MI813/ME601/ME701 - Fundamentos Estatísticos da Ciência de Dados
Segundo Semestre de 2020
Professor Responsável: Aluísio de Souza Pinheiro Sala 237
Aulas virtuais: terças e quintas, das quatorze às dezesseis
Atendimento virtual: quartas, 12h40 às 13h40

Plano de Trabalho Virtual para MI813/ME601/ME701

(a) Aulas virtuais com envio de material em pdf por email para os discentes e comunicação com os alunos durante o horário de aula através de G-talk, G-Classroom e G-Meeting, de acordo com o funcionamento dessas ferramentas e das redes envolvidas.

(b) Uso de G-talk, G-Classroom e G-Meeting durante os horários de atendimento, de acordo com o funcionamento dessas ferramentas e das redes envolvidas.

(c) Não haverá atividades presenciais, a não ser por decisão de órgãos superiores.

(d) Se for mantida a possibilidade de uso dos conceitos "S" e "R" para a pós ("S" e "I" para a graduação) e em lugar dos conceitos "A", "B" etc. (notas de zero a dez), todos os conceitos serão "S" e "R" para a pós ("S" e "I" para a graduação).

(e) Os conceitos (ou notas) serão conseqüências das atividades descritas na Seção III Forma de Avaliação.

MI813/ME601/ME701 - Fundamentos Estatísticos da Ciência de Dados

I. Objetivo

O curso proporciona aos alunos uma introdução formal e operacional aos conceitos da estatística em ciência de dados. Ao final do semestre, é esperado que o(a) aluno(a) seja capaz de:

1. Identificar problemas em que a alta dimensão tenha relevância;
2. Interpretar análises de dados;
3. Analisar conjuntos de dados, tanto do ponto de vista descritivo como inferencial, com ênfase em técnicas estatísticas sólidas; e
4. Propor e construir, inferencialmente, modelos para a análise de dados.

II. Conteúdo Programático

1. Áreas de Aplicação e Exemplos
2. Questões Estatísticas, Numéricas e Computacionais
3. Regressões
4. Mínimos Quadrados Penalizados
5. Modelos Lineares Generalizados e Verossimilhannça Penalizada
6. M-Estimadores Penalizados
7. Dados Funcionais
8. Inferência em Alta Dimensão
9. Seleção de Atributos
10. Regularização de Covariâncias e Modelos Gráficos
11. Aprendizado por Covariâncias e Modelos Fatoriais
12. Applicações de Modelos Fatoriais e ACP
13. Aprendizado Supervisionado
14. Aprendizado Não-Supervisionado
15. Aprendizado Profundo

III. Forma de Avaliação

A avaliação será realizada da seguinte forma:

(a) Três trabalhos individuais serão entregues pelos discentes, respectivamente denotados por P1, P2 e P3.

- Os temas serão escolhidos pelos alunos, num formulário automático disponibilizado na internet, pela ordem das escolhas feitas pelos alunos. Cada tema poderá ser escolhido por apenas um(a) aluno(a). Os temas se encontram em anexo, numerados de 1 a 250.
- Os dados serão providos pelos próprios alunos, em conformidade com o tema.

(b) Os trabalhos devem chegar à caixa postal de *pinheiro@ime.unicamp.br* nos seguintes prazos -

- P1 - dezesseis horas (horário de Brasília) do dia quatro de novembro de 2020;
- P2 - dezesseis horas (horário de Brasília) do dia 25 de novembro de 2020;
- P3 - dezesseis horas (horário de Brasília) do dia dezoito de dezembro de 2020.
- Exame Final (caso necessário) - dezesseis horas (horário de Brasília) do dia 21 de janeiro de 2021.

(c) Cada trabalho entregue no prazo estabelecido em (b) receberá conceito **S** ou **I/R**, conforme os preceitos abaixo -

Formato do trabalho - o texto principal do trabalho deve ser entregue como um relatório auto-contido num arquivo pdf de até 1MB, escrito em português. O arquivo deve permitir que trechos sejam copiados e colados e ferramentas de análise sejam usadas.

Metodologia - ligação do tema do projeto aos paradigmas da disciplina.

Exemplo - deve ter relevância científica, tecnológica ou metodológica e o porquê de sua solução por ciência de dados deve ser estabelecido no texto. A descrição do conjunto de dados deve fazer parte do texto do trabalho.

Rigor Estatístico - as questões do rigor estatístico da solução devem ser discutidas e referências bibliográficas devem ser apresentadas.

Coerência Textual - o texto deve exprimir uma visão coesa e clara sobre o tema abordado, por uma construção lógica com idéias elementares, complementares e consistentes. Ao final do texto, o leitor deve ter entendido **o quê**, **por quê** e **como**.

Arquivos suplementares (3) -

(1) Um arquivo com o conjunto de dados deve ser fornecido com instruções claras para descompactação etc..

(2) Um arquivo com o programa utilizado deve ser fornecido com instruções claras para seu uso. O programa deve reproduzir os resultados apresentados no texto e deve rodar, sem erros.

(3) Um arquivo deve conter o fluxograma que explique o método de análise e o algoritmo implementado no programa.

(d) Receberá conceito **I/R** o trabalho que não for entregue conforme (b)-(c).

(e) O conceito geral do aluno(a) será igual a

S, se ele(a) tiver exatamente três conceitos em P1, P2 e P3 iguais a **S**.

S, se ele(a) tiver exatamente dois conceitos em P1, P2 e P3 iguais a **S**.

I/R, se ele(a) tiver exatamente um conceito em P1, P2 e P3 igual a **S**.

I/R, se ele(a) não tiver qualquer conceito em P1, P2 e P3 igual a **S**.

(f) O(a) aluno(a) que tiver conceito geral **S** também terá conceito final igual a **S** e estará **aprovado(a)**, sem exame.

(g) O(a) aluno(a) que tiver conceito geral **I/R** poderá fazer um exame final da seguinte forma (itens (h)-(j)).

(h) O(a) aluno(a) deve entregar refeito cada trabalho cujo conceito original foi **I/R**, no prazo estipulado em (b) e formato estabelecido em (c). O conceito desse trabalho seguirá (b)-(d).

(i) O conceito **S** de um trabalho original será repetido para o mesmo trabalho no exame. Cada aluno(a) terá então seis conceitos (dois por trabalho - original e exame).

(j) O conceito final do(a) aluno(a), após o exame final, segundo os itens (g)-(h), será igual a

S, se ele(a) tiver exatamente quatro conceitos iguais a **S**.

S, se ele(a) tiver exatamente três conceitos iguais a **S**.

I/R, se ele(a) tiver exatamente dois conceitos iguais a **S**.

I/R, se ele(a) tiver exatamente um conceito igual a **S**.

I/R, se ele(a) não tiver qualquer conceito igual a **S**.

(k) O(a) aluno(a) que tiver conceito final igual a **S** estará **aprovado(a)**.

IV. Datas Importantes

16/09	Início do semestre letivo
04/11	Prazo Final para Entrega do Trabalho P1 (Dezesseis horas, horário de Brasília)
25/11	Prazo Final para Entrega do Trabalho P2 (Dezesseis horas, horário de Brasília)
18/12	Prazo Final para Entrega do Trabalho P3 (Dezesseis horas, horário de Brasília)
19/01/21	Prazo Final para Cumprimento do Programa
21/01/21	Prazo Final para Entrega do Exame Final (Dezesseis horas, horário de Brasília)

V. Conceitos e Notas

A tabela a seguir apresenta a equivalência entre notas e conceitos.

Nota Geral	Sistema Usual		Sistema Alternativo	
NG	Graduação	Pós-Graduação	Graduação	Pós-Graduação
[8,50; 10,0]	NG	**A**	**S**	**S**
[7,00; 8,49]	NG	**B**	**S**	**S**
[5,00; 6,99]	NG	**C**	**S**	**S**
[0,00; 4,99]	NG	**D**	**I**	**R**
Abandono	0,00	**E**	**I**	**R**

VI. Bibliografia

[0] Notas de aula

[1] Fan, J., Li, R., Z, C-H. & Zhou, H. (2020). *Statistical Foundations of Data Science*. CRC Press, Boca Raton, Florida.

[2] Bühlman, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer-Verlag, Berlin.

[3] Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer-Verlag, New York.

[4] Hastie, T., Tibshirani, R. & Wainwright, M.J. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, Boca Raton, Florida.

[5] Morettin, P.A., Pinheiro, A. & Vidakovic, . (2017). *Wavelets in Functional Data Analysis*. Springer-Verlag, New York.

[6] Ramsay, J. & Silverman, B. (2005). *Functional Data Analysis*. Second Edition, Springer-Verlag, New York.

[7] Ramsay, J., Hooker, G. & Graves, S. (2009). *Functional Data Analysis with R and Matlab*. Springer-Verlag, New York.

[8] Wainwright, M.J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Uiversity Press, Cambridge.

Nota:

- Os tópicos de **1 a 83** serão disponibilizados para **P1**.
- Os tópicos de **84 a 166** serão disponibilizados para **P2**.
- Os tópicos de **167 a 245** serão disponibilizados para **P3**.

Anexo - Tópicos para os trabalhos

1. Rise of Big Data and Dimensionality
2. Big Data in Biological Sciences
3. Big Data in Health Sciences
4. Big Data in Computer
5. Big Data in Information Sciences
6. Big Data in Economics
7. Big Data in Finance
8. Big Data in Business and Program Evaluation
9. Big Data in Earth Sciences
10. Big Data in Astronomy
11. Impact of Big Data
12. Impact of Dimensionality
13. Big Data and Computation
14. Big Data and Noise Accumulation
15. Big Data and Spurious Correlation
16. Big Data and Statistical theory
17. Aim of High-dimensional Statistical Learning
18. What big data can do

19. Regression Model Building and Basis Expansions
20. Regression Bias-Variance Tradeoff
21. Ridge Regression Solution Path
22. Kernel Ridge Regression
23. Regression in Reproducing Kernel Hilbert Space
24. Leave-one-out and Generalized Cross-validation
25. Bayesian Interpretation of Regression
26. Penalized Least-Squares - Selection of regularization parameters
27. Folded-concave Penalized Least Squares
28. Penalized Least-Squares - Orthonormal designs
29. Penalized Least-Squares - Penalty functions
30. Penalized Least-Squares - Thresholding by SCAD and MCP
31. Penalized Least-Squares - Risk properties
32. Penalized Least-Squares - Characterization of folded-concave PLS
33. Penalized Least-Squares - Lasso and L Regularization
34. Penalized Least-Squares - Nonnegative garrote
35. Penalized Least-Squares - Elastic Net
36. Penalized Least-Squares - Dantzig selector
37. Penalized Least-Squares - SLOPE and Sorted Penalties
38. Penalized Least-Squares - Concentration inequalities and uniform convergence
39. Penalized Least-Squares - Model selection
40. Penalized Least-Squares - Bayesian Variable Selection
41. Penalized Least-Squares - Bayesian view of the PLS

42. Penalized Least-Squares - Numerical Algorithms
43. Penalized Least-Squares - Quadratic programs
44. Penalized Least-Squares - Least angle regression
45. Penalized Least-Squares - Local quadratic approximations
46. Penalized Least-Squares - Local linear algorithm
47. Penalized Least-Squares - Penalized linear unbiased selection
48. Penalized Least-Squares - Cyclic coordinate descent algorithms
49. Penalized Least-Squares - Iterative shrinkage-thresholding algorithms
50. Penalized Least-Squares - Projected proximal gradient method
51. Penalized Least-Squares - ADMM
52. Penalized Least-Squares - Iterative Local Adaptive Majorization and Minimization
53. Penalized Least-Squares - Regularization parameters for PLS
54. Penalized Least-Squares - Degrees of freedom
55. Penalized Least-Squares - Extension of information criteria
56. Penalized Least-Squares - Residual variance and refitted cross-validation
57. Residual variance of Lasso
58. Penalized Least-Squares - Refitted cross-validation
59. Penalized Least-Squares - Extensions to Nonparametric Modeling
60. Penalized Least-Squares - Structured nonparametric models
61. Penalized Least-Squares - Group penalty
62. Penalized Least-Squares - Performance Benchmarks
63. Penalized Least-Squares - Performance measures
64. Penalized Least-Squares - Impact of model uncertainty

65. Penalized Least-Squares - Bayes lower bounds for orthogonal design
66. Penalized Least-Squares - Minimax lower bounds for general design
67. Penalized Least-Squares - Performance goals, sparsity and sub-Gaussian noise
68. Penalized Least-Squares - Penalized L Selection
69. Penalized Least-Squares - Lasso and Dantzig Selector
70. Penalized Least-Squares - Selection consistency
71. Penalized Least-Squares - Prediction and coefficient estimation errors
72. Penalized Least-Squares - Model size and least squares after selection
73. Penalized Least-Squares - Properties of the Dantzig selector
74. Penalized Least-Squares - Regularity conditions on the design matrix
75. Penalized Least-Squares - Properties of Concave PLS
76. Penalized Least-Squares - Properties of penalty functions
77. Penalized Least-Squares - Local and oracle solutions
78. Penalized Least-Squares - Properties of local solutions
79. Penalized Least-Squares - Global and approximate global solutions
80. Penalized Least-Squares - Smaller and Sorted Penalties
81. Penalized Least-Squares - Sorted concave penalties and its local approximation
82. Penalized Least-Squares - Approximate PLS with smaller and sorted penalties
83. Penalized Least-Squares - Properties of LLA and LCA
84. Generalized Linear Models - Maximum likelihood
85. Generalized Linear Models - Iteratively reweighed least squares
86. Generalized Linear Models - Deviance and Analysis of Deviance
87. Generalized Linear Models - Residuals

88. Generalized Linear Models - Bernoulli and binomial models
89. Generalized Linear Models - Models for count responses
90. Generalized Linear Models - Models for nonnegative continuous responses
91. Generalized Linear Models - Normal error models
92. Generalized Linear Models - Sparest solution in high confidence set
93. Generalized Linear Models - A general setup
94. Generalized Linear Models - Main Properties
95. Generalized Linear Models - Variable Selection via Penalized Likelihood
96. Generalized Linear Models - Algorithms
97. Generalized Linear Models - Local quadratic approximation
98. Generalized Linear Models - Local linear approximation
99. Generalized Linear Models - Coordinate descent
100. Generalized Linear Models - Iterative Local Adaptive Majorization and Minimization
101. Generalized Linear Models - Tuning parameter selection
102. Generalized Linear Models - Sampling Properties in low-dimension
103. Generalized Linear Models - Notation and regularity conditions
104. Generalized Linear Models - The oracle property
105. Generalized Linear Models - Sampling Properties with Diverging Dimensions
106. Generalized Linear Models - Asymptotic properties of GIC selectors
107. Generalized Linear Models - Properties under Ultrahigh Dimensions
108. Generalized Linear Models - The Lasso penalized estimator and its risk property
109. Generalized Linear Models - Strong oracle property
110. Generalized Linear Models - Risk properties

111. Variable selection in quantile regression

112. A fast algorithm for penalized quantile regression

113. Penalized composite quantile regression

114. Variable selection in robust regression

115. Variable selection in Huber regression

116. Rank regression and its variable selection

117. Variable Selection for Survival Data

118. Variable selection via penalized partial likelihood and its properties

119. Theory of folded-concave penalized M-estimator

120. Conditions on penalty and restricted strong convexity

121. Statistical accuracy of penalized M-estimator with folded concave penalties

122. Penalized M-estimators - Computational accuracy

123. Tools for exploring functional data

124. From functional data to smooth functions

125. Derivatives and functional linear models

126. Differential equations, operators, and functional data

127. Fitting differential equations to functional data: Principal differential analysis

128. Wavelet bases

129. Wavelet shrinkage

130. Wavelet-based Andrews' Plots

131. Functional ANOVA Models

132. Debias of regularized regression estimators

133. High Dimensional Inference in linear regression - Choices of weights

134. High Dimensional Inference in linear regression - Inference for the noise level
135. High Dimensional Inference in generalized linear models
136. High Dimensional Inference in linear regression - Desparsified Lasso
137. High Dimensional Inference in linear regression - Decorrelated score estimator
138. High Dimensional Inference in linear regression - Test of linear hypotheses
139. High Dimensional Inference in linear regression - Asymptotic efficiency
140. High Dimensional Inference in linear regression - Statistical efficiency and Fisher information
141. High Dimensional Inference in linear regression - random design
142. High Dimensional Inference in linear regression - Partial linear regression
143. High Dimensional Inference in linear regression - Gaussian graphical models
144. High Dimensional Inference in linear regression - Inference via penalized least squares
145. High Dimensional Inference in linear regression - Sample size in regression and graphical models
146. High Dimensional Inference in linear regression - General solutions
147. High Dimensional Inference in linear regression - Local semi-LD decomposition
148. High Dimensional Inference in linear regression - Data swap
149. High Dimensional Inference in linear regression - Gradient approximation
150. Generalized and Rank Correlation Screening
151. Feature Screening for Parametric Models
152. Generalized linear models
153. A unified strategy for parametric feature screening
154. Conditional sure independence screening

155. Nonparametric Screening

156. Feature Screening - Additive models

157. Feature Screening - Varying coefficient models

158. Feature Screening - Heterogeneous nonparametric models

159. Model-free Feature Screening

160. Sure independent ranking screening procedure

161. Feature screening via distance correlation

162. Feature screening for high-dimensional categorial data

163. Screening and Selection

164. Feature screening via forward regression

165. Feature Screening - Sparse maximum likelihood estimate

166. Feature screening via partial correlation

167. Refitted Cross-Validation

168. Refitted Cross-Validation algorithm

169. Refitted Cross-Validation in linear models

170. Refitted Cross-Validation in nonparametric regression

171. Sparse Covariance Matrix Estimation

172. Covariance regularization by thresholding and banding

173. Covariance Regularization - Asymptotic properties

174. Nearest positive definite matrices

175. Robust covariance inputs

176. Sparse Precision Matrix and Graphical Models

177. Covariance Regularization - Gaussian graphical models

178. Covariance Regularization - Penalized likelihood and M-estimation
179. Covariance Regularization - Penalized least-squares
180. Latent Gaussian Graphical Models
181. Factor model and high-dimensional PCA - Methods for selecting number of factors
182. Factor model and high-dimensional PCA - Robust initial estimation of covariance matrix
183. Augmented factor models and projected PCA
184. Factor model and high-dimensional PCA - Properties for estimating loading matrix
185. Factor model and high-dimensional PCA - Properties for estimating covariance matrices
186. Factor model and high-dimensional PCA - Properties for estimating realized latent factors
187. Factor model and high-dimensional PCA - Properties for estimating idiosyncratic components
188. Applications of Factor Models and PCA - FarmSelect
189. Applications of Factor Models and PCA - Asymptotic theory for FarmSelect
190. Applications of Factor Models and PCA - Factor-adjusted robust multiple testing
191. Applications of Factor Models and PCA - False discovery rate control
192. Applications of Factor Models and PCA - Multiple testing under dependence measurements
193. Applications of Factor Models and PCA - Power of factor adjustments
194. Applications of Factor Models and PCA - FarmTest
195. Applications of Factor Models and PCA - Factor Augmented Regression Methods
196. Applications of Factor Models and PCA - Principal Component Regression

197. Applications of Factor Models and PCA - Augmented Principal Component Regression
198. Applications of Factor Models and PCA - Matrix completion
199. Applications of Factor Models and PCA - Item ranking
200. Applications of Factor Models and PCA - Gaussian Mixture models
201. The standard support vector machine
202. Generalizations of SVMs
203. Sparse Classifiers via Penalized Empirical Loss
204. Supervised Learning - The importance of sparsity under high-dimensionality
205. Sparse support vector machines
206. Sparse large margin classifiers
207. Sparse Discriminant Analysis
208. Nearest shrunken centroids classifier
209. Features annealed independent rule
210. Selection bias of sparse independence rules
211. Regularized optimal affine discriminant
212. Linear programming discriminant
213. Direct sparse discriminant analysis
214. Solution path equivalence between ROAD and DSDA
215. Feature Augmention and Sparse Additive Classifiers
216. Feature augmentation
217. Penalized additive logistic regression
218. Semiparametric sparse discriminant analysis

219. Variable Selection in Clustering

220. Sparse clustering

221. Sparse model-based clustering

222. Sparse mixture of experts model

223. Inconsistency of the regular PCA for High Dimensional Data

224. Consistency of High Dimensional PCA under sparse eigenvector model

225. Sparse Principal Component Analysis

226. Unsupervised Learning - An iterative SVD thresholding approach

227. Unsupervised Learning - A penalized matrix decomposition approach

228. Unsupervised Learning - A semidefinite programming approach

229. Unsupervised Learning - A generalized power method

230. Deep Learning - Generative adversarial networks

231. Sampling view of Generative adversarial networks

232. Minimum distance view of Generative adversarial networks

233. Training deep neural nets

234. Deep Learning - Stochastic gradient descent

235. Deep Learning - Mini-batch SGD

236. Deep Learning - Momentum-based SGD

237. Deep Learning - SGD with adaptive learning rates

238. Deep Learning - Easing numerical instability

239. Deep Learning - ReLU activation function

240. Deep Learning - Skip connections

241. Deep Learning - Batch normalization

242. Deep Learning - Regularization techniques

243. Deep Learning - Weight decay

244. Deep Learning - Dropout

245. Deep Learning - Data augmentation