Lista nº 7 – Análise de Conglomerados e de Discriminante

1. Considere o conjunto de dados do arquivo *crime.xls*. As variáveis tratadas são taxas de criminalidade (em cada 100.000 habitantes) em vários estados dos Estados Unidos no ano de 1980. Foram avaliadas nove categorias de crime: violência, crime contra propriedade, assassinato, estupro, roubo, assalto, roubo de residência, furto e roubo de veículo.
   .
   a) Faça uma análise de conglomerado por um método aglomerativo à sua escolha (justifique a sua escolha). Indique qual, em sua opinião, seria o número de grupos mais provável da partição dos estados no que se refere às taxas de criminalidade analisadas. Dê os grupos formados e faça um gráfico para visualizar os agrupamentos dos elementos.
   b) Faça uma análise de conglomerado pelo método das k-médias com estes dados considerando. Compare a solução encontrada com aquela do item (a). Verifique a estabilidade de sua solução final com relação à escolha das sementes iniciais
   c) Faça análise de conglomerados agrupando variáveis e indique quantos grupos de variáveis teríamos
      i. Em cada grupo que você construiu escolha no máximo duas variáveis para representar o grupo.
      ii. Se você tivesse que escolher um número de variáveis menor que nove para agrupar os estados, quantas variáveis você sugeriria? Quais seriam essas variáveis?
   d) Faça um escalonamento multidimensional com estes dados, considerando q = 2. Compare a solução encontrada com aquela do item (a).
2. Considere o conjunto de dados apresentado no exercício 11.30, pág. 661, Johnson e Wichern, 2007. Os dados estão no arquivo *petroleo.xls*.
   a) Analise a normalidade individual e conjunta das variáveis envolvidas na análise.
   b) Faça uma análise discriminante considerando as 5 variáveis do problema. Nessa análise você deverá o método stepwise para encontrar as variáveis mais importantes na discriminação. Caso nem todas as 5 variáveis sejam significativas, você deverá fazer também uma análise discriminante com as variáveis significativas do problema. Você acha que a função discriminante construída com as variáveis significativas está conseguindo de fato discriminar as populações?
   c) Analise a qualidade da(s) função(ões) discriminante(s) construída(s) em (b) usando os métodos da ressubstituição, o método da validação cruzada e o método de colação de elementos "à parte".
   d) Use a regra de classificação construída em (b) para classificar uma amostra que tem os seguintes valores: $X_1 = 3{,}7$; $X_2 = 34{,}0$; $X_3 = 0{,}3$; $X_4 = 6{,}33$; $X_5 = 4{,}00$.
   e) Dê sua opinião sobre a qualidade geral da(s) função(ões) discriminante(s) que foi(ram) construída(s)
   f) Faça uma análise de conglomerados para estes dados usando as 5 variáveis explicativas e o método das K-médias, usando k = 3.Analise os grupos que foram formados e a concordância destes grupos com as respectivas populações a que os respectivos elementos amostrais pertencem, ou seja, você acredita que o método das k-médias conseguiu identificar corretamente as 3 populações das quais os dados foram coletados? Justifique.
   g) Se você tivesse que indicar um dos dois métodos para resolver esse problema de discriminação, você optaria pelas funções construídas em (b) ou pelo método usado em (f)?
   h) Faça um escalonamento multidimensional com os dados considerando q = 3. Compare a solução encontrada com aquela do item (b).

Exercícios do Johnson e Wichern (2007).

3. 11.1, pág. 650
4. 11.7, pág. 651
5. 12.5, pág. 748.
6. 12.7, pág. 748.
7. 12.16, pág. 750.