Universidade Federal de Viçosa

Tópicos Especiais: Ciência de Dados

Apresentação da Disciplina e Introdução

Prof. Marcos H. F. Ribeiro

Departamento de Informática - UFV

Descrição Geral

- Atualmente, muitas disciplinas vêm sendo oferecidas no DPI com alguma relação a Inteligência Artificial e Aprendizado de Máquinas
- Esta disciplina pertence, também, a este contexto
- Mas, normalmente, tais disciplinas possuem foco nas técnicas e algoritmos
- No entanto, esta disciplina, apesar de também lidar com, e tratar de algoritmos, tem seu foco principal no processo todo de Ciência de Dados, enfatizando aspectos de análise, engenharia preparação de dados, bem como em aspectos da análise e comparação de resultados e sua aplicação
- Além dos aspectos acima, terá foco também na metodologia de treinamento e avaliação de modelos
- Desta forma, ela é complementar com outras disciplinas da área, como Mineração de Dados, Aprendizado de Máquina, Visão Computacional, *Deep Learning* e outras

Objetivos da disciplina

- Proporcionar uma visão geral da área de Ciência de Dados (*Data Science*), com foco nas etapas de preparação de dados, análise de resultados e modelos e também em aplicações reais ou realísticas.
- Relacionar a visão ampla do conceito de Ciência de Dados com as áreas correlatas de Aprendizado de Máquina e Mineração de Dados.
- Apresentar um repertório de técnicas que possibilite a atuação em contextos envolvendo suporte à tomada de decisões, seja na área acadêmica ou não.

Metodologia

- Aulas expositivas, com slides
- Aulas e exemplos práticos, usando a linguagem *Python* e *Python Notebooks*
- Não haverá foco na implementação da grande maioria dos algoritmos de mineração de dados e aprendizado de máquina. Ao invés disso:
  - Serão apresentados os princípios básicos de funcionamento dos algoritmos
  - Serão utilizadas bibliotecas já existentes que implementam os algoritmos vistos, o que não elimina o emprego de implementação e programação de computadores durante as aulas
  - As implementações feitas durante a disciplina visam o uso e aplicação dos métodos em problemas reais ou realísticos (cenários fictícios que emulam cenários reais)
- Desenvolvimento de projetos de aplicação
- Para a pós-graduação: seminários relacionando os temas da disciplina aos projetos de dissertação/tese

Outras Informações Gerais

- Horário das aulas:
  - Segundas-feiras, às 16:00h
  - Quintas-feiras, às 14:00h
- Professores:
  - Marcos Ribeiro (marcosh.ribeiro@ufv.br)
  - Daniel Louzada (daniel.louzada@ufv.br)

Conteúdo

1. Introdução
   - O que é Ciência de Dados (*Data Science*)
   - Conceitos importantes: *Data Mining* versus *Machine Learning* versus *Data Science* versus IA
   - Visão geral do curso e suas lições
2. Fundamentos de análise de dados
   - Representação de dados
   - Tipos de atributos
   - Fundamentos de estatística
     - População e amostra
     - Variáveis aleatórias
     - Média, mediana e moda
     - Amplitude, quantis e quartis
     - Desvio padrão, coeficiente de variação, variância
     - Covariância e correlação
     - Distribuições
     - Teste de hipóteses, testes não paramétricos e ANOVA
   - Visualização de dados na análise exploratória
   - Análise uni e multivariada

Conteúdo

3. Pré-processamento e preparação de dados
    - Normalização (escalonamento e padronização)
    - Identificação e remoção de dados discrepantes (*outliers*)
    - Tratamento de valores faltantes
    - Pré-processamento de dados categóricos
    - Categorização de dados numéricos
4. Redução de dimensionalidade e seleção de atributos
    - Remoção de redundância por correlação
    - Métodos de seleção de atributos
        - *Mutual information*
        - Seleção por variância
        - Seleção usando modelos simples e Extra Trees
        - Teste-F
        - Métodos *Embedded*, *Filter* e *Wrapper*
    - Análise de Componentes Principais (PCA)

Conteúdo

5. Agrupamentos (*Clusters*)
    - Definição / conceitos básicos
    - Tipos de agrupamentos
    - Visão geral dos principais algoritmos
    - Métricas de avaliação e validação de agrupamentos
    - Visualização de dados na análise de agrupamentos

Conteúdo

6. Classificação
   - Definição / conceitos básicos
   - Construindo um Classificador Linear Simples
   - Algoritmos
     - *White box* versus *Black box*
     - Principais famílias e algoritmos
   - Metodologia para treinamento de classificadores
     - Treinamento e teste
     - Métricas de avaliação e validação de classificadores
     - Validação Cruzada
     - Comparação de modelos
   - Classificação multi classe e multi rótulo
   - Casos especiais
     - *Underfitting*
     - *Overfitting*
     - Classes não balanceadas
     - Classes ausentes
   - Visualização de dados na análise de classificadores

Conteúdo

7. Regressão
    - Definição / conceitos básicos
    - Paralelos e diferenças com classificação
    - Construindo um Regressor Linear Simples
    - Principais famílias de algoritmos
    - Metodologia para treinamento de regressores
        - Treinamento e teste
        - Métricas de avaliação e validação de classificadores
        - Validação Cruzada
        - Comparação de modelos
    - Visualização de dados na análise de classificadores

Conteúdo

8. Comitês
    - Tipos de comitês e tipos de votação
    - Avaliação de comitês
9. Explicabilidade de Modelos
    - Interpretabilidade intrínseca e extrínseca
    - Abordagem específica e agnóstica
    - Técnicas local e global
    - Principais métodos
        - LIME
        - SHAP
        - ELI5
    - Desafios e limitações
    - Estudos de caso

Conteúdo

10. Noções de outras técnicas de aprendizado
    - Aprendizado semi-supervisionado
    - Séries temporais
    - Aprendizado por reforço
    - Regras de Associação
11. Visão geral de áreas de aplicação modernas
    - Processamento de Linguagem Natural
    - Visão Computacional
    - Modelos Generativos
    - Federated Learning

Para saber mais

Tem dúvidas ou quer saber mais? Faça contato.