



DEPARTAMENTO DE ESTATÍSTICA

10 abril 2023

Lista 1

Prof. Dr. George von Borries

Análise Multivariada 1

Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

1)

O Artigo Lozano (2016) cita que alguns estatísticos vêem a estatística mais como uma ferramenta multidisciplinar e que deve se ater à cálculos probabilísticos para inferência do que propriamente para ser resignificada como ciência de dados. Citando Friedman (1998), diz que mineração de dados (*data mining*) seria mais especificamente sobre encontrar padrões e correspondências nos dados, com foco em grandes bases de dados. Com citação à Dhar (2013), diz que a grande diferença entre estatística e ciência de dados é o volume da informação. Enquanto estatísticos estão acostumados com um volume de dados que caberia em uma calculadora de bolso, a ciência de dados trabalharia com volume de dado gigantesco, necessitando de grande poder computacional. Cita a limitação de algumas técnicas estatísticas, que tendem a serem sensíveis demais com um volume de dados muito grande (*upsampling problem*). Mas mais importante, cita a comum baixa capacidade de estatísticos de lidarem com dados não numéricos e/ou não estruturados, que compoem grande parte da massa de dados produzido ultimamente. Para este problema, a engenharia de dados é importante para estruturar estes dados e torná-los trabalháveis. O texto do Kiliç (2019) foca mais em detalhar e descrever as técnicas estatísticas utilizadas pela ciência de dados para construção de modelos (PCA, regressão, clusterização e afins), bem como a discussão sobre a aplicabilidade dessas técnicas em grandes bases de dados. Ao fim do artigo, os autores tentam traçar correlações entre estatística e ciência de dados. O último artigo Zola (2019), traz mais uma abordagem mercadológica dos termos e as necessidades do mercado quanto a qualificação dos profissionais que pretendem trabalhar em áreas relacionadas a dados.

Ciência de dados é grande volume de dados. Estatística tem as técnicas mais sofisticadas, mas por vezes anda devagar em relação as transformações da realidade do exponencial crescimento do poder computacional e das massas de dados. A engenharia de dados é a ciência que transforma um dado desestruturado em algo “trabalhável” com técnicas estatísticas numéricas, enquanto a mineração de dados foca em encontrar dentro de uma grande massa de dados, os fragmentos mais relevantes para uma análise.

A grande semelhança é que no fundo essas ciências são quase a mesma coisa e o profissional que exerce alguma dessas funções têm capacidade de compreender e até realizar as outras ciências correlatas; inclusive, num ambiente pequeno-corporativo, a realidade é que provavelmente será o mesmo profissional à executar todos esses serviços. Em Kiliç (2019), é citado que “O ponto em comum da estatística com a ciência de dados é aprender com dados, ou transformar dados em conhecimento” (Kuonen, 2004 apud Kiliç, 2019)

2) Escolha uma área de pesquisa de interesse. Pesquise artigos publicados em revista indexadas e descreva (resumidamente) um exemplo indicando o tipo de problema (ou problemas) entre os listados abaixo. Inclua referência bibliográfica e indique as características dos dados e estudo que relacionam ao tipo de problema (ou problemas) indicado.

a) Análise multivariada clássica ($n < p$).

Análises fatoriais e escalas psicométricas. “Com o uso dessa técnica, o investigador discerne os aspectos isolados da estrutura latente em uma matriz de dados e, logo depois, indica o quanto que cada variável é explicada por cada dimensão; ou seja, ele é capaz de descrever as correlações entre muitas variáveis observáveis em termos de um número menor de variáveis não observáveis (linearmente relacionadas com as variáveis originais observadas) denominadas dimensões ou variáveis latentes” (VASCONCELOS, Rafaela Balliana Bragança de. (2016) Apud Hair Jr. et al., 2009; Johnson e Wichern, 2007)

b) Mineração de dados (Data Mining) (n elevado).

Referências:

[1] VASCONCELOS, R. **Avaliação psicométrica de um instrumento para mensuração de autoeficácia**. Monografia (Bacharelado em Estatística) – Departamento de Estatística, Universidade de Brasília. Brasília, p. 21. 2016.