Trabalho 2

Tópicos de Aprendizagem Automática

Contents

Descrição:

Modelo: Trabalho de grupo de 3 alunos

Projeto: Neste documento serão disponibilizados alguns temas possíveis a desenvolver neste projeto. Contudo os alunos são fortemente encorajados a propor um problema de aprendizagem automática da sua preferência, que não esteja listado e que reflita melhor os seus interesses. Por favor, discuta a sua ideia com o docente.

Pretende-se que o grupo aplique algoritmos de aprendizagem automática adequados, aprendidos em aula ou de forma autónoma. O problema em estudo fará uso de algoritmos de ciência de dados (classificação, regressão, clustering).

Os resultados deverão ser apresentados em formatos gráficos/tabelares, acompanhados da respetiva análise e conclusões.

Âmbito: Trabalho prático, onde os alunos deverão resolver um problema de machine learning, que será proposto pelo grupo, recorrendo a bases de dados disponíveis nas plataformas web, fazendo uso das ferramentas lecionadas.

É permitido o uso de ferramentas de AI, desde que mencionadas, privilegiando as que permitem fazer tracking das referências (por exemplo: https://scite.ai/ , https://consensus.app/ , https://elicit.com/).

Entregáveis: 3 entregas programadas

- Entrega 1 – primeira semana de abril – Definição de grupo, tema, título, objetivos
- Entrega 2 – primeira semana de maio - Resultados preliminares (algoritmo selecionado, metodologia de tratamento de dados)
- Entrega 3 – última semana de maio – Documento final, código e apresentação

Avaliação:

- Entrega 1 e 2: sem avaliação quantitativa, apenas qualitativa em caso de necessidade de ajuste
- Documento: qualidade do documento, rigor técnico/ científico, dificuldade do problema, capacidade critica. Máximo 6 páginas, formato IEEE.

- Apresentação: qualidade da apresentação, capacidade de discussão do tema
- Avaliação entre pares dentro do grupo
- Avaliação pelos colegas da turma
- Documento (40%), apresentação (40%), Avaliação dentro do grupo (10%), Avaliação turma (10%)

Entrega: Última semana de maio, deverão entregar documento, e código desenvolvido no âmbito do projeto.

Apresentação: 2 últimas aulas do semestre, 10 min de apresentação + 5 min de discussão.

Objetivo: Desenvolver as capacidades de comunicação oral e escrita, capacidade critica, raciocínio lógico e discussão.

Desenvolver capacidade de uso de ferramentas de ML, construção de pipeline de processamento, avaliação de performance e discussão técnica dos resultados.

Estrutura do Documento:

O documento deverá refletir o trabalho executado de forma clara, permitindo perceber como foram tratados os dados, e implementada toda a metodologia. Os resultados deverão ser apresentados de forma critica e baseada na evidência.

(Dado o material estar em inglês, a estrutura seguinte está em inglês. Contudo o relatório poderá ser apresentado em português.)

- **State of the art review**

Search and review of at least 5-6 references (papers, reports, thesis, etc.) handling the same or similar problem. Make a review of different techniques used to solve the problem you want to explore.

- **Data description, visualization and statistical analysis**

Describe the problem you want to solve, the features and visualize the data (if it is difficult due to high dimension, show only some samples). Provide some statistical analysis such as metadata (e.g. features range of variation), histograms, try to identify if there are some data quality problems, detect interesting subsets.

- **Data preprocessing (if relevant)**

Describe possible preprocessing steps to construct the final input to the machine learning algorithm from the initial data, such as data normalization, feature selection or dimensionality reduction in case of redundant features.

- **Description of the applied machine learning algorithm(s)**

Apply a suitable ML algorithm (learned in class or self-learned) to solve the problem with the chosen dataset. Introduce the method shortly, define its parameters. Make a selection of the most important model hyper parameters after their variation in a selected range. Show graphically the results of this search.

Clear statement what is the ML problem (classification, regression, what are inputs (how many), and the outputs of the model).

- **Presentation and discussion of results**

Presentation of the results preferably in a graphical format. Analysis, discussion, interpretation. Compare your results with the results in the reviewed references or apply and compare at least two ML methods on the same problem.

- **Conclusions**

Critical discussion of the gained knowledge regarding the advantages/disadvantages of the applied methods on the problem in hand. Describe the problem complexity. Suggestions for potential future directions of study.

Possíveis temas (mas não restrito a estes):

Sign language understanding

Hand gestures and sign language are the most commonly used methods by deaf and non-speaking people to communicate among themselves or with speech-able people. However, understanding sign language is not a universal skill. For this reason, building a system that recognizes hand gestures and sign language can be very useful to facilitate the communication gap between speech-able and speech-impaired people.

Project proposal 1: Identification of digits from sign language images

Data source: https://www.kaggle.com/datasets/ardamavi/sign-language-digits-dataset

(Google: Sign Language Digits Dataset)

Project proposal 2: American sign language understanding

Data source: https://www.kaggle.com/datasets/datamunge/sign-language-mnist (Google: Sign Language MNIST)

Take the data from this fixed source !!!

https://github.com/darkin1/sign-language-digits-ml/blob/master/dataset_fixed.zip

Project proposal 3: Mammographic Mass Data Set

This data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. The aim is to discriminate benign from malignant cases assuming that all cases with BI-RADS assessments greater or equal a given value (varying from 1 to 5), are malignant and the other cases are benign.

Data source: http://archive.ics.uci.edu/ml/datasets/mammographic+mass

Project proposal 4: Heart Disease Data Set

This dataset contains 4 heart disease related datasets. For the present project you will use the Cleveland database and the referred subset of 14 features form a total of 76 attributes. The goal is to distinguish presence (values 1,2,3,4) from absence (value 0) of heart disease in the patient.

Data source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Project proposal 5: Bank Marketing Data Set

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe or not a term deposit.

Data source: https://www.kaggle.com/janiobachmann/bank-marketing-dataset

Project proposal 6: The German Traffic Sign Benchmark (GTSB)

GTSB is a multi-class, single-image classification challengeheld at IJCNN 2011. Automatic recognition of traffic signs is required in advanced driver assistance systems and constitutes a challenging real-world computer vision and pattern recognition problem. A comprehensive, lifelike dataset of more than 50,000 traffic sign images has been collected. It reflects the strong variations in visual appearance of signs due to distance, illumination, weather conditions, partial occlusions, and rotations. The dataset comprises 43 classes with unbalanced class frequencies.

Data source: https://www.kaggle.com/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign

Project proposal 7. Kaggle Credit Card Fraud Detection

https://www.kaggle.com/mlg-ulb/creditcardfraud

Recommended Data Repositories:

- Kaggle Data Repository : https://www.kaggle.com/datasets
- UCI Machine Learning Repository : https://archive.ics.uci.edu/ml/index.php