Introdução a Modelagem e Aprendizado
ELTD1 - DAELN UTFPR-CT

Prof. Heitor S. Lopes (2024) - Exercício #4

Objetivo: utilizar métodos de agrupamentos para a análise de dados

Parte 1: Experimento duplo-cego

a. Em experimentos clínicos do tipo “duplo-cego”, uma pessoa coleta e anonimiza os dados. Depois, uma outra pessoa faz a análise, de modo que não exista nenhuma comunicação entre elas. Isto evita a “contaminação” dos dados colhidos por algum tipo de atitude, mesmo que involuntária, dos envolvidos na coleta e análise dos dados. Neste caso em particular, o objetivo final é identificar a qual do grupo pertence cada um dos 300 pacientes dos quais os dados foram coletados em experimentos.

b. Utilize o *dataset* fisioterapia.csv que contém 6 variáveis físicas mensuradas dos pacientes. A única informação disponível é que há pacientes com problemas músculo-esqueléticos e pacientes saudáveis (grupo controle). Devido à natureza do estudo, não se sabe quais são os pacientes saudáveis e quais são os que têm problemas músculo-esqueléticos. Além disto, também não se sabe quantos problemas diferentes podem ter nesta amostra e quais variáveis são afetadas pelos problemas físicos.

c. Utilizando dois métodos de agrupamento de dados (*K-means* e hierárquico) descubra quantos grupos (concisos) distintos há nos dados coletados, identificando qual dos grupos é, possivelmente, o grupo controle (saudável) e quais são os grupos de pacientes com enfermidades. Para o algoritmo *K-means*, ajuste o valor de *k* para os melhores valores do coeficiente de silhoueta. Observe que não necessariamente o máximo valor do coeficiente vai levar ao melhor agrupamento de dados. Use também o algoritmo de agrupameto hierárquico e ajuste seus parâmetros (Min, Max, Avg, Ward) para encontrar partições nos dados que tenham sentido. IMPORTANTE: o objetivo final é determinar quais são pacientes são saudáveis e quais têm cada uma das enfermidades diferentes.

Parte 2: Caracterização de clientes de shopping

a. O objetivo deste estudo é caracterizar em grupos os clientes que frequentam um determinado shopping de Curitiba, de modo a identificar *clusters* de clientes. Desta maneira, entendendo as características dos grupos pode-se potencializar as vendas com campanhas de *marketing* dirigidas a cada grupo. Cada cliente tem uma identificação, informação de gênero, idade, renda anual e gastos realizados no shopping em um determinado período (os dois últimos dados estão em uma escala arbitrária).

b. Utilize o dataset ClientesShopping.csv e aplique os métodos de agrupamento, basicamente *K-means* e hierárquico. Tenha em mente que é necessário separar os clientes em grupos **que tenham significado**. O objetivo da análise dos dados, obviamente, é aumentar as vendas através de estratégias de marketing focadas em grupos específicos, **claramente categorizados** (p.ex. pessoas com alta renda, mas com gastos medianos/baixos). Deve-se caracterizar todos os grupos encontrados. Mostrar os parâmetros utilizados nos algoritmos e os resultados com gráficos comentados.