Lista de Exercícios 1

Introdução - Aprendizado Supervisionado e Não Supervisionado

João Victor de Oliveira Nogueira

15-04-2024

Questão realizada por João Victor de Oliveira Nogueira e Vinicius Martins Tostes.

1. (a) Análise Estatística Paramétrica

O artigo tem como objetivo ajustar modelos de predição lineares para estimar o número esperado de gols para times mandantes e visitantes, estimando pârametros de força de ataque e de defesa.

O artigo traz o ajuste de um modelo de Poisson para estimar os parâmetros de força de ataque e defesa dos times mandantes e visitantes. O modelo de Poisson é um modelo paramétrico e por isso o link com a pergunta da lista.

Durante o artigo é mostrado diversas tabelas com as estimativas e também é mostrado algumas probabilidades de vitória, empate e derrota para os times mandantes e visitantes além de probabilidades de resultados exatos.

Achei muito interrensante essa abordagem por ser feita com algo que eu gosto bastante além de que modelos de previsão de eventos é algo que pode ser usado de diversas maneiras no nosso ramo. O artigo também cita as diversas maneiras que se pode estimar os resultados, além de deixar claro que teriam modelos melhores e que foi usado um modelo simples que performou muito bem.

Também gostei da forma com que os dados foram analisados, mostrou como é importantes também conhescer sobre o assunto que está sendo tratado, no caso futebol deu para notar que os parâmetros defensivos de um time tendem ser mais importante do que os ofensivos quando se observa a colocação dos times no campeonato.

Fonte: **ESTIMATING RATINGS IN FOOTBALL: BRAZILIAN CHAMPIONSHIP 2017** (https://biometria.ufla.br/index.php/BBJ/article/view/403/253)

(b) Análise Estatística não Paramétrica

Neste estudo foi analisado analisado diferenças estatísticas de hipertensos nas Regiões de Saúde do estado do Rio de Janeiro do período de 2002 a 2012. Após a utilização do teste Shapiro-Wilk que rejeitou a hipótese de normalidade para sete das nove regiões estudadas, foi necessário utilizar métodos não paramétricos (Teste de Kruskal-Wallis e pós teste Dunn) para verificar se há diferenças estatísticas entre as regiões.

Fonte: **[APLICAÇÕES DE TESTES ESTATÍSTICOS NÃO PARAMÉTRICOS PARA ANÁLISE DE HIPERTENSOS NAS REGIÕES DE SAÚDE DO RIO DE JANEIRO] (www.e-publicacoes.uerj.br/cadest/article/view/67352/43926)**

(c) Reconhecimento de Padrões e (d) Aprendizado de Máquinas e/ou Estatístico Supervisionado

O artigo tem como objetivo aplicar métodos de aprendizado de máquina para classificar e analisar o desfecho de pacientes com COVID-19 que receberam alta ou óbito e descrever o perfil dos pacientes infectados pelo vírus.

Os dados utilizados tem variáreis de entrada, sendo: sexo, idade, tempo de internação, tipo de atendimento (ambulatorial, pronto atendimento, externo e interno) e unidade federativa a que pertence o paciente, e como desfecho a variável de saída que é a alta ou óbito do paciente.

O artigo faz a comparação de 4 algoritmos de classificação: Máquina de vetores de suporte (SVM), K-vizinhos mais próximos (KNN), Redes Naïve Bayes e Random florest. Todos esses algoristimos são não paramétricos e supervisionados e essas caracteristicas são as que linkam com a pergunta da lista.

Achei interessante a abordagem do artigo em comparar os algoritmos de classificação, os dados utilizados tinham uma grande diferença entre as duas possíveis saídas, o número de óbitos era de apenas 23 enquanto o de altas era de 3879. Isso pode ser um problema comum em futuras análises que um estatístico possa fazer, então a abordagem de comparar os modelos demonstra a importância de se escolher um modelo que se adeque melhor aos dados.

No final do artigo é feita uma análise dos resultados dos modelos e o modelo de Redes Naïve Bayes apresentaram um desempenho superior e capacidade de calcular indicadores de acurácia, como sensibilidade, especificidade e coeficiente kappa, os outros modelos não conseguiram acertar a classificação de nenhum óbito. Esses resultados são úteis caso em alguma ocasião futura eu tenha um caso parecido de classificação binária com um desbalanceamento grande entre as classes.

Fonte: **Classification and Analysis of Patients with COVID-19 Using Machine Learning** (https://biometria.ufla.br/index.php/BBJ/article/view/588/362)

(e) Aprendizado não supervisionado

O artigo utiliza técnica de aprendizado de máquina não-supervisionado para agrupar, compreender e detectar ataques distribuídos de negação de serviço, ou Distributed Denial of Service (DDoS). O método utilizado foi o k-means.

O autor fala sobre algumas limitações que sofreu em termos de hardware para realizar a análise exploratória dos dados e executar o algoritmo de aprendizado não supervisionado, devido o tamanho das bases de dados e mostra interesse em uma continuidade do trabalho focada em outros tipos de ataques e analisar melhor os agrupamentos gerados pelo estudo.

Fonte: **Análise exploratória da base de ataques DDoS com a utilização de aprendizado de máquina não-supervisionado** (https://repositorio.usp.br/directbitstream/d4bd7d27-6beb-45c2-90f7-f2e4b8781306/Rafael%20Henrique%20Quaresma%20.pdf)

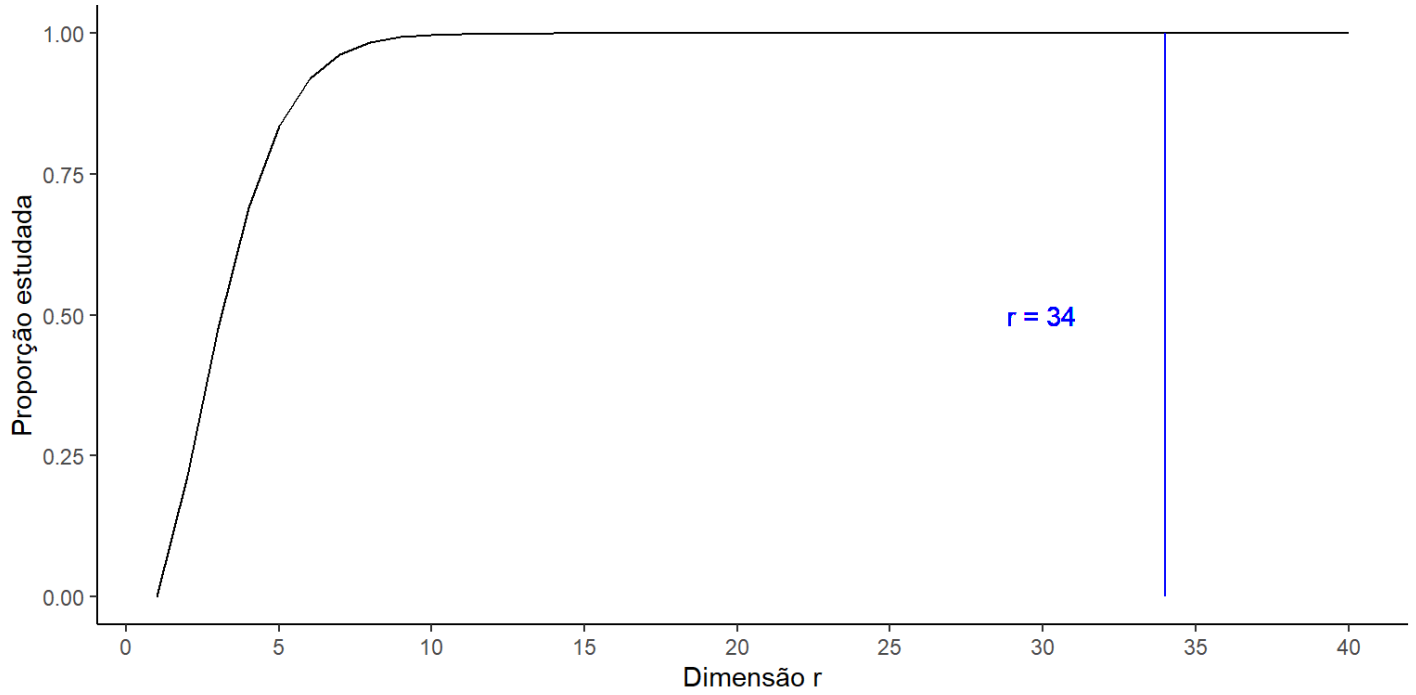
2. Considere um hipercubo de dimensão r e lados de comprimento 2A. Dentro deste hipercubo temos uma hiperesfera r-dimensional de raio A. Encontre a proporção do volume do hipercubo que está fora da hiperesfera e mostre que a proporção tende a 1 a medida que a dimensão r cresce. Escreva um programa R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

Primeiramente temos que o volume do hipercubo é dado por $V_{cubo} = (2A)^r$ e o volume da hiperesfera é dado por $V_{esfera} = \frac{\pi^{r/2}}{\Gamma(r/2+1)}A^r$. A proporção do volume do hipercubo que está fora da hiperesfera é dada por $P = 1 - \frac{V_{esfera}}{V_{cubo}}$. Irei calcular a proporção para r = 1, 2, …, 40 e verificar se a proporção tende a 1 a medida que a dimensão r cresce.

O volume do hipercubo segue $V_{cubo} = (2A)^r$, já a hiperesfera segue $V_{esfera} = \frac{\pi^{r/2}}{\Gamma(r/2+1)}A^r$. Para encontrar a proporação do volume do hipercubo que está fora da hipersfera é necessário seguir $P = 1 - \frac{V_{esfera}}{V_{cubo}}$.

```
r = 1:40
A = 1
vCubo = (2*A)^r
vEsfera = (pi^(r/2))/(gamma(r/2 + 1))*A^r
Prop = 1-vEsfera/vCubo
#
library(ggplot2)

data = data.frame(n = r, Prop = Prop)
ggplot(data, aes(x = r, y = Prop)) +
  geom_line() +
  geom_line(data = data.frame(x = c(34, 34),y = c(0, 1)), aes(x = x, y = y),
            color = "blue") +
  geom_text(aes(x = 34, y = 0.5,label = "r = 34            "),
            color = "blue", hjust = 1) +
  scale_x_continuous(breaks = seq(0, 40, 5)) +
  labs(x = "Dimensão r",
       y = "Proporção estudada") +
  theme_classic()
```



Com o crescimento da dimensão r é perceptível que a proporção estudada tende a 1, e a partir do r = 5 já tem o valor de 0,835 e cresce rapidamente para valores próximos de 1 e com r = 34 atinge o número 1. Este fenômeno ocorre, pois com o crescimento da dimensão R o hipercubo aumenta seu volume exponencialmente e a hiperesfera diminui seu volume.