UNIVERSIDADE FEDERAL DE GOIÁS
Pós Graduação em Sistemas e Agentes Inteligentes

Profa. Dra. Juliana Felix
julianafelix@ufg.br

TRABALHO FINAL DA DISCIPLINA DE APRENDIZADO DE MÁQUINA APLICADO A DADOS ESTRUTURADOS

Descrição

O projeto final visa reforçar e aplicar os diversos conceitos e técnicas estudadas ao longo da disciplina, além de capacitar o aluno a apresentar, de forma detalhada e demonstrando conhecimento, as informações e os resultados encontrados com as técnicas utilizadas. Como objetivo secundário, essa atividade visa preparar o aluno para o desenvolvimento do projeto final de curso (TCC).

Neste projeto, a ser realizado em grupos de até 4 alunos, vocês deverão:

1. Navegar no repositório de bases de dados da University of California Irvine (UCI), disponível neste endereço:
    - https://archive.ics.uci.edu/datasets
    - Alternativamente, você poderá utilizar outro repositório disponível online, desde que os dados estejam disponibilizados de modo gratuito (ex.: kaggle, physionet.org)
2. Escolher um (1) dataset que:
    - seja composto de pelo menos 100 atributos/features numéricas,
    - tenha pelo menos 500 amostras,
    - possa ser categorizado em duas ou mais classes, e
    - possua dados faltantes (ou você deverá simular esses dados faltantes).
    - O dataset MNIST não poderá ser utilizado neste trabalho final.
    - Opcionalmente, você pode utilizar um banco de dados formado por imagens ou séries temporais. Neste caso, **o requisito de quantidade de features consiste em obter um vetor de características de pelo**

Profa. Dra. Juliana Felix
julianafelix@ufg.br

menos 100 features, obtido a partir de transformações ou extração de características a partir de cada série/imagem.

3. Desenvolver, em Python, e utilizando as bibliotecas de sua preferência, um sistema para lidar com o problema de classificação delimitado por você (pela base de dados escolhida). Sua solução deverá:
    - explorar e apresentar visualizações dos dados disponíveis na base;
    - utilizar alguma técnica para lidar com os dados faltantes, ou justificar a dispensa dessa fase mesmo havendo dados faltantes na base;
    - realizar transformações/escalas nos dados, quando julgar necessário;
    - explorar pelo menos uma técnica de redução de dimensionalidade (PCA, LDA, ou T-SNE, por exemplo);
    - comparar a performance de pelo menos 3 classificadores diferentes, sendo que:
        - cada classificador escolhido (ex.: SVM, KNN, Random Forest…) deverá ser avaliado sob pelo menos duas variações de ajustes finos, através de alterações nos hiperparâmetros. O uso do `gridsearch` é recomendado para essa etapa, mas outras técnicas podem ser implementadas.
        - a comparação deve ser feita tanto no conjunto de validação (utilizando k-fold cross-validation) como no conjunto de teste separado especialmente para este propósito;
        - a matriz de confusão, a acurácia e pelo menos outras 2 métricas de avaliação de classificadores (de sua escolha e justificadas) devem ser apresentadas para comparar o desempenho dos modelos.
4. Escrever um artigo/relatório com os seguintes tópicos:
    - Introdução
        - Descrever o problema (não o conjunto de dados), fornecendo uma motivação para investigar o problema escolhido.

Profa. Dra. Juliana Felix
julianafelix@ufg.br

Preferencialmente, busque artigos científicos que já lidam com o tema e a base de dados para construir a sua motivação do estudo. Sugestão de locais para busca de artigos:

- https://scholar.google.com/
- https://ieeexplore.ieee.org/Xplore/home.jsp

- Materiais e Métodos
  - Descrever o conjunto de dados com informações numéricas e gráficas das informações disponíveis no dataset.
    - Ex.: Apresente figuras (sempre de forma comentada!) que ilustram o dataset, as features, e a relação entre as features ou classes contidas na base. Nesta fase, é importante deixar claro o total de amostras, tipo e total de características, como as informações foram obtidas (no caso de haver essa explicação na fonte da base de dados), e possíveis relações de (des)balanceamento entre classes.
  - Descrever todas as etapas realizadas nas diversas fases do desenvolvimento:
    - pré-processamento de dados (limpeza, transformação, escala dos dados)
    - extração e/ou seleção de características (ex.: combinação de atributos, técnicas de redução de dimensionalidade, escolhas empíricas, etc)
    - classificação e métricas de avaliação
      - informe quais os classificadores escolhidos, justificando brevemente a escolha, bem como quaisquer ajustes finos realizados,
      - apresente como foi realizada a divisão dos dados em conjunto de treinamento, validação e teste,

Profa. Dra. Juliana Felix
julianafelix@ufg.br

    - descreva quais as métricas de avaliação foram utilizadas para comparar os modelos implementados.

- Resultados
    - A análise dos resultados deve ser feita individualmente, para cada classificador analisado.
    - Em cado caso, apresente uma matriz de confusão com os resultados obtidos **na fase de treinamento**, com a validação-cruzada empregada. Comente eventuais tendências ou observações encontradas a partir da matriz.
    - Após a apresentação do resultado dos 3+ classificadores, compare brevemente os resultados encontrados **no conjunto de teste** utilizando as métricas de avaliação escolhidas (acurácia e mais 2). Você também pode utilizar a matriz de confusão nessa fase.
- Conclusões
    - Fornecer as conclusões ou considerações finais sobre o trabalho desenvolvido. Em especial, forneça uma recomendação, baseado nos dados e resultados observados, de qual classificador deve ser escolhido para a tarefa (se identificar algum).
- Referências
    - Apresentar a bibliografia utilizada para apoiar seu trabalho (livros, artigos, sites).
        - Procure referenciar as citações nos locais adequados, ao invés de simplesmente fornecer o link/referência utilizado.

O que deverá ser submetido:

Profa. Dra. Juliana Felix
julianafelix@ufg.br

- Um Python notebook (.ipynb) com o código desenvolvido, com cada etapa organizada e comentada de forma clara, e com o resultado da execução de cada célula visível (quando for o caso).
    - Antes de fazer a submissão, limpe o kernel do notebook e execute todas as células de uma vez. Envie o notebook apenas após a execução de todas as células ser finalizada.
- Um arquivo .pdf com o artigo elaborado conforme instruções acima. O artigo deve ser desenvolvido utilizando o modelo de artigos da SBC, disponível em: https://www.sbc.org.br/documentos-da-sbc/summary/169-templates-para-artigos-e-capitulos-de-livros/878-modelosparapublicaodeartigos ou o modelo IEEE para conferências (A4 Word ou LaTeX), disponível para download em: https://www.ieee.org/conferences/publishing/templates.html

Sua solução será avaliada considerando a corretude e a clareza do código, além da clareza, estrutura, coerência argumentativa e nível de detalhamento do artigo/relatório.

Data limite de entrega: até 23h59 do dia 21/09/2025, via google SIGAA.

Abaixo, indico alguns artigos que lidam com aprendizado de máquina e que podem ser utilizados como referência sobre **como** apresentar a motivação do problema estudado, a base de dados utilizada, as técnicas de pré-processamento e análise de dados, e como reportar os resultados obtidos. Tenha em mente que os artigos publicados geralmente tem um limite de páginas definido pela conferência ou revista e, portanto, podem apresentar certas seções de forma bastante sucinta. No trabalho desta disciplina, não há limite de páginas e o objetivo é verificar a sua compreensão dos tópicos estudados e do pipeline completo de análise de dados. Portanto, procure descrever com detalhes todos os passos executados na sua atividade.

- A Parkinson's Disease Classification Method: An Approach Using Gait Dynamics and Detrended Fluctuation Analysis

Profa. Dra. Juliana Felix
julianafelix@ufg.br

- An Intelligent System to Improve Diagnostic Support for Oral Squamous Cell Carcinoma
- Insect Predation Estimate Using Binary Leaf Models and Image-Matching Shapes
- An automatic method for estimating insect defoliation with visual highlights of consumed leaf tissue regions
- Rede Neural Multicamadas para Classificação de Doenças Neurodegenerativas a partir de Sinais de Marcha
- Redução do Viés Intra-Sujeito na Classificação da Doença de Parkinson pela Voz com Otimização por Colônia de Formigas
- Aplicação de Algoritmos de Aprendizado de Máquina na Análise da Vulnerabilidade Social e Insegurança Alimentar
- Uma nova abordagem de padrões binários em radiografias de tórax para avançar o diagnóstico de tuberculose