MACHINE LEARNING II

Debates & Coding Competition

Topics & Instructions

Aude Billard

Topics for Debates:

Success of ML depends most on :

a) the data;

b) the algorithm;

c) none of these

Which approach is the most effective for real-world applications?

a) Supervised learning;

b) Unsupervised Learning

Real-world deployment will remain hindered until we have:

a) better methods for incremental learning;

b) good interfaces for lay users;

c) more efficient storage methods and faster computing resources

Generalisation in machine learning:

a) can be readily assessed with several metrics and benchmarks;

b) is ubiquitous;

c) is an ill-posed problem

Biases in machine learning :

a) can be resolved with heavier computation & curated datasets;

b) is not an issue for the vast majority of applications;

c) is by essence irresolvable

Machine learning and climate change:

a) growth of the former is incompatible with the latter,

b) both can be made compatible under certain constraints;

c) this is a non-issue

Debates Guidelines: Machine Learning Tradeoffs

The goal of these debates is to engage in an in-depth discussion on controversial topics related to machine learning techniques and their real-world applications.

Preparation Requirements:

- **Research:** Read relevant literature to build a well-informed perspective.
- **Materials:** Prepare a slide deck and a 2-4 pages summary outlining your key arguments.

Debate Structure:

1. Opening Statements:

- Each debater presents their position using 1-2 slides.

2. Live Debate (15-20 minutes):

- Participants engage in discussion, debating with one another and responding to audience questions.
- Debaters must support their arguments with additional slides, providing concrete examples, quantifiable results, etc.

The debates are an opportunity to critically analyze different viewpoints, defend your position with well-researched arguments, and learn on how to engage in a dynamic discussion and build convincing arguments.

Coding Competitions - Topics

Average Monthly surface temperature (1940-2024)

https://www.kaggle.com/datasets/samithsachidanandan/average-monthly-surface-temperature-1940-2024

This Dataset contains details of Average Monthly surface temperature (1940-2024). Current climate change is primarily caused by human emissions of greenhouse gases. This warming can drive large changes in sea level, sea ice and glacier balances, rainfall patterns, and extreme temperatures. This has potentially devastating impacts on human health, farming systems, the stability of societies, and other species.

Coding Competitions - Topics

Sleep Health and Digital Screen Exposure Dataset

https://www.kaggle.com/datasets/arifmia/sleep-health-and-digital-screen-exposure-dataset

This dataset contains multiple health-related attributes collected from individuals, including their sleep quality, stress levels, heart rate, and screen exposure habits. It can be used for statistical analysis, machine learning modeling, and health-related research.

Coding Competitions - Topics

- **Full Netflix Dataset**

https://www.kaggle.com/datasets/octopusteam/full-netflix-dataset/data

This dataset provides a comprehensive collection of all titles (Movies and TV Series) available on Netflix. In addition to basic information, it includes IMDb-specific data like IMDb ID, Average Rating, and Number of Votes.

Coding Competitions - Topics

- **Traffic Accidents**

https://www.kaggle.com/datasets/oktayrdeki/traffic-accidents

This dataset contains detailed information on traffic accidents across various regions and time periods. It includes various metrics such as accident date, weather conditions, lighting conditions, crash types, injuries, and vehicle involvement. The data span multiple locations and accident types, offering a comprehensive view of traffic incidents and their causes.

Coding Competitions - Topics

Food Nutrition Dataset

https://www.kaggle.com/datasets/utsavdey1410/food-nutrition-dataset/data

The dataset provides detailed nutritional information for a wide range of food items commonly consumed around the world. This dataset aims to support dietary planning, nutritional analysis, and educational purposes by providing extensive data on the macro and micronutrient content of foods.

Coding Competitions - Topics

Calories Burned During Exercise and Activities

https://www.kaggle.com/datasets/aadhavvignesh/calories-burned-during-exercise-and-activities/data

This dataset contains the number of calories burned by a person while performing some activity/exercise. It contains 248 activities and exercises ranging from running, cycling, calisthenics, etc.

Coding Competitions - Topics

AI/ML Salaries

https://www.kaggle.com/datasets/cedricaubin/ai-ml-salaries/data

The salaries are from ai-jobs. Ai-jobs collects salary information anonymously from professionals all over the world in the AI/ML and Big Data space and makes it publicly available for anyone to use, share and play around with. The data is being updated regularly with new data coming in, usually on a weekly basis.

Coding Competitions - Topics

Education & Career Success

https://www.kaggle.com/datasets/adilshamim8/education-and-career-success

This dataset explores the relationship between academic performance and career success. It includes 5000 records of students' educational backgrounds, skills, and career outcomes. The dataset can be used for various analyses, such as predicting job success based on education, identifying key factors influencing salaries, and understanding the role of networking and internships in career growth.

Coding Competitions - Topics

French employment, salaries, population

https://www.kaggle.com/datasets/etiennelq/french-employment-by-town/data

This dataset explores the demographics of salary distribution across various regions of France. It provides several demographic information to enable to draw patterns.

Coding Competition Guidelines: Machine Learning for Structure Discovery

The objective of this competition is to apply machine learning techniques for structure discovery, as covered in class, to a real-world dataset.

Expectations:

1. Initial Data Analysis:

- Perform a preliminary analysis of the dataset using standard statistical methods.
- Visualize data distributions and compute key statistics (e.g., mean, median).

2. Application of Machine Learning Techniques:

- Implement at least one (or multiple) structure discovery techniques from class.
- You may also explore alternative approaches, such as manifold learning, kernel-based clustering, or dimensionality reduction methods.

3. Identifying Key Insights:

- Evaluate the results and determine the most meaningful findings.
- Prepare a 2-4 pages report and accompanying slides that address the following:
  - Techniques Used: Justify your choice of method(s).
  - Hyperparameter Selection: Explain how you tuned the model’s parameters.
  - Findings & Insights: Analyze the results and their implications.
  - Limitations & Challenges: Discuss any shortcomings or potential biases in your analysis.

The goal is not only to apply machine learning techniques but also to critically assess their effectiveness and limitations in uncovering meaningful structure within the dataset.