DS 207 Final Project Guidelines

Purpose

The DS 207 final project is designed to develop your skills in building machine learning pipelines for real-world applications. This assignment represents 40% of your final grade.

Grades will be calibrated by group size and individual contributions.

Week 03: Project Proposal (5%)

Step 1: Form a project team with classmates from your enrolled section only. The maximum is 4 teams per section (~4 students per team).

Step 2: Create and submit a PDF document that includes:

- Full names of all team members.
- Berkeley.edu email addresses for each member.
- Project description (400-500 words) covering:
  - **Motivation:** What is the question you will be working on? Why is it interesting?
  - **Data:** Provide a description of your dataset, including source(s) with links.
  - **Related work:** Include at least one example of related work (use Google Scholar for papers, or provide links to website(s) for other products)
  - **Methodology (optional):** What ML algorithms are you planning to apply, what experiments are you planning to run, and how do you plan to evaluate your ML algorithms?
  - **Project GitHub repo:** Create a GitHub repository for your group project. All project code must be maintained in this repository. Include the repository link in your proposal and make sure to add your instructor as a contributor (username will be provided)

Step 3: Submit one proposal per team to Gradescope. Please make sure to tag all the team members in your submission.

Note: You have complete flexibility in choosing your project topic and dataset. If your dataset requires minimal preprocessing (such as many Kaggle datasets), we expect you to place significantly more emphasis on the Methodology section by implementing additional models and aiming for competitive leaderboard metrics.

Week 07: Project Milestone (10%)

The project milestone ensures you are making progress on your project and serves as a checkpoint for your ML pipeline development process.

Step 1: Create and submit a PDF document (**at most 3 pages long**) that includes:

- Full names of all team members.
- Berkeley.edu email address for each member
- Project progress covering:
  - **Motivation:** What is the question you are tackling? Why is it interesting?
  - **Data:** Describe in detail your dataset, including its source and size (both before and after preprocessing). Clearly specify what the input and target of interest is, and the number of examples allocated to your training, validation, and test sets.
    - Data preprocessing: Explain all preprocessing techniques applied to your data, including: handling of missing values, outlier detection and treatment, feature scaling or normalization methods, encoding of categorical variables, feature selection approach.
    - Exploratory Data Analysis (EDA): Document your EDA process, including: distribution analysis of key features (visualize and discuss patterns, skewness, etc.), correlation analysis between features and target variables, identification of potential relationships and insights, feature importance assessment, time-based patterns or trends (if applicable).
    - Visualization requirements: Include multiple detailed plots that effectively communicate your data insights: ensure all plots have properly labeled x and y axes; include descriptive titles; add legends where appropriate; consider using multiple plot types (histograms, scatter plots, box plots, heatmaps, etc.) to highlight different aspects of your data; accompany each visualization with interpretations of what the patterns reveal.
    - Data challenges: Discuss any challenges encountered with your dataset and how you addressed them.
  - **Methodology:** Describe your desired machine learning approach, including which algorithms you plan to implement (at least three: a baseline and two improvements over baseline) and why they make sense for your problem. Describe the experiments that you plan to run, including any hyperparameter tuning strategies, validation methods, and evaluation metrics you will use to assess model performance. Explain how you will compare different models and determine which one is the best for your outcome of interest. If you have any preliminary results, include a brief discussion of these findings.
  - **Contributions:** Provide a detailed breakdown of each team member's specific contributions to the project, including the Jupyter notebooks created/maintained by each person. Note that we allow only one contributor per notebook, and all notebooks must have cell numbering starting at 1 in your GitHub repository. This documentation ensures team members are carrying a fair share of the workload and helps us assess individual contributions.

Step 2: Submit one milestone per team to Gradescope. Please make sure to tag all the team members in your submission.

Guidelines for PDF submissions: You can use either single-column or two-column layouts. Paper size should be Letter (8.5 x 11 inches) with font size between 10-12 point and margins of at least 0.5 inch on all sides. Your submission must be in PDF format.

Week 14: Project Report (20%)

Step 1: Create and submit a PDF document (**at most 5 pages long**) that includes:

- Full names of all team members.

- Berkeley.edu email addresses for each member.
- Project report covering:
    - **Abstract (1 paragraph):** Summarize your project's motivation, machine learning methodology, and key findings.
    - **Introduction (0.5 pages):** Explain the problem and its importance. Clearly define inputs and outputs: "The input to our algorithm is *{specific data type}*, and we use *{specific models}* to predict *{specific output}*."
    - **Related work (0.5 pages):** Review at least 2 relevant papers or products. Group them by approach, discuss their strengths and weaknesses, and compare them to your work. Identify state-of-the-art methods and explain what makes certain approaches particularly effective. Use any citation format consistently.
    - **Dataset (at most 1 page):** Include dataset description, source, size (before and after preprocessing), data preprocessing steps, and distribution across training, validation and test sets. Thoroughly discuss all EDA visualizations and their implications for your modeling approach (see: Project Milestone data section).
    - **Methods (1-1.5 pages):** Describe at least three algorithms you implemented (a baseline and two improvements). Explain why each is appropriate for your problem and provide a paragraph on how each works. If you implement a linear regression model, logistic regression model, or any neural network, all code must be written in TensorFlow.
    - **Experiments, Results and Discussion (1-2 pages**): Describe the experiments you run for each model presented in the Methods section, including any hyperparameter tuning strategies, validation methods, and evaluation metrics you used to assess model performance. Show subgroup performance evaluations as well (e.g., if your data includes gender, show evaluation results for male vs. female). Discuss overfitting concerns and mitigation strategies. Analyze why certain models outperformed others and justify your final model selection.
    - **Conclusions (1-2 paragraphs):** Summarize your work and key findings. Identify the best-performing algorithms and explain why they succeeded. Discuss limitations and suggest future improvements if you had more time or computational resources.
    - **Contributions:** Provide a detailed breakdown of each team member's specific contributions to the project, including the Jupyter notebooks created/maintained by each person. Note that we allow only one contributor per notebook, and all notebooks must have cell numbering starting at 1 in your GitHub repository. Each team member should contribute to the Methods and Experiment section (for a classification task, a majority class classifier doesn't count as credit). This documentation ensures team members are carrying a fair share of the workload and helps us assess individual contributions.
    - **Appendix (optional):** Include any additional details that support your analysis but aren't essential to the main report, such as additional visualizations, detailed hyperparameter searches, or supplementary results.

Step 2: Submit one report per team to Gradescope. Please make sure to tag all the team members in your submission.

Guidelines for PDF submissions: You can use either single-column or two-column layouts. Paper size should be Letter (8.5 x 11 inches) with font size between 10-12 point and margins of at least 0.5 inch on all sides. Your submission must be in PDF format.

Week 14: In-Class Project Presentation (5%)

Step 1: Prepare a slide presentation (PowerPoint or PDF) that effectively summarizes your Project Report. Your presentation should address all key components: question, motivation, related work, dataset (preprocessing, EDA), methods, experiments, results, and conclusions. There is no need to upload your presentation to Gradescope.

Step 2: Present your work during the live session, keeping your presentation to a maximum of 12 minutes per team. Be prepared for questions following your presentation.