**Machine Learning 1- Yilmaz**

**2024-2025 Quarter 2 Project**

### **Project Topics**

Your first task is to pick a project topic. The project needs to be an **algorithmic project**: Pick a problem or family of problems, and develop a new learning algorithm, or a novel variant of an existing algorithm, to solve it. You can concentrate on some subfield of machine learning that you want to explore more.

Once you have identified a topic of interest, it can be useful to look up existing research on relevant topics by searching related keywords on an academic search engine such as: [http://scholar.google.com](http://scholar.google.com/). You may also concentrate on the algorithms we have learned so far and may want to read up on their variations to come up with your own variation. However, you are NOT restricted to the algorithms we have learned in class so far.

After coming up with your own algorithm and implementing it, I expect you to compare its performance against the related algorithms. For that purpose, you can use any data set you want.

Please refrain from choosing your topic merely based on the code you could find using sites like paperswithcode.com. The topics can be simple but you could still do valuable research if you could grasp the topic itself.

**Grading:**

The final report and presentation will combine to be the majority of the grade. Projects will be evaluated based on:

* **The technical quality of the work**. (I.e., Does the technical material make sense? Are the things tried reasonable? Are the proposed algorithms or applications clever and interesting? Do the authors convey novel insight about the problem and/or algorithms? Does the project have sufficient scope for the given team size?)
* **Originality**. (Did the authors add their own data processing, methods, or analysis? Does the final project avoid being a mirror image of existing papers/projects with no net new work?)
* **Communication.** (Are the authors able to clearly and effectively explain the work that they did, including context, methods, and results? Do the paper and presentation balance clarity with rigor?)

In order to highlight these components, it is important you present a solid discussion regarding the learnings from the development of your method, and summarizing how your work compares to existing approaches.

**Project Parts: Proposal, Final Report, and Presentation**

**Proposal (due December 12 or 13, 2024 based on your section):**

Your project proposal should include the following information:

* **Motivation**: What problem are you tackling?
* **Method**: What machine learning techniques are you planning to improve upon?
* **Intended experiments**: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?

Presenting pointers to one relevant dataset and one example of prior research on the topic are a valuable addition.

**Final Report (January 15 or 16, 2025 depending on your section)**

**0. Title page: Project title, your names, date, class-teacher, and “Quarter 2 Project”**

**1. Abstract [≈ 1 paragraph]** It should consist of 1 paragraph consisting of the motivation for your paper and a high-level explanation of the methodology you used/results obtained.

**2. Introduction [≈ 0.5 pages]** Explain the problem and why it is important. Discuss your motivation for pursuing this problem. Give some background if necessary. If you are using a dataset for performance evaluations, clearly state what the input and output is. Be very explicit: “The input to our algorithm is an {image, amplitude, patient age, rainfall measurements, grayscale video, etc.}. We then use a {SVM, neural network, linear regression, etc.} to output a predicted {age, stock price, cancer type, music genre, etc.}.”

**3. Related work [≈ 0.5 pages]** You should find existing papers, group them into categories based on their approaches, and discuss their strengths and weaknesses, as well as how they are similar to and differ from your work. In your opinion, which approaches were clever/good? What is the state-of-the-art? Do most people perform the task by hand? You should aim to have at least 5 references in the related work. Include previous attempts by others at your problem, previous technical methods, or previous learning algorithms. Google Scholar is very useful for this: https://scholar.google.com/ (you can click “cite” and it generates MLA, APA, BibTeX, etc.)

**4. Dataset and Features [≈ 0.5 − 1 pages]** Describe your dataset you are using for performance analysis: How many training/validation/test examples do you have? Is there any preprocessing you did? What about normalization or data augmentation? How is data discretized? Include a citation on where you obtained your dataset from. Depending on available space, show some examples from your dataset. You should also talk about the features you used.

**5. Methods [≈ 1 − 1.5 pages]** Describe your learning algorithms, proposed algorithm(s). Make sure to include relevant mathematical notation. For example, you can briefly include the SVM optimization objective/formula or say what the softmax function is. It is okay to use formulas from the lecture notes. For each algorithm, give a short description (≈ 1 paragraph) of how it works. Again, I am looking for your understanding of how these machine learning algorithms work. Additionally, if you are using a cutting-edge algorithm (e.g. long short-term memory, SURF features, or anything else not covered in the class), you may want to explain your algorithm using 1/2 paragraphs.

**6. Experiments/Results/Discussion [≈ 1 − 3 pages]** You should give details about what parameters you chose (e.g. why did you use a certain k for KNN or K-means and why) and how you chose them. Before you list your results, make sure to list and explain what your primary metrics are: accuracy, precision, etc. Provide equations for the metrics if necessary. For results, you want to have a mixture of tables and plots. If you are solving a classification problem, you should include a confusion matrix or ROC curves. Include visualizations of results, examples of where your algorithm failed and a discussion of why certain algorithms failed or succeeded. In addition, explain whether you think you have overfit to your training set and what, if anything, you did to mitigate that. Make sure to discuss the figures/tables in your main text throughout this section. Your plots should include legends, axis labels, and have font sizes that are legible when printed.

**7. Conclusion/Future Work [≈ 1 − 2 paragraphs]** Summarize your report and reiterate key points. Which algorithms were the highest performing? Why do you think that some algorithms worked better than others? For future work, if you had more time, more team members, or more computational resources, what would you explore?

**8. Contributions** The contributions section is not included in the 5 page limit. This section should describe what each team member worked on and contributed to the project.

**9. References/Bibliography (No page limit)** This section should include citations for: (1) Any papers mentioned in the related work section. (2) Papers describing algorithms that you used which were not covered in class. (3) Code or libraries you downloaded and used. This includes libraries such as scikit-learn, Matlab toolboxes, Tensorflow, etc. Acceptable formats include: MLA, APA, IEEE. If you do not use one of these formats, each reference entry must include the following (preferably in this order): author(s), title, conference/journal, publisher, year. If you are using TeX, you can use any bibliography format which includes the items mentioned above. Main body text, figures, and any discussions are strictly forbidden from this section. The references section is excluded from the page limit to encourage students to perform a thorough literature review/related work section without being space-penalized if they include more references.

**Presentations due: January 15 or 16, 2025 (depending on your section)**

**Submission**: Please create a Google drive folder and share it with me. I need you to submit the link to the folder when you submit your proposal. You will place your final report and presentation slides as well as dataset under this folder.

**Some sample posters and reports can be found** [**here**](https://cs229.stanford.edu/proj2020spr/)**.**

**Some of my previous machine learning students’ reports that were published in TJ’s TEKNOS journal :**

2022:

* “Novel OVO Strategy for SVMs: Neural Network Classifier (NNCSVM)” *by Jeb Chi, Raffu Khondaker, Andy Wang.*
* “Combatting the Normality Assumption in Naive Bayes: A Novel Method of Classification Combining Gaussian and Beta Distributions” *by Allison Seo, Shreya Singh.*

2023:

* “Finding Order in Chaos: A Systematic Alternative to Random Forest” by *Katie Merrill, Ramya Reddy*.
* “K-nearest Neighbors Traversal Clustering: A Method for Clustering Data of Differing Densities and Non-linear Partitions”, Kathline Newland.
* “Enabling Jump Point Search to Work on Weighted Graphs” by Andrei Basto.

**Optional:** A very good project will be a publishable or nearly-publishable piece of work. Each year, some number of students continue working on their projects after completing this course, submitting their work to conferences or journals. Thus, for inspiration, you might also look at some recent machine learning research papers. Two of the main machine learning conferences are ICML and NeurIPS. You can find papers from the recent ICML <https://icml.cc/Conferences/2020/Schedule> and NeurIPS conference <https://neurips.cc/Conferences/2020/Schedule>.

**More exhaustive list of ML conferences and journals is** [**here**](http://www.wikicfp.com/cfp/call?conference=machine%20learning)**.**

Here are a few tips to keep in mind:

- Journals are hard to get accepted but they provide comprehensive feedback.

- New conferences may be easier to get accepted.

- If you are not sure about a particular conference, check out their web site and subjects they are interested in. It may also be a good idea to check out the papers they have previously accepted.

- You may choose a conference based on the location: if you want to visit a certain city at a certain time etc.

- Please keep in mind that publishing a paper may take several attempts. But your paper may improve and get better at each attempt.