



Cálculo Numérico – Prova 2

Projeto A – Disfonia em Pacientes com Parkinson

Contexto: Os distúrbios neurológicos, como a Doença de Parkinson (DP), Alzheimer e epilepsia, impactam profundamente a vida dos pacientes e de suas famílias. Embora não haja cura, medicamentos podem aliviar significativamente os sintomas, especialmente nas fases iniciais. Porém, a maioria dos pacientes com Parkinson (PcP) depende de monitoramento e intervenções clínicas contínuas.

Visitas físicas frequentes ao médico podem ser difíceis para muitos PcP. O avanço da internet e das telecomunicações possibilita o monitoramento remoto desses pacientes através da telemedicina, o que pode reduzir o custo e a inconveniência das visitas. Uma área promissora para esse monitoramento é a avaliação da voz, já que cerca de 90% dos pacientes apresentam algum grau de comprometimento vocal. A medição da voz é simples e não invasiva, tornando-se uma ferramenta útil para detectar e acompanhar a progressão da doença.

Sua tarefa é auxiliar na análise dos dados e interpretar essas informações. Os dados e análises podem ser observados no trabalho de Little et al. (2009). O trabalho não é sobre predição da doença, mas sim, a análise das variáveis auxiliam na previsão.

Dados: Um conjunto de dados foi obtido durante um estudo com PcP e é composto por uma variedade de medições biomédicas de voz de 31 pessoas, sendo 23 com DP. Cada coluna na tabela representa uma medida específica da voz, e cada linha corresponde a uma das 195 gravações de voz desses indivíduos. Isso significa que cada sujeito foi medido mais de uma vez.

Os dados originais são chamados "parkinson_mdvp.csv", pois são dados utilizados relacionados ao *multidimensional voice program* (MDVP), totalizando 10 categorias extraídas das vozes. Quando os dados são carregados no arquivo p2-a.m na linha 8, as colunas estão na ordem abaixo, sendo a última coluna o "status".

As colunas são:

- 1 - MDVP:Fo(Hz)** - Frequência fundamental vocal média
- 2 - MDVP:Fhi(Hz)** - Frequência fundamental vocal máxima
- 3 - MDVP:Flo(Hz)** - Frequência fundamental vocal mínima
- 4 - MDVP:Jitter(%)** – Jitter¹ em porcentagem

¹ Variação ou instabilidade na frequência das vibrações das pregas vocais de um ciclo para o outro. Essencialmente, é uma medida da inconsistência no tom (ou frequência fundamental) da voz.



5 - MDVP:Jitter(Abs) – Jitter absoluto em microssegundos

6 - MDVP:RAP – Perturbação Relativa da Amplitude

7 - MDVP:PPQ – Quociente de perturbação no período de cinco-pontos

8 - MDVP:Shimmer – Shimmer² local

9 - MDVP:Shimmer(dB) – Shimmer em decibéis

10 - MDVP:APQ – Quociente de perturbação no período de onze-pontos

11 - Status – indica se o sujeito possui Parkinson, sendo “0” para não e “1” para sim.

Dúvidas nas métricas podem ser verificadas nos trabalhos de Little et al. (2009) e BOERSMA (2001).

Questionamentos: seu trabalho deve conter, obrigatoriamente, as análises que seguem, mas outras avaliações que o grupo julgar pertinentes podem ser consideradas. Utilize o Octave para responder:

Análise 1: Seleção das variáveis

- 1.1. Tendo em mente uma regressão linear e usando uma abordagem gráfica, quais variáveis você selecionaria para analisar com mais atenção na previsão de Parkinson? Justifique sua resposta. Sugestão: faça mais plots do tipo scatter.
- 1.2. Essas variáveis são os mesmos em sujeitos com e sem Parkinson? Se forem diferentes, selecione ao menos duas para cada caso (com e sem a doença).
- 1.3. Tomando os pares na questão (1.2.), utilize análises numéricas vistas em aula para ajudar a embasar sua decisão sobre qual par de variáveis melhor representa cada grupo (com e sem). Justifique sua resposta comentando os resultados obtidos.

Ao final dessa análise, você deve ser capaz de mostrar numericamente que certas duas variáveis possuem alta relação para pacientes com Parkinson e outras duas (ou as mesmas) para sujeitos sem Parkinson.

Análise 2: Comparação estatística das métricas

- 2.1. Em uma nova análise, tome suas variáveis de decisão que melhor predizem pacientes com Parkinson obtidos em (1.2). Com essas métricas, crie duas figuras, uma para cada variável, em que possamos ver o boxplot entre as variáveis do paciente com Parkinson comparadas com a mesma métrica do paciente sem Parkinson.
- 2.2. O que podemos observar em termos de valores absolutos para cada métrica em relação ao paciente com e sem Parkinson para cada métrica? Por exemplo,

² Variação do volume da voz ao longo do tempo. Assim como o **jitter** mede a instabilidade da frequência, o shimmer avalia a inconsistência na força ou intensidade da vibração das cordas vocais.



em qual grupo existe maior dispersão daquela variável? O que o desvio padrão e a média te dizem aqui? Inclua outras análises que julgar pertinente.

Análise 3: Predição

3.1. Faça dois modelos de regressão linear: um que contabiliza $y_1 = a_{0,1} + a_{1,1}x_1$ e outro $y_2 = a_{0,2} + a_{1,2}x_2$ onde x_1 e x_2 são as variáveis obtidas na Análise 1 para paciente com Parkinson. Calcule as somas dos resíduos S_r , r^2 e $s_{y/x}$ para cada modelo. Qual é melhor? Justifique.

3.2. Agora, faça um terceiro modelo de regressão linear múltipla de duas variáveis. Essa predição é para pacientes com Parkinson deve usar seu status como "1" para ser a variável de regressão de saída (i.e., $y_i = 1$).

Isso significa que você deve encontrar os parâmetros $a_{0,3}$, $a_{1,3}$ e $a_{2,3}$ de tal maneira que a regressão $y_3 = a_{0,3} + a_{1,3}x_1 + a_{2,3}x_2$ seja o melhor possível para x_1 e x_2 sendo as variáveis obtidas na Análise 1.

Calcule qual a soma dos resíduos S_r , r^2 e $s_{y/x}$ para o modelo y_3 . Compare-o com y_1 e y_2 . Qual é o melhor? Justifique.

3.3. Depois dessas análises, você diria que esse é um bom modelo de regressão para prever o paciente com a doença? Justifique sua resposta

Atenção: essa não é uma prova de aprendizado de máquina, mas de cálculo numérico. Logo, não espero uma análise utilizando métricas rebuscadas, discussões sobre modelos mais robustos ou que seu modelo performe bem. Quero que seu grupo foque na utilização do modelo construído, usando a ferramenta Octave e a interpretação dos resultados obtidos segundo a teoria dada em aula.

Códigos: o arquivo que deve resolver seu projeto é dado "p2_a.m". Esse código já carrega os dados e faz o plot de dispersão (scatter) das métricas duas a duas. Você pode criar outros scripts .m caso precise, mas eu só devo precisar rodar "p2_a.m" para verificar os entregáveis do seu projeto.

Funções proibidas:

- polyfit, linsolve, regress, interp1, interp2, interp3, interpn, spline, fitlm, compact, fitrlinear, mvregress, mvregresslike, plsregress.
- Também é proibido resolver sistemas lineares com o método da inversa (função inv()) ou pelo operador barra invertida (\). Caso você precise, utilize a função lu() para obter a decomposição e implemente uma pequena função que resolve o sistema dado que você possui L e U.



Referências:

BOERSMA, P. (2001). Praat, a system for doing phonetics by computer. *Glott. Int.*, 5(9), 341–345. <https://cir.nii.ac.jp/crid/1572261550900588928.bib?lang=en>

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease. *IEEE Transactions on Biomedical Engineering*, 56(4), 1015–1022. <https://doi.org/10.1109/TBME.2008.2005954>

ANEXO

Formato dos dados:

| | MDVP:F0(Hz) | MDVP:F1(Hz) | MDVP:F2(Hz) | MDVP:Jitter(%) | MDVP:Jitter(Abs) | MDVP:RAP | MDVP:PPQ | MDVP:Shimmer | MDVP:Shimmer(dB) | MDVP:APQ | status |
|---|-------------|-------------|-------------|----------------|------------------|----------|----------|--------------|------------------|----------|--------|
| 0 | 119.992 | 157.302 | 74.997 | 0.00784 | 0.00007 | 0.00370 | 0.00554 | 0.04374 | 0.426 | 0.02971 | 1 |
| 1 | 122.400 | 148.650 | 113.819 | 0.00968 | 0.00008 | 0.00465 | 0.00696 | 0.06134 | 0.626 | 0.04368 | 1 |
| 2 | 116.682 | 131.111 | 111.555 | 0.01050 | 0.00009 | 0.00544 | 0.00781 | 0.05233 | 0.482 | 0.03590 | 1 |
| 3 | 116.676 | 137.871 | 111.366 | 0.00997 | 0.00009 | 0.00502 | 0.00698 | 0.05492 | 0.517 | 0.03772 | 1 |
| 4 | 116.014 | 141.781 | 110.655 | 0.01284 | 0.00011 | 0.00655 | 0.00908 | 0.06425 | 0.584 | 0.04465 | 1 |

Figura 1. Cinco primeiras linhas dos dados em parkinson_mdvp.csv

Pairplots das variáveis:



Figura 2. Plots dois a dois das variáveis. Status = 0 para círculos azuis e Status = 1 para círculos laranjas.