Lista 1

César A. Galvão - 190011572

Gabriela Carneiro - 180120816

João Vitor Vasconcelos - 170126064

Índice

Questão 1

Escolha uma área de pesquisa de interesse (engenharia, medicina, economia, ecologia, computação ou outra área de interesse). Para cada tipo de problema da lista abaixo, apresente um artigo publicado em revista indexada e indique as características do estudo que o fazem relacionar o artigo ao problema em questão. Indique pontos fortes e fracos de sua formação em estatística para realizar estudos semelhantes.

Análise Estatística Não Paramétrica e Reconhecimento de Padrões

O artigo de Pitombo e Costa (2015) propõe a utilização de uma abordagem híbrida de técnicas não paramétricas e paramétricas para a previsão de escolha modal. Tomando como foco a análise não paramétrica, o estudo aborda a técnica de Árvore de Decisão a qual tem como objetivo subdividir o banco de dados em um número finito de classes. De um modo geral esse modelo parte de um nó inicial de classes e a partir de uma variação do algoritmo CART (do inglês, Classification and Regression Tree), subdivide o banco de dados em subconjuntos cada vez mais homogêneos em relação a variável resposta a partir de divisões binárias.O particionamento dos dados se faz a partir da minimização do desvio em todas as divisões permitidas nos nós da árvore. Além de essa técnica ter sido utilizada para reconhecer padrões dentro do banco de dados, também foi usada como forma de facilitar a discretização de variáveis independentes.

A base, ou pelo menos a ideia inicial da técnica de árvore de decisões, foi abordada em Análise Multivariada, como análise de cluster e a utilização de discretização e categorização, criação de variáveis *dummy* e análise de variáveis categóricas na matéria de Dados categorizados. O curso deu embasamento para entender a forma geral dessas técnicas, tendo como deficiência a falta de melhor aprofundamento em análise de algoritmos, que poderia ser melhor abordado na matéria de Estatística computacional. A ementa atual dadisciplina, mesmo abordando algoritmos, não foi o bastante para o aprendizado de algoritmos e problemas de otimização.

Aprendizado Estatístico Supervisionado e Análise Estatística Paramétrica

O estudo apresentado por Armstrong e Sloan (1989) se relaciona com a análise estatística paramétrica e o aprendizado estatístico supervisionado de várias maneiras. Primeiramente, o estudo adota suposições paramétricas ao utilizar modelos de regressão logística, o que implica na aceitação de uma forma paramétrica para esses modelos, incluindo a suposição de distribuições específicas dos dados. Destaca-se que a análise estatística paramétrica também é evidente na estimação dos parâmetros dos modelos, como o modelo de odds cumulativas e o modelo logit, utilizados para analisar dados de resposta categóricas.

Além disso, o estudo emprega técnicas de aprendizado estatístico supervisionado para lidar com dados epidemiológicos. Esse tipo de aprendizado envolve o treinamento de um modelo com dados rotulados, onde a variável resposta é conhecida, permitindo a previsão de novos

dados. No contexto do estudo, isso inclui o ajuste dos modelos de regressão logística aos dados e a avaliação de seu desempenho.

Pontos fortes da formação em estatística para realizar estudos semelhantes incluem a capacidade de compreender e aplicar modelos estatísticos paramétricos, como a regressão logística, para analisar dados complexos, como os epidemiológicos discutidos no artigo. Além disso, a formação em estatística proporciona conhecimento em técnicas de aprendizado supervisionado, fundamentais para prever resultados com base em dados rotulados.

Por outro lado, uma possível lacuna a ser destacada é a falta de disciplinas mais práticas na formação em estatística, que poderiam aprimorar as habilidades aprendidas na teoria e oferecer uma experiência mais aplicada na análise de dados reais.

Aprendizado de Máquinas e/ou Estatístico não Supervisionado

Em seu artigo, Anderlucci, Montanari, e Viroli (2019) contribuições publicadas em revistas de estatística de alto prestígio entre os anos de 1970 e 2015 com o propósito de propor uma "taxonomia" dinâmica dos principais tópicos desenvolvidos. Foram considerados clusters de assuntos a cada década, com a possibilidade de divisão e fusão entre grupos, bem como surgimento de novos grupos e extinção de outros. Os grupos são compostos de artigos que possuem similaridade de assunto. Diversas estratégias estatísticas e de informação são adotadas para lidar com a sparsidade das matrizes envolvidas e estimação da distribuições necessárias para adequadamente lidar com as características dos dados e compreender a heterogeneidade inerente a cada agrupamento gerado.

Enquanto de fato foram explorados nas disciplinas do curso técnicas de análise multivariada, uma visão geral de técnicas disponíveis para análise de dados com as características listadas, assim como teoria da informação ainda é muito superficial, se é que foram abordadas. Outro assunto que é tangencial são modelos dinâmicos, que não são abordados no curso. Mesmo assim, são tratadas as ferramentas fundamentais para compreensão e pesquisa do conhecimento probabilístico necessário para uma engenharia reversa do estudo.

Questão 2

Considere um hipercubo de dimensão r e lados de comprimento $2A$. Dentro deste hipercubo temos uma hiperesfera r-dimensional de raio A. Encontre a proporção do volume do hipercubo que está fora da hiperesfera e mostre que a proporção tende a 1 a medida que a dimensão r cresce. Escreva um programa R para verificar o resultado encontrado. O que este resultado significa?

Resolução

O volume de uma hiperesfera r-dimensional de raio A no espaço Euclideano é definido por

$$V_r(A) = \frac{\pi^{r/2}}{\Gamma\left(\frac{r}{2}+1\right)} A^r, \tag{1}$$

em que $\Gamma(.)$ é a função gama. Por sua vez, o volume de um hipercubo de dimensão r e lados de comprimento $2A$ é dado por

$$C_r(2A) = (2A)^r. \tag{2}$$

Assim, a proporção do volume do hipercubo que está fora da hiperesfera é dada por

$$F_c = \frac{C_r(2A) - V_r(A)}{C_r(2A)} = \frac{(2A)^r - \frac{\pi^{r/2}}{\Gamma(\frac{r}{2}+1)} A^r}{(2A)^r} = 1 - \frac{\pi^{r/2}}{2^r\,\Gamma\left(\frac{r}{2}+1\right)}. \tag{3}$$

Se tomamos o limite $r \longrightarrow +\infty$, temos que

$$\lim_{r \longrightarrow +\infty} \frac{\pi^{r/2}}{2^r\,\Gamma\left(\frac{r}{2}+1\right)} = 0, \tag{4}$$

pois a função do denominador é dominante. Portanto, a proporção do volume do hipercubo que está fora da hiperesfera tende a 1 a medida que a dimensão r cresce. Isso significa que a hiperesfera de raio A no hipercubo de dimensão r e lados de comprimento $2A$ se torna cada vez mais insignificante à medida que a dimensão do espaço cresce.

O programa a seguir ilustra o resultado:

```r
pacman::p_load(purrr,dplyr, ggplot2)

# Função que calcula a proporção do volume do hipercubo que está
# fora da hiperesfera

prop_volume <- function(r){
  return(1 - pi^(r/2)/(2^r * gamma(r/2 + 1)))
}

tabela_proporcao <- tibble(
dimensoes = c(1:20),
proporcao = map_vec(dimensoes, prop_volume)
)
```

Dimensões	Prop. vol. fora da hiperesfera
1	0.0000000
2	0.2146018
3	0.4764012
4	0.6915749
5	0.8355066
6	0.9192545
7	0.9630878
8	0.9841457
9	0.9935576
10	0.9975096
11	0.9990800
12	0.9996740
13	0.9998888
14	0.9999634
15	0.9999884
16	0.9999964
17	0.9999989
18	0.9999997
19	0.9999999
20	1.0000000

Tabela 1: Proporção do volume do hipercubo que está fora da hiperesfera em função da dimensão do espaço.

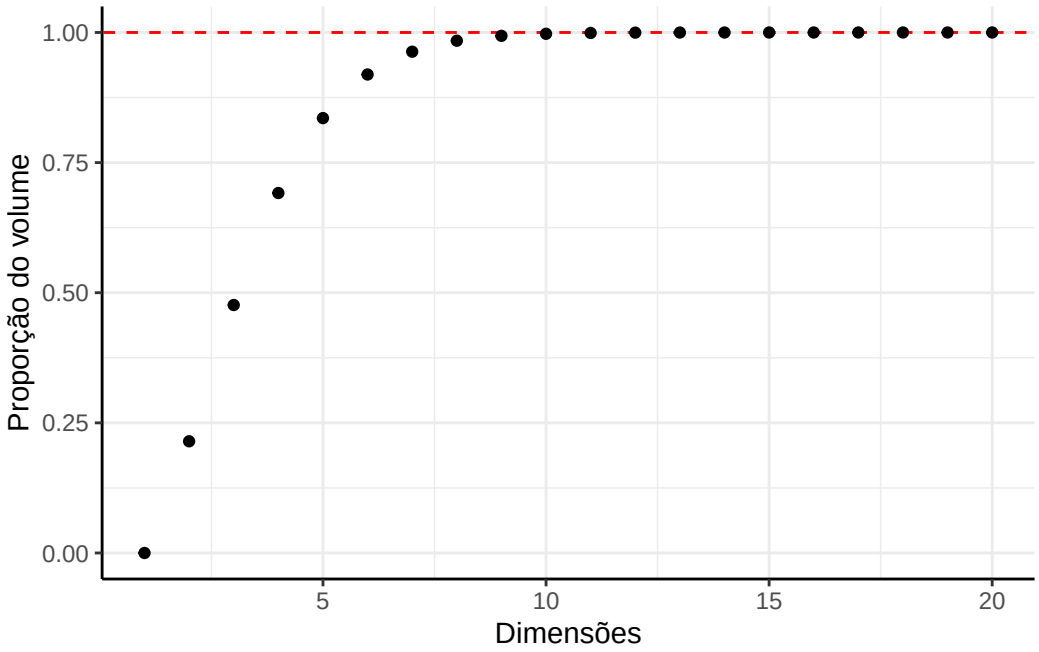

Figura 1: Proporção do volume do hipercubo que está fora da hiperesfera em função da dimensão do espaço.

Referências

Anderlucci, Laura, Angela Montanari, e Cinzia Viroli. 2019. «The Importance of Being Clustered: Uncluttering the Trends of Statistics from 1970 to 2015». *Statistical Science* 34 (2). https://doi.org/10.1214/18-sts686.

Armstrong, Ben G., e Margaret Sloan. 1989. «Ordinal Regression Models for Epidemiologic Data». *American Journal of Epidemiology* 129 (1): 191–204. https://doi.org/10.1093/oxfordjournals.aje.a115109.

Pitombo, Cira Souza, e Aline Schindler Gomes da Costa. 2015. «Aplicação conjunta de modelos não paramétricos e paramétricos para previsão de escolha modal». *Journal of Transport Literature* 9 (1): 30–34. https://doi.org/10.1590/2238-1031.jtl.v9n1a6.