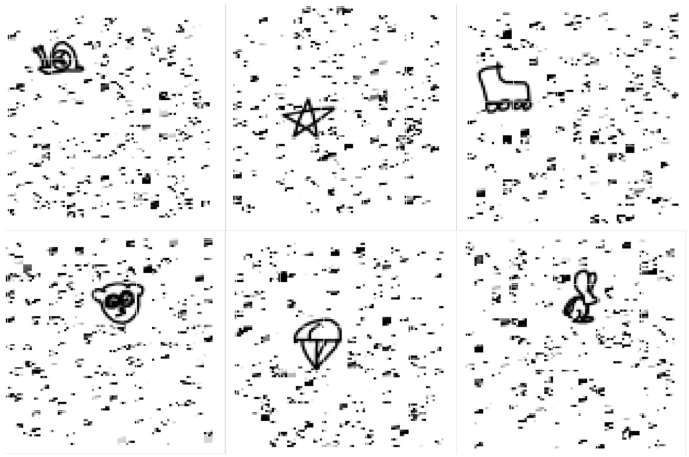
INF8953CE - MACHINE LEARNING (FALL 2020), KAGGLE COMPETITION

Deadline for team formation: November 12, 08:00 pm, EST
Kaggle submission closing: December 09, 08::00 pm EST
Deadline for submitting report: December 12, 08:00 pm EST

1 BACKGROUND

For this project, you will take part in a Kaggle competition based on image analysis. The goal is to design a machine learning algorithm that can automatically identify hand-drawn images as well as reason about their appearance. The dataset we have prepared is a variant of google's quick draw dataset. For that dataset, a popular goal has been to simply identify the given human drawn images. For our variant, we've reduced the number of classes to 31. To create an image, we've appended the drawing to a random location on a larger 100x100 pixel blank canvas image. Additionally, we've added randomly generated noise around the drawing. Note that one of the classes is called "empty" and consists of only noise and no human drawings. The dataset consists of 10k images of size (100,100) for the training set and 10k for the test set. You will be evaluated on test accuracy. Examples of the training samples are shown here:



First, you need to create an account on the Kaggle website, if you haven't already. Next, you can access the competition, including the data, through this private link:

https://www.kaggle.com/c/f2020-INF8953CE/

We expect you to be working in groups of **exactly 3**.

For this competition, you are NOT supposed to use any external data (other than what is provided by us) and also you are NOT supposed to use any pre-trained models. If we find you using external data or pre-trained models, you will get ZERO mark for this competition.

2 KAGGLE TEAM FORMATION

Each team should consist of exactly 3 members. To form a team:

- Fill out the google form (https://forms.gle/ok4XYU9tMRgJXrUK7) with your team information by Nov 12^{th} at 08:00 pm, EST. Any team not registered or registered later will not be graded.
- Register as an individual Kaggle user
- Enter the competition and accept terms and conditions.
- Go to https://www.kaggle.com/c/f2020-INF8953CE/team
- In the "Invite Others" section, enter your teammates' names, or team name.
- Your teammate has the option to accept your merge. The person accepting a merge is the team leader.

****** IMPORTANT NOTE ******

The maximum amount of submissions is 2 per day, per TEAM.

You should not make any submissions before the team formation deadline which is 12 November, 2020 08:00 pm EST. If you make any submissions before this, you will get a ZERO for this competition.

All the team members will receive same marks for this competition. It is your duty to make sure everyone has contributed to the competition equally.

3 INSTRUCTIONS

To participate in the competition, you must provide a list of predicted outputs for the instances on the Kaggle website. To solve the problem, we expect you to try the following methods:

- A baseline linear learner consisting of SVM or logistic regression, implemented by hand or using a library.
- Any other ML method of your choice. Be creative! Some suggestions are neural network trained by back-propagation, k-NN, random forests, kernelized SVM, CNN's, etc.

For the Kaggle competition, you can submit results from your best performing system, whichever method (from the above two categories) it may fall under. Note that there is a maximum of 2 prediction submissions per day on Kaggle. You can submit 2 predictions per day over the course of the competition, so we suggest you start early, allowing yourself enough time to submit multiple times and get a sense of how well you are doing.

4 REPORT

In addition to your methods, you must write up a report that details the preprocessing, validation, algorithmic, and optimization techniques, as well as providing your Kaggle results that we compare them with. The report should contain the following sections and elements:

- Project title
- Team name on Kaggle, as well as the list of team members, including their full name, email, and matricule.
- Introduction: briefly describe the problem and summarize your approach and results.
- Feature Design: Describe and justify your pre-processing methods, and how you designed and selected your features.
- Algorithms: Give an overview of the learning algorithms used without going into too much detail in the class notes (e.g. SVM derivation, etc.), unless necessary to understand other details.
- Methodology: Include any decisions about training/validation split, distribution choice for Naive Bayes, regularization strategy, any optimization tricks, setting hyper-parameters, etc.

- Results: Present a detailed analysis of your results, including graphs and tables as appropriate. This analysis should be broader than just the Kaggle result: include a short comparison of the most important hyper-parameters and all the methods you implemented.
- Discussion: Discuss the pros/cons of your approach & methodology and suggest areas of future work.
- Statement of Contributions. Briefly describe the contributions of each team member towards each of the components of the project (e.g. defining the problem, developing the methodology, coding the solution, performing the data analysis, writing the report, etc.) At the end of the Statement of Contributions, add the following statement: "We hereby state that all the work presented in this report is that of the authors."
- References (optional).
- Appendix (optional). Here you can include additional results, more detail of the methods, etc.

  The main text of the report should not exceed 8 pages. References and appendix can be in excess of the 8 pages. You should uss the ICLR format. You can find the template in the following link: https://www.overleaf.com/latex/templates/template-for-iclr-2021-conference-submission/mmpfhsxmqdkp

5 SUBMISSION REQUIREMENTS

- You must submit the code developed during the project. The code must be well-documented. The code should include a README file containing instructions on how to run the code. Submit the code as an attachment (see Submission Instructions).
- Your submission should contain a notebook named "final.ipynb" which would reproduce your predictions exactly. Make sure to fix the random seeds so that the generated predictions are exactly matching your submitted prediction file.
- The prediction file must be submitted online at the Kaggle website. **Please make sure your submitted result file has the correct structure and format. You should submit your result in .csv format.** More information about the correct structure and format could be found in Kaggle website (go to 'overview' $\rightarrow$ 'Evaluation').
- You must submit a written report according to the general layout described earlier.

6 SUBMISSION INSTRUCTIONS

For this project, you will submit the report to Gradescope, and the code to Moodle.

- Submit a zipped folder to Moodle.
  - Your zip file should contain a folder called "code" that contains all code and data. Make sure all the data files needed to run your code is within the folder and loaded with a relative path. We should be able to run your code without making any modifications.
- Your group report should be submitted to Gradescope.

One submission per team is sufficient for both code and report.

7 LATE SUBMISSION POLICY

Late submission policy is the same as default policy used for the other assignments.

8 EVALUATION CRITERIA

Marks will be attributed based on 30% for performance on the private test set in the competition, 70% for the written report. For the competition, the performance grade will be calculated as follows: The top team, according to the score on the private test set, will receive 100%. A simpler

classifier, entered by the instructor, will score 0%. All other grades will be calculated according to the interpolation of the private test set scores between those two extremes. For the written report, the evaluation criteria include:

- Technical soundness of the methodology (pre-processing, feature selection, validation, algorithms, optimization).
- Technical correctness of the description of the algorithms (may be validated with the submitted code).
- Meaningful analysis of final and intermediate results.
- Clarity of descriptions, plots, figures, tables.
- Organization and writing. Please use a spell-checker and don't underestimate the power of a well-written report!!

Do note that the grading of the report will emphasize the rationale behind the pre-processing and optimization techniques. The code should be clear enough to reflect the logic articulated in the report. We are looking for a combination of insight and clarity when grading the reports.

9 QUESTIONS AND CLARIFICATIONS

For additional questions, please use Piazza, or ask questions during the TAs office hours.