Statistik 2 2010, 3.-7.5.2010

Stansens PC-klass ASA-huset.

Material del III

Klusteranalys och klassificering

Klusteranalys används generellt för att upptäcka dolda grupper av data där observationerna liknar varandra mer än vad de liknar observationer hos andra grupper. Det finns ett enormt forskningsfält kring detta både inom statistik och datalogi, typiskt inom maskininlärning. Här betraktar vi två vanliga metoder för klusteranalys: K-means och hierarkisk klustring som finns tillgängliga i SPSS. K-means algoritmen skapar K grupper av n datavektorer så att skillnaderna mellan grupperna maximeras och skillnaderna inom grupperna minimeras. K måste anges i analysen och man kan givetvis utföra flera analyser för att avgöra ett lämpligt värde på K. Det bör noteras att det även finns 100-tals metoder för modellbaserad klustring där metoden skall automatiskt lära sig ett lämpligt värde på K på basen av datat. Dylika metoder kräver vanligtvis skräddarsydd programvara och många finns implementerade på R. Hierarkisk klustring skapar ej direkt gruppering av data, utan ett indexerat träd som kallar dendrogram. Dendrogrammet visar närhetsstruktur hos datavektorer och kan användas för att urskilja mönster i data. Genom att skära dendrogrammet på en viss nivå, erhåller man en gruppering av data i likhet med K-means metoden.

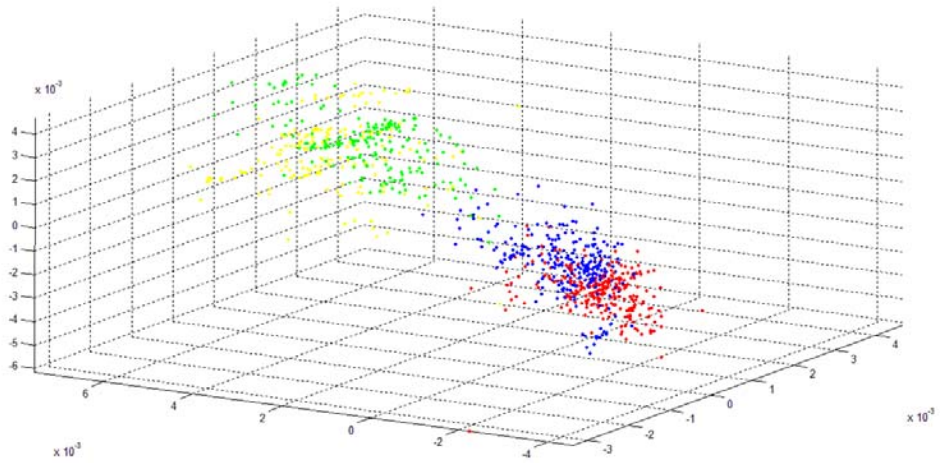
Öppna filen brainData.sav. Den innehåller normaliserade expressionsnivåer hos 1000 gener för 115 sampel av människohjärnvävnad som är tagna genast efter att individerna avlidit. Det finns 3 grupper av individer: kontroll, schizofreni och man-depressiv. Vi klustrar datat med K-means och hierarkisk klustring för att se om skillnader m.a.p. generna är kopplade till diagnosen. Även Diskriminantanalys kommer att betraktas.

Flerdimensionell skalning

Flerdimensionell skalning eller MDS (multidimensional scaling) är ett alternativ till PCA (principalkomponentanalys) ifall man har som syfte att upptäcka mönster i hödimensionellt data, t ex avslöja dolda undergrupper bland observationerna. Metoden beskrivs relativt utförligt här:

http://www.mathpsyc.uni-bonn.de/doc/delbeke/delbeke.htm

Nedan finns en 3D bild av MDS resultat för ca 1300 bakterieisolat p.b.a. sju MLST geners sammansatta variation (totalt ca 3250 nukleotider lång DNA sekvens). MDS uppvisar en tydlig separation mellan två grupper (gul/grön mot röd/blå). Färgerna står för tillhörighet gällande olika delpopulationer av bakterier för varje isolat.



Faktoranalys

Faktoranalys beskrivs i nötskal här:

http://faculty.chass.ncsu.edu/garson/PA765/factor.htm

Se även beskrivning av Strukturella EkvationsModeller (SEM):
http://faculty.chass.ncsu.edu/garson/PA765/structur.htm

Läs in filen Uscrime.sav. Filen innehåller uppgifter om diverse brottsfrekvenser i olika delstater i USA. Sammanlagt finns uppgifter om följande sju brottstyperna: murder, rape, robbery, assault, burglary, larceny, autotheft.

Vi skall använda explorativ faktoranalys för att studera datat. Notera skillnaden mellan explorativ och konfirmatorisk faktoranalys – i den senare specificerar man modellstrukturen i förväg och testar dess lämpligthet för ett visst datamaterial. I explorativ faktoranalys försöker man urskilja strukturer i data med hjälp av faktormodeller som identifieras som lämpliga. Ofta utnyttjar man Scree-plot från principalkomponentanalys som ett första steg, för att avgöra hur stort antal faktorer kan vara lämpligt för datamaterialet.

Välj dimension reduction – factor analysis, metoden principal factor analysis och välj sedan antalet faktorer lika med 2. Vi bestämmer att inte använda rotering av lösningen.

Titta på skattningar av *communality* för varje variabel. Ett högt värde på detta indikerar att de allmänna faktorerna förklarar största delen av variationen för variabeln i fråga. Som en indikation på modellens anpassningsgrad, kan man betrakta resultatet från ett Chi^2-test, där nollhypotesen motsvarar den specificerade faktormodellen och mothypotesen att inga begränsningar läggs på kovariansmatrisen för variablerna. Ett signifikant testresultat (på nivån α) innebär att modellens struktur inte stöds av datat. Det bör dock observeras

att med väldigt stora datamaterial erhålls signifikanta resultat i regel för alla tänkbara restriktioner på kovariansmatrisen. Datat här i uppgiften är förhållandevis litet.

Titta sedan på faktorladdningarna (Factor loadings) för variablerna. Går det att urskilja något tydligt mönster? Ett intressant mönster i faktoranalyssammanhang är att delgrupp av variabler får relativt höga laddningsvärden på en viss faktor, medan alla de andra variablerna får relativt låga laddningsvärden på just denna faktor.

För att förenkla tolkningen av faktorladdningarna, upprepas nu samma analys som ovan, men med en viss rotering av faktorerna enligt sk Varimax-metoden. Upprepa stegen ovan, men använd Factor rotation – Varimax. Titta på skattningarna och laddningarna, går det nu att urskilja ett tydligare mönster bland laddningarna? Vi ser att Assault och Murder får höga laddningar på en faktor, samt att Burglary, Larcery och Autotheft får höga värden på den andra faktorn. Dessa faktorer motsvarar alltså olika dimensioner m.a.p. variation i datat, t ex den senare faktorn kunde tolkas representera egendomsbrott.

Notera att i Varimax metoden tvingas faktorerna att vara oberoende av varandra, dvs. de är ortogonala i geometrisk mening. I en sk oblique rotering, tillåts faktorerna att vara korrelerade. Det finns möjlighet att erhålla en oblique rotering med hjälp av metoden Promax. Upprepa analysen en gång med Promax metoden och jämför faktorladdningarna. Bildas här samma mönster som med Varimax roteringen?

I faktoranalys kan man skatta de icke-observerade värdena på faktorerna för varje ”individ” i datamaterialet. Detta görs genom att spara Factor scores, då man valt faktoranalys från menysystemet. Upprepa de två faktoranalyserna ovan och välj att erhålla faktorpoängen (spara dem med olika namn i den andra analysen, så att värden inte skrivs över). Rita upp faktorpoängen med hjälp av Graphs-Scatterplot. Notera hur deras relation skiljer sig mellan Varimax (oberoende faktorer) och Promax (korrelerade faktorer) roteringarna.

Öppna datafilen AidsPatienter.sav Filen innehåller uppgifter om hur aids-patienter har betraktat sina vårdande läkarei en frågeformulär med 13 olika frågor. Vi undersöker bl a hur sekventiellt test av faktormodeller kan leda till valet av antalet faktorer. Anpassa faktormodellen utan rotering med 1,2,3 och 4 faktorer och notera p-värden för Chi^2-testet. Man ser att modellen med tre faktorer blir nätt och jämt acceptabel på 5%-nivån. Använd sedan tre faktorer och anpassa modellen både med Varimax och Promax roteringen. Titta på laddningsmönstren. Det kan konstateras att inget enkelt struktur kan lätt hittas för datat med hjälp av faktoranalysen, eftersom flera variabler laddas relativt högt på fler än en faktor och för att somliga variabler laddas lågt på samtliga faktorer.