Aula 1 – Mineração de Dados
Apresentação

Profa. Elaine Faria

UFU

Dados do Professor e da Sala Virtual

- Elaine Ribeiro de Faria Paiva
- Email: elaine@ufu.br
- Sala: 1B137
- Horário de atendimento (agendado por email com 24h de antecedência):
  - Segunda: 08:50h às 10:30h
  - Quarta: 13:30 às 15:10h

Dados da Sala Virtual

- Microsoft Teams
  - Mineração de Dados - 2022-1
  - Código: **ojznhse**

Sala Virtual

- No Microsoft Teams será divulgado
  - Notas, faltas e avisos
  - Slides das aulas e material complementar
  - Link para as aulas gravadas
  - Enunciado das atividades assíncronas
  - Enunciado das atividades avaliativas

- Os alunos deverão fazer as entregas de trabalhos e exercícios usando o Microsoft Teams

Como serão as aulas?

- Dias e Horários:
  - Quinta: 19:00 às 20:40 - Teórica
  - Terça: 20:50 às 22:20 - Prática (a maioria delas em laboratório - Laboratório 03)

- Calendário:
  - 26/09/22 a 10/02/23
  - 18 semanas de aulas
    - 4 aulas extras para cumprir a carga horária

Objetivo

- Introduzir o aluno às
  - Principais tarefas e técnicas de Mineração de Dados.
- Habilitar o aluno a
  - Aplicar ferramentas de Mineração de Dados em problemas práticos.
  - Implementar suas próprias ferramentas de Mineração de Dados.

Conteúdo

- **Introdução**
  - O que é Mineração de Dados ; o que é Descoberta de Conhecimento (KDD)
  - As fases do processo de KDD
  - Principais Tarefas de Mineração de Dados.
- **Preparação dos Dados**
  - Sumarização dos dados
  - Limpeza dos dados: valores ausentes, tratamento de ruídos
  - Integração e Transformação dos dados
  - Redução dos dados: seleção de atributos, redução de dimensionalidade
  - Discretização, Normalização

Conteúdo

- **Associação**
  - Mineração de Regras de Associação – Algoritmo Apriori e variantes.
  - Mineração de Sequências – Algoritmos GSP e Prefix
- **Classificação**
  - O que é um classificador?
  - Árvore de Decisão
  - Classificadores baseados nos vizinhos mais próximos (KNN).
  - Classificadores baseados em Redes Bayesianas de Crença.
  - Avaliação de Performance: Método Holdout, Cross-Validation, Bootstrap

Conteúdo

- **Agrupamentos (Clusters)**
  - Diferentes tipos de clusters
  - Diferentes tipos de clusterização
  - Método K-Means e K-Medóides – análise de performance, complexidade.
  - Método hierárquico aglomerativo – análise de performance, complexidade.
- **Análise de Clusters**
  - Medidas: coesão, separação, SSE, coeficiente de silhueta.
  - Técnicas para determinar o número correto de clusters.
  - Técnicas para determinar a tendência de clusters nos dados.

Conteúdo

- **Detecção de Anomalias (Outliers)**
  - Introdução: causas de anomalias
  - Técnicas para detecção de anomalias: estatísticas e baseadas em proximidade
- **Pós-Processamento: Análise, Interpretação e Visualização**

Referências Bibliográficas

- **Básica**
  - TAN, Pang-Ning. Introdução ao Data Mining: mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

https://www-users.cs.umn.edu/~kumar001/dmbook/index.php (alguns capítulos estão disponíveis)

  - J. Han, M. Kamber: Data Mining: Concepts and Techniques, 2a. Ed., Morgan Kaufmann, 2006.
  - I. H. WiΣen, E. Frank: Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.

Referências Bibliográficas

- **Complementar**
  - CHAKRABARTI, Soumen. Mining the Web: discovering knowledge from hypertext data. San Francisco: Morgan Kaufman, c2003.
  - KANTARDZIC, Mehmed. Data mining: concepts, models, methods, and algorithms. 2nd ed. Hoboken; Piscataway: John Wiley & Sons: IEEE Press, c2011. E-Book.
  - LAROSE, Daniel T. Data mining methods and models. Hoboken; Piscataway: John Wiley & Sons: IEEE Press, c2006. E-Book.
  - JENSEN, Finn V. Bayesian networks and decision graphs. 2nd ed. New York: Springer, c2007.
  - CHERKASSKY, Vladimir S. Learning from data: concepts, theory, and methods. 2nd ed. Hoboken; Piscataway: John Wiley & Sons: IEEE Press, c2007. E-Book.

Referências Bibliográficas

- **Complementar**
  - Link para o livro do prof. Wagner Meira da UFMG

    Data Mining and Machine Learning: Fundamental Concepts and Algorithms

    http://dataminingbook.info/

  (também estão disponíveis videos sobre o conteúdo do livro)

Avaliação de assiduidade

- **Presença nas aulas**

- **Entrega das atividades avaliativas**

Atenção: os alunos poderão ser reprovados se não atingirem o mínimo de presença nas aulas

Sistema de Avaliação

- 1ª Avaliação: 17/11/22 - 35 pontos
- 2ª Avaliação: 12/01/23 - 35 pontos

- Trabalho em grupo - 30 pontos
  - 1o: 03/11/22
  - 2o: 08/12/22
  - 3o: 05/01/23

Sistema de Avaliação

- **Observações**
  - As Implementações poderão ser feitas em qualquer linguagem de programação
  - As entregas das atividades avaliativas e não-avaliativas será feita pelo Microsoft Teams em prazo pré-determinado
  - Não serão aceitas entregas atrasadas!
  - Não serão aceitas atividades copiadas dos colegas ou da Internet
  - A responsabilidade pela entrega correta da atividade é do aluno

Recuperação do conteúdo

- O aluno deverá
  - Procurar o professor em horário extra
- O professor disponibilizará
  - um atendimento individual
  - lista de exercícios
  - material complementar para estudo dirigido

Recuperação de Nota

- Prova substitutiva ao fim do semestre
  - Conterá todo o conteúdo da disciplina
  - Substituirá as notas das provas ao longo do semestre
    - Pode substituir a 1a, 2a ou ambas as provas
  - Será também usada para os alunos que perderam alguma prova durante o semestre
  - **Data: 26/01/23**

Cronograma

SEMANA	MÓDULOS	ATIVIDADES PREVISTAS	CARGA-HORÁRIA
1	Apresentação e Introdução a Mineração de Dados	Apresentação da disciplina e do professor Introdução a Mineração de Dados Etapas do KDD; Tarefas da Mineração de Dados	4 horas/aula
2	Preparação de Dados	Representação e Preparação de Dados Aula Prática: Representação e Preparação de Dados	4 horas/aula
3	Preparação de Dados	Preparação de Dados e Medidas de Distância Aula Prática: Ferramenta Weka	4 horas/aula
4	Classificação	Medidas de distância e Classificação: K-vizinhos mais próximos (KNN) Aula Prática: Exercícios sobre KNN	4 horas/aula

Cronograma

5	Classificação	Classificação: Árvores de decisão e Naive Bayes Aula Prática: Árvore de Decisão e Naive Bayes	4 horas/aula
6	Classificação	Classificação: Medidas e Métodos de Avaliação Aula Prática: Exercícios sobre Medidas e Métodos de Avaliação	4 horas/aula
7	Classificação	Classificação: Medidas e Métodos de Avaliação Revisão para a Prova	4 horas/aula
8	Avaliação	1a prova Resolução da prova **Aula Extra:** Vista de Prova	6 horas/aula

Cronograma

9	Agrupamento	Agrupamento: Algoritmos Particionais Aula Prática: Exercícios sobre Agrupamento: Algoritmos Particionais	4 horas/aula
10	Agrupamento	Agrupamento: Algoritmos Hierárquicos Aula Prática: Exercícios sobre Agrupamento: Algoritmos Hierárquicos	4 horas/aula
11	Agrupamento	Agrupamento: Medidas de Validação Aula Prática: Agrupamento - Medidas de Validação	4 horas/aula
12	Regras de Associação	Regras de Associação: Apriori Aula Prática: Exercícios sobre Regras de Associação - Apriori	4 horas/aula

Cronograma

13	Regras de Associação	Regras de Associação: Mineração de sequências Aula Prática: Exercícios sobre Regras de Associação - Mineração de Sequências	4 horas/aula
14	Anomalias	Detecção de Outliers e Anomalias Pós-processamento: análise, interpretação e visualização	4 horas/aula
15	Avaliação	2a prova Resolução da 2a prova **Aula Extra:** Vista de Prova	6 horas/aula
16	Trabalhos	Apresentação de trabalhos Apresentação de trabalhos	4 horas/aula
17	Avaliação	Prova substitutiva Resolução prova substitutiva	4 horas/aula

Observações

- Sobre o trabalho:
  - Não serão aceitos trabalhos entregues fora do prazo estipulado pelo professor (exceto em caso de atestado)
  - É responsabilidade do aluno entrar no sistema, aprender a usá-lo e fazer a submissão
  - Plágio é crime!
    - Cópias da internet ou dos colegas não serão aceitas

Observações

- Sobre as aulas
  - A presença aulas é muito importante, pois é o momento que temos para tirar dúvidas e fazer perguntas
  - Use o momento de aula para tirar dúvidas ou conversar com o professor
  - Aproveite as aulas do laboratório para resolver exercícios em grupo e tirar dúvidas
    - A Internet deve ser usada para pesquisar sobre o conteúdo da disciplina

Observações

- Sobre o acompanhamento da disciplina
  - Os slides do professor não são suficientes para o estudo da disciplina
  - Estude usando materiais online, como por exemplo, livros e artigos disponíveis na internet
  - O estudo diário é imprescindível para o bom andamento do aluno na disciplina

Observações

- Sobre a frequência
  - A frequência nas aulas será cobrada
  - Atestado médido de um dia não abona falta
    - Procure a coordenação caso tenha atestado médico para muitos dias
  - Se você manifestou qualquer sintoma e precisa se ausentar das aulas, converse com o professor o mais rápido que puder

O que os alunos esperam??

- Sugestões?
- Expectativas?