MI407 / ME732 A / ME921 – Análise Multivariada / Métodos em Análise Multivariada II / Métodos em Aprendizado Não-Supervisionado de Máquinas

Plano de Desenvolvimento

1 Ementa

Princípio: As disciplinas serão oferecidas de maneira mista e toda decisão sobre tópicos abordados no curso se dará entre docente e Coordenações de Graduação e Pós-Graduação. As respectivas Coordenações estão cientes do plano de atividades.

MI407: Distribuição normal multivariada e distribuição de Wishart. Inferência sobre vetor de médias e matriz de variância e covariância. Análise de componentes principais. Análise fatorial. Análise de variáveis canônicas e regressão. Análise de variância multivariada. Análise discriminante. Análise de conglomerados.

ME732: Análise de agrupamentos: tipos de algoritmos, similaridade, "k-means", hierárquicos e outros, métodos de validação. Análise de Associação: conjuntos de itens frequentes, regras booleanas, outros tipos de regras, avaliação de padrões de associação. Classificação: árvores de decisão, aprendizado de máquina, vizinho mais próximo, Bayesiano, métodos de grupos e avaliação de classificadores. Detecção de anomalias. Elaboração de um relatório final que inclua análise de dados.

ME921: Introdução à distribuição normal multivariada. Agrupamentos baseados em distância: agrupamentos hierárquicos e agrupamentos de K-médias. Generalizações de distância. Agrupamentos baseados em modelos. Agrupamentos baseados em densidades. Agrupamentos espectrais. Detecção de anomalias.

2 Calendário

13/04	**Atividade Teórica 1** (T_1)
20/04	**Atividade Prática 1**, entrega (P_1)
25/05	**Atividade Teórica 2** (T_2)
01/06	**Atividade Prática 2**, entrega (P_2)
17/06	Entrega do relatório do trabalho final (R, graduação)
17/06 a 06/07	Apresentações de artigos (A, pós-graduação)
08/07	**Avaliação substitutiva**, se for necessária (S)
20/07	**Exame** (E) (graduação)

3 Atividades e avaliação

- As aulas serão em formato síncrono através do Google Meet. As gravações serão disponibilizadas em formato de video aula no Moodle, de modo privado. A discussão dos materiais da aula e atendimento será feito em fórum no Moodle.
- Não será permitido comunicação através de e-mail: todas as consultas devem ser feitas através do Moodle.
- As notas da graduação serão numéricas (0 a 10) e as notas da pós-graduação serão conceitos (A, B, C, D, E), sem exceções.
- As Atividades Teóricas 1 e 2 devem ser submetidas exclusivamente através do Moodle. O aluno terá um período de 2 horas para entregar a atividade nas datas indicadas. O monitoramento do tempo de atividade é responsabilidade do aluno.
- As Atividades Práticas 1 e 2 devem ser submetidas exclusivamente através do Moodle. O aluno terá um período de 1 semana para entregar a atividade nas datas indicadas. O monitoramento do tempo de atividade é responsabilidade do aluno.
- O uso de meios fraudulentos para lograr sucesso nas avaliações resultará em nota 0 na avaliação correspondente para todos os alunos envolvidos. As atividades são individuais.
- O trabalho final é diferente entre graduação e pós-graduação. Alunos inscritos no Programa Integrado de Formação (PIF) devem seguir as instruções de alunos de **pós-graduação**.

- A apresentação de trabalho final para alunos da pós-graduação é um resumo e estudo detalhado de artigo recente (preferencialmente publicado após 2010), que envolva desenvolvimento metodológico em tópicos de aprendizado não-supervisionado, e que esteja publicado em revistas de alta seletividade internacionais (Annals of Statistics, Annals of Probability, Journal of the American Statistical Association, Journal of the Royal Statistical Society: Series B, Journal of Computational and Graphical Statistics, Bayesian Analysis, Journal of Machine Learning Research, Biometrics, outras revistas com Q1 no ScimagoJR, em Estatística, Matemática ou Ciência da Computação). Casos individuais serão analisados com base nos autores e complexidade da publicação. A busca do artigo é parte do projeto. Não há necessidade de submissão de relatório. A apresentação será feita nos dias indicados, em ordem aleatória.

- O relatório final para alunos da graduação é um relatório de no máximo 10 páginas, em formato de artigo (sem folha de rosto), individual, desenvolvendo a análise de clusteres em um conjunto de dados individual escolhido (ou preparado) pelo aluno. A seleção dos dados é parte da avaliação, e a complexidade dos dados será considerada na avaliação (dados extraídos de livros terão pontuação mínima, dados preparados pelo aluno em consulta a bancos de dados brasileiros e/ou coletados pelo próprio aluno terão pontuação máxima com respeito a esse critério). O aluno deve desenvolver a análise de clusteres justificando amplamente sua decisão com respeito a técnicas e meta-parâmetros do problema. O trabalho deve citar corretamente fontes. As fontes de dados devem ser apresentadas durante o semestre, para garantir que todos os alunos executarão análises individuais.

- Na pós-graduação a nota média M é $M = 0.15T_1 + 0.15P_1 + 0.15T_2 + 0.15P_2 + 0.4A$.

- O conceito final na pós-graduação será o melhor conceito entre dois critérios:

  - Critério I:
    - Se $8.5 \leq M \leq 10$, conceito é A (aprovado).
    - Se $6 \leq M < 8.5$, conceito é B (aprovado).
    - Se $5 \leq M < 6$, conceito é C (aprovado).
    - Se $M < 5$, conceito é D (reprovado).
  - Critério II:

    - Se M for maior ou igual ao percentil 90% das notas, conceito é A (aprovado).
    - Se M for maior ou igual ao percentil 60% das notas, mas estritamente menor que o percentil 90%, conceito é B (aprovado).
    - Se M for maior ou igual ao percentil 30% das notas, mas estritamente menor que o percentil 60%, conceito é C (aprovado).
    - Se M for estritamente menor que o percentil 30%, conceito é D (reprovado).

- Na graduação a nota média M é $M = 0.15T_1+0.15P_1+0.15T_2+0.15P_2+0.4R$. Se $M \geq 5$, o aluno está dispensado do exame, e $NF = M$ será a nota final.

- Se $M \leq 2.5$, o aluno não poderá fazer exame e está reprovado (Regimento Geral, Art. 52, item II). Caso contrário, a nota final será $NF = (M + E)/2$. O aluno com $NF \geq 5$ está aprovado.

- O aluno no PIF tem direito ao Exame na disciplina de graduação, mas o Exame não será computado no cálculo da nota de pós-graduação. O conceito será calculado na pós-graduação com base unicamente na nota M.

- Uma avaliação substitutiva só será aplicada em casos excepcionais e justificados (Regimento Geral, Art.72). A justificativa deverá ser feita em até 15 dias a partir do ocorrido, com cópia para a Coordenação responsável. Se a demanda pela substitutiva for baixa, poderá ser usado o exame no lugar da prova substitutiva. Somente uma atividade pode ser substituída por semestre.

4 Demais informações sobre o curso

A página do curso encontra-se em `http://www.ime.unicamp.br/~gvludwig/2021s1-me921`, com o calendário. Usaremos a página do curso no Moodle para divulgação de material.

O livro de Everitt *et al.* (2011), *Cluster Analysis, 5th edition*, será a referência principal durante a primeira metade do curso, mas posteriormente usaremos Bouveyron *et al.* (2019). A leitura das outras referências é fortemente recomendada.

Referências

C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-Based Clustering and Classification for Data Science: with applications in R*, volume 50. Cambridge University Press, 2019.

M. A. Cox and T. F. Cox. *Multidimensional Scaling.* Chapman and Hall / CRC, 2000.

M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, volume 96, pages 226–231, 1996.

B. Everitt and T. Hothorn. *An Introduction to Applied Multivariate Analysis with R.* Springer, 2011.

B. S. Everitt, S. Landau, M. Leese, and D. Stahl. *Cluster Analysis, 5th edition.* John Wiley & Sons, 2011.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning.* Springer, 2008.

M. Hahsler, M. Piekenbrock, and D. Doran. `dbscan`: Fast density-based clustering with R. *Journal of Statistical Software*, 91(1):1–30, 2019.

M. Hubert, P. J. Rousseeuw, and S. Van Aelst. High-breakdown robust multivariate methods. *Statistical science*, pages 92–119, 2008.

A. J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning.* Springer, 2008.

L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons, 2009.

J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons, 2019.

K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis.* Academic Press, 1979.