Statistik 2 2008, 29.4.08

Övning 5a

Hämta följande datafil: http://www.abo.fi/fak/mnf/mate/jc/statistik2/mathmarksClustering.txt
Filen innehåller tentapoäng för 88 mattestuderande m.a.p. 5 olika kurser som representerar olika områden i matematik.
Starta R och sedan R commander. Vi försöker nu med hjälp av en klusteranalys upptäcka om datamaterialet innehåller ”dolda” delgrupper av observationer. Välj Statistics-Dimensional...-Cluster analysis-k-means clustering. Sätt #clusters lika med 2 och kryssa för Assign clusters to the data set. Ange namnet Kmeans2 för assignment variabeln. Programmet skapar nu en ny variabel som indikerar klustertillhörighet för varje observation. K-means-algoritmen försöker hitta en optimal uppdelning av datapunkter till ett givet antal delgrupper (cluster). Man bör dock komma ihåg att k-means algoritmen och andra dylika klustringsmetoder kan lätt missa vettiga grupperingar av data, då datastrukturen är komplicerad (relativt stort antal kluster).

Givet att vi fått en klustring, kan vi titta på hur fördelningarna ser ut för tentapoängen i olika kluster. Välj Graphs-Boxplot och t ex variabel V1 och sedan Plot by groups (här väljer man den korrekta klustringsindikatorn). Vi ser att studerande i det ena klustret tenderar ha lägre tentapoäng. Samma mönster upptäcks för de övriga variablerna om man ritar låddiagram på motsvarande sätt.

För att se på de samtliga variablerna och deras relationer på en gång, givet klustringen, välj Graphs-Scatteplot matrix, sätt in variablerna (V1-V5) och Plot by groups. Ta bort krysset från Smooth lines. Hurdana mönster upptäcker man i datat?

Upprepa nu k-means analysen med 3 cluster istället för 2 (kom ihåg att namnge klustringsindikatorn Kmeans3 för att skilja åt resultaten). Hur ser resultaten ut jämfört med de tidigare erhållna?

Övning 5b

Fortsätt med datat i föregående exempel. Vi gör nu en hierarkisk klusteranalys, dvs. välj Statistics-Dimensional...-Cluster analysis-Hierarchical... Ange variablerna V1-V5, välj metoden Single Linkage, samt ett lämpligt namn åt klustringen, t ex sätt förkortningen SL efter Hclust i namnrutan. Titta på dendrogrammet som ritas upp av programmet, kan man urskilja tydliga delgrupper i datat? Betrakta närmare på den hierarkiska klustringen genom Statistics-Dimensional...-Cluster...-Summarize... Välj 3 kluster och tryck på Ok. Antalet observationer, samt medelvärden på variabler visas nu klustervis i output-fönstret. Klusterindikatorn kan läggas till datat genom Statistics-Dimensional...-Cluster...-Add hierarchical... Välj 3 kluster och ange ett lämpligt namn till indikatorn i rutan Assigned cluster label.
Upprepa den hierarkiska klusteranalysen ovan med metoderna Ward och Complete Linkage och jämför resultaten med det tidigare erhållna. Kom ihåg att ange vettiga namn åt klustringslösningarna för skilja åt dem! Blir det några skillnader? Rita upp datat med Graphs-Scatterplot matrix och använd klusterindikatorn från analysen med Ward-metoden som gruppering (Plot by groups). Hur ser klustren ut?