhii cluster*.

Cea mai sintetică soluţie a unei structuri cluster obţinute cu ajutorul metodelor de clasificare ierarhică este formată *dintr-un singur cluster*, care include toate obiectele analizate. Cea mai detaliată soluţie a unei structuri cluster de acest fel include unnumăr maxim de clustere, *egal cu numărul de obiecte analizate*, fiecare cluster conţinând un singur obiect.

**Ce sunt clasificatorii de tip liniar?**

Un clasificator liniar este o combinaţie liniară de trăsături, care sunt componentele unui vector x. Clasificatorul poate fi scris ca:

g(x) = wT x + w0 = ∑wi xi + w0

Unde w este vectorul de ponderi, w0 este deplasamentul (bias), sau ponderea pragului.

**Formulati problema generala a recunoasterii supervizate a formelor si mentionati cateva domenii de activitate.**

În mod frecvent,în analiza datelor apare necesitatea studierii unor populaţii care sunt eterogene d.p.d v. al caracteristicilor analizate. Expresia cea mai semnificativă a populaţilor de tip eterogen este întâlnită în special în domeniul statisticii, econometriei şi analizei datelor,fiind repr. chiar de cantităţile foarte mari de informaţie care trebuie prelucrată,sintetizată şi interpretată.În cazul cercetării unor populaţii de acest tip,pentru ca rezultatele investigării să capete consistenţă şi relevanţă,este necesară o împărţire, o divizare a acestor populaţii în subpopulaţii cu un anumit grad de omogenitate,urmând ca analizele şi procesul de modelare implicate în studierea respectivei populaţii să se facă în mod diferenţiat, pentru fiecare subpopulaţie în parte.

**Definiti functiile discriminant liniare,variabilele discriminant si scorurile discriminant.**

***Functiile discriminant*** (***Fisher***) sunt combinaţii liniare de variabilele descriptor, de forma:

D(x)=B0+Bt\*x, unde x este vectorul variabilelor descriptor, iar B este vector propriu al matricii.

***Variabilele discriminant*** sunt combinaţii liniare de variabilele descriptor, de forma:

d=B0+BT\*x, unde x si B au semnificaŢia din definiţia precedentă.

Scorurile discriminant sunt utilizate ca *indicatori* în clasificarea propriu-zisă a obiectelor.

**Descrieti clasif. Bayesian si aratati cum poate fi utilizat in predictia apartenentei formelor.**

Principala probl care trebuie rezolvata in cadrul recunoasterii supervizate a formelor este aceea a determinarii criteriilor (regulilor) de clasificare astfel incat noi forme a caror apartenenta nu este cunoscuta sa poata fi incadrate in clase in care acestea se grupeaza in mod natural. Determinarea clasificatorilor au o natura algoritmica, numindu-se proces de formare a clasificatorului.

**Descrieti clasificatorul liniar fisher si aratati cum poate fi utilizat acesta in predictia apartenetei formelor.**

Potrivit acestor metode de solutionare ale probl de clasificare supervizata,multimea initiala de obiecte Ω este impartita in ω1, ω2,.. ωk pe baza comparatiei dintre variabilitatea intraclasa si variabilitatea interclasa, inregistrate in cadrul claselor din populatia analizata.Analiza discriminanta pp. det. functiei discriminant care sa asigure separabilitatea cat mai buna a claselor si fiecare clasa sa fie caract. printr-un grad de omogenitate cat mai mare, echivalent cu variab intraclasa cat mai mica.

**Descrieti modul de stabilire a abilitatii predictive a unui clasificator si matricea corectitudinii clasificarii.**

Daca vom considera un esantion format din T obiecte,care apartin claselor ω1, ω2,... Ωk,atunci va rezulta o matricea de clasificare.Un element al matricii de clasif.arata nr. de obiecte apartinând în mod real clasei ωi care, prin utiliz.

tehnicilor de recunoastere a formelor, sunt clasif. în clasa ωj.

Populaţia sau colectivitatea generală este reprezentată de mulţimea tuturor măsurătorilor efective sau conceptuale care prezintă interes pentru cercetător sau experimentator.

Eşantionul reprezintă o submulţime de măsurători selectate dintr-o populaţie, o submulţime a populaţiei statistice supusă investigaţiei ştiinţifice.

Variabila reprezintă o abstractizare a mulţimii de valori posibile pe care le poate înregistra o caracteristică a unui anumit fenomen. După natura pe care o au, variabilele pot fi de două tipuri: variabile calitative şi variabile cantitative.

**Scala nominală** este o scală non-metrică, pe baza căreia valorile variabilelor sunt definite prin intermediul simbolurilor nenumerice. Măsurarea variabilelor pe scala nominală este echivalentă cu procesul de codificare a variabilelor.

Scala ordinală este o scală non-metrică, prin intermediul căreia valorilor posibile ale caracteristicilor li se atribuie numere de ordine sau ranguri, în funcţie de poziţia acestor valori într-o ierarhie

Scala raport este o scală metrică, prin intermediul căreia valorilor posibile pe care le pot lua caracteristicile măsurate li se atribuie numere definite in raport cu o origine prestabilită.

**Tendinta centrala** poate fi evidentiata prin intermediul: *media, mediana* si *modulul*. ***Media*** - se obţine împărţind suma valorilor individuale la efectivul populaţiei sau al eşantionului. ***Mediana*** - este valoarea care, în cadrul seriei statistice, separă efectivul populaţiei în două părţi egale. ***Modulu****l* - se utilizează numai cînd se lucrează cu frecvenţe, fiind valoarea luată cu cea mai mare frecvenţă.

***Varianta simpla*** este o masura pt devierea de la medie, deviatie standard fata de medie. **Vs=Σi=1n  (xi- xmediu)2** . **Varianta totala** masoara variabilitateace caracterizeaza observatiile unei multimi de variabile si se defineste ca suma a variantelor individuale ale variabilelor: **Vt=ΣSi2**

**Varianta generalizata** masoara variabilitatea ce carac observatiile multimii de variabile, atat din pct de vedere individual, cat si din pct de vedere al simultanietatii, al interaxtivitatii informationale variabilele.

**Covarianta** este o masura a variatiei simultane a doua variabile, ea fiind, in valoare absoluta, cu atat mai mare cu cat valorile absolute ale variatiilor celor doua variabile in jurul mediei sunt mai apropiate ca magnitudine,.

**Definiti conceptul de distanta si descrieti cateva modalitati de evaluare a distantelor dintre forme**

*Indicatorii de disimilaritate* sunt marimi numerice care exprima cât de *deosebite* sau cât de *departate* sunt doua obiecte sau doua variabile. Indicatorii de disimilaritate se mai numesc si indicatori sau coeficienti de *deosebire* sau de *distantare* a obiectelor sau variabilelor.

***Distan****t****a Euclidian****a* se calculeaza ca *radacina patrata a sumei patratelor diferentelor coordonatelor* celor doua obiecte sau variabile pentru care se evalueaza distansa.

**Distanta *Manhattan***, numita si distanta *rectangulara*, se calculeaza ca suma a valorilor absolute ale diferentelor coordonatelor celor doua obiecte sau celor doua variabile analizate

***Distanta Cebî****s****ev***Deste o distanta de tip *valoare absoluta* si se determina ca fiind valoarea maxima a valorilor absolute ale diferentelor dintre coordonatele obiectelor sauvariabilelor.***Distanta Mahalanobis***

Distanta *Mahalanobis* este una dintre cele mai cunoscute, mai importante si mai frecvent utilizate distante. Ea este o forma generalizata a conceptului de distanta.