Statistik 2, del 6:

Principalkomponentanalys

Principalkomponentanalys används till mycket inom data-analys, men dess huvudsakliga syfte är oftast att hitta/urskilja intressanta mönster i flerdimensionella data. En allmän beskrivning med länkar till vidare material finns på http://en.wikipedia.org/wiki/Principal_component_analysis.

Öppna datafilen SedlarData.sav och gör en principalkomponentanalys (Dimension reduction/Factor). Sätt först antalet komponenter till 6, och sedan till 3 och 2. Kom ihåg att spara komponenterna. Gör ett spridningsdiagram för datat utifrån den erhållna principalkomponenterna och använd variabeln Grupp som markör. Kan du urskilja de två grupperna?
Filen innehåller sammanlagt 200 observationer på 6 olika typer av mätvärden för äkta och förfalskade sedlar. Den första variabeln indikerar om en sedel är äkta (värdet = 1) eller förfalskad (värdet = 2). Scree plot visar hur mycket varje komponent förklarar av variationen i data.

I följande övning betraktar vi genetiska data där varje case är en individ (97 totalt) samplats från en av fyra geografiska regioner (GenetiskData.sav) Den första kolumnen i datamatrisen indikerar regionen och var och en av de övriga kolumnerna (85 stycken) indikerar om individen bär en viss mutation i arvsmassan. Försök utföra principalkomponentanalys som ovan med samtliga mutationsindikatorer (V2-V86). Vad händer? Det finns fem indikatorer som är konstanta i datamaterialet (V20, V55, V63, V70, V71). Tag bort dem från listan av variabler som sätts in i principalkomponentanalysen och utför sedan analysen (Scree plot & scores). Gör ett 3D spridningsdiagram med 3 principalkomponenter och undersök om det verkar finnas skillnader i genetiska profiler mellan regionerna.

I följande artikeln av Novembre et al i Nature hittar man ett exempel på dylik användning av principalkomponentanalys i ett genetiskt sammanhang, där man har ett stort antal samplade individer (ca 3,000) och variabler (ca 500,000):

http://www.nature.com/nature/journal/v456/n7218/full/nature07331.html

Klusteranalys

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.