CS 329S: Problem Set 1

Due 11.59PM Thursday, Jan 27, 2022 (PST)

Welcome to the first problem set for CS 329S! We hope you've been having fun with the course so far. If you aren't, come talk with us!

In this problem set, you'll have the opportunity to connect what we've seen in lecture to examples of ML systems "in the wild" and to run some tests to determine the strengths and weaknesses of some of these systems.

Note: This problem set is open-ended. Many of the questions have more than one correct answer. We might still give points to answers that are considered incorrect if you demonstrate thoughtfulness in your approach.

Honor code

You're free to search, read papers, even ask questions on public forums as long as you include all the resources you've used in your submission.

1. If you read it in a paper, cite it.
2. If you ask a question on Quora/Reddit/etc., include a link to it.
3. If you discuss the problem set with other students, disclose who you discussed with, and which problems you discussed with each collaborator.

Submission instructions

Once you've completed the assignment, please upload a PDF with your answers to Gradescope. Please start each problem on a new page.

Working in groups

You can form a working group of two people. If you work in a group, **every person in the group must:**

1. **submit their homework individually and**
2. **mention the name of their partner**.

Late days

Each student is allowed **two 24-hour late days in total** that they can use for any deadline except the final project's demo and end-of-quarter report deadlines.

Written with □ and □ by the CS 329S course staff.

Problem 0 (warm-up): Understanding performance requirements [5 points]

For the following systems, determine which error is worse: a false positive or a false negative. For each case, explain your reasoning in one or two sentences. It's possible that both false positives and false negatives are really bad for your system. If that is the case, you should explain why each is bad and in what situations you would prioritize one over another.

(a) [**1 point**] Spam filtering for emails.

In this case false positive is worse because we don`t want to miss an important emails.

(b) [**1 point**] Spam filtering for search engines.

It known there are people, who wants to trick search engines make them show their spam. In this case false negative will be slight worse because we don`t want a spam (or only spam) in search results.

(c) [**2 points**] Fraud detection for online transactions.

In this case we don`t want to block a valid transactions, because a clients will be dissatisfied. However we don`t want to miss a fraud transactions. Thereby it will be better to work with both these metrics (as alternative – f1-score)

(d) [**1 point**] Classifying skin lesions as benign (negative) or malignant (positive) from a photograph.

In this case false negative is worse because we don`t want to miss a decease. It will be better to repeat tests if something goes wrong.

Analyzing ML Systems

The next two problems will require you to analyze various machine learning systems. Choose three examples from the list below (each from a different category) to use in your responses to these problems.

---

1. **Search:** DuckDuckGo, You.com, Airbnb Search, Twitter Search, Quora Search, Reddit Search.
2. **Recommendation systems:** YouTube homepage recommendations, Spotify Discover Weekly, DoorDash restaurant recommendations, DoorDash item recommendations, Weee! recommendations, Reddit recommendations.
3. **Speech recognition**: Siri, Google Assistant, Google Voice Typing for Google Docs, Alexa, Cortana, etc.
4. **Spam filtering**: Gmail/Outlook spam filtering.
5. **Caption generation**: YouTube auto-caption.
6. **Smart assistants:** Siri, Google Assistant, Alexa, Cortana, etc.
7. **Other conversational AI bots:** Mitsuku, Woebot, Replika, etc.
8. **Machine translation:** Google Translate, DeepL, etc.
9. **Predictive texting**: Phone's predictive texting, autocorrection, Gmail smart replies.
10. **Sentiment analysis:** Google's Natural Language API (scroll down to **Natural Language API demo**), Twinword's Sentiment Analysis, etc.

Problem 1: Understanding objectives and models [10 points + 3 points extra credit]

For each of your chosen systems, answer the following questions.

(a) **[1 point]** List the 3 examples you have chosen from the list above (each should be from a different category). You will analyze them in the following questions.
Spam filtering, smart assistants (Alexa), search (reddit search)

(b) **[6 points - 2 points each system]** Describe some of the metrics you can use to evaluate the ML component of that system. How do you measure these metrics, both during training and inference? Make sure you discuss metrics in the context of the particular application.

(c) **[3 points - 1 point each system]** Describe a baseline method for the system. For instance, for Google Search, a baseline method may rank pages sorted by how many incoming links there are to a page (e.g. using the PageRank algorithm).

(d) **[Extra credit - 3 points - 1 point each system]** Briefly summarize how the system works (e.g. in terms of machine models used or input sources used for training) in 50-200 words. We understand that it might be impossible to determine how a system works under the hood without talking to their engineering team, so the answers won't be graded on the absolute correctness, but on how thorough your consideration is. You are encouraged and expected to have to do some research for this part. If there's a resource

---

that can back your answer up (e.g. publication from the system's developers), please include it. (Hint: searching on Google might help a lot!)

Problem 2: Understanding limitations of ML systems [17 points + 5 points extra credit]

Now comes the fun part - playing around with the systems you've selected to discover their limitations. Feel free to do research on this!

You're welcome to select 3 systems different from the one you've analyzed in Problem 1 - but if you do, please tell us what systems you are choosing at the beginning of your response to this problem.

We'll be focusing on two types of limitations:

- **Edge cases (instance-level):** Individual samples that cause failures in an ML system. Examples:
    a. Siri failed when asked to compute 10 trillion to the power of 1000 minus 1
    b. Google Photos' auto panorama feature failed to blend these ski trip photos and these cliff diving photos correctly.
- **Biases (system-level):** Patterns of inputs that consistently trigger failures in an ML system. Examples (other than examples already shown in Lecture 1, Slide #43):
    a. Gender bias in a translation system.
    b. Racial bias in risk assessment.
    c. Racial bias in speech recognition.

Answer the following:

(a) **[6 points - 2 points each system]** Identify one edge case for each system. Provide evidence of the limitation you have found (e.g. a screenshot, a recording).

(b) [**6 points - 2 points each system**] Putting yourself in the shoes of a machine learning engineer working on this system, how would you go about fixing the errors you identified in 2(a)?

For one chosen system (you can choose any), do the following:

(c) [**5 points**] Identify one system-level error of this system. Provide evidence of the bias you've found (e.g. a screenshot, a recording). You should additionally identify the pattern of failure, and therefore, the class of inputs which reliably produce system failure.

(d) **[Extra credit - 5 points]** Putting yourself in the shoes of a machine learning engineer working on your chosen system, how would you go about fixing the error you identified in 2(c)? You should first propose a hypothesis of the cause of this error, then suggest ways

---

to validate whether this hypothesis is true, and propose a solution to address this.

Problem 3: Bring a research model into production [12 points]

Find a recent (2018 or later) research paper describing a machine learning system which you believe would have potential to be deployed as a system in the real world.

Here are some examples of the papers that you might want to consider. This list is provided to help you with the search. There's no difference in grading whether you choose a model from this list or somewhere else:

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2018)
- DALL·E: Creating Images from Text (OpenAI, 2021) (I can't get over these avocado chairs)
- CTRL: A Conditional Transformer Language Model for Controllable Generation (Keskar et al., 2019)
- GauGAN: Semantic Image Synthesis with Spatially-Adaptive Normalization (Park et al., 2019)
- SEQUENCER: Sequence-to-Sequence Learning for End-to-End Program Repair (Chen et al., 2018)
- Through-Wall Human Pose Estimation Using Radio Signals (Zhao et al, 2018)
- GPT-3 (Language Models are Few-Shot Learners) (Brown et al., 2020)
- CLIP: Connecting Text and Images (Radford et al., 2021)
- Evaluating Large Language Models Trained on Code (Chen et al., 2021)
- AV-HuBERT: Audio-Visual Hidden Unit BERT (Meta AI, 2022)

If you want to look for more papers, we'd recommend checking out the following sources:

- Machine Learning Conferences: NeurIPS (formerly NIPS), ICML, ICLR, MLSys, RecSys.
- NLP Conferences: ACL, NAACL, EMNLP.
- Computer Vision Conferences: CVPR, ICCV.

(a) **[3 points]** Provide a link to the paper you've found, and summarize, in no more than 150 words, the relevant background, approach, and findings of the paper.

(b) **[1 point]** Briefly describe, in no more than 100 words, why the model presented in the paper is a good candidate for being brought into production: what compelling use case(s) do you envision such a system would fulfill, for which segment(s) of users? Users can be both individuals and enterprises. We expect only one compelling use case, but welcome it if you can provide more.

(c) **[5 points]** Pick one use case you've identified in 3(b). Imagine that you are a consultant advising a company on how to incorporate this paper into this use case. Drawing on what you learned from the first two lectures, list and describe the three most important considerations for this company to consider. Your considerations should demonstrate a strong understanding of the project/paper and the use case you described in 3(b); answers which discuss high-level points of consideration from the course slides without grounding in the context of the project you have selected will be awarded minimal credit. (Hint: the breakout exercise from Lecture 2 could be helpful!)

(d) **[3 points]** Given that the team does not have the resources to develop a full end-to-end machine learning system in the immediate future, describe, in no more than 150 words, a minimum viable product (MVP) they might use to test their initial assumptions that this model can help with the specified use case.

Problem 4: Handling missing values [10 points + 5 points extra credit]

Download the Titanic dataset here. This is a modified version of Kaggle's **Titanic - Machine Learning from Disaster** dataset. **Please do NOT download the dataset directly from Kaggle**.

The data schema is as follows:

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

---

There are four columns that have missing values in this dataset. **For each of the columns with missing values, answer the following questions**:

1. [**6 points**] What type of missing values does the column have: missing not at random, missing at random, or missing completely at random? If MAR, specify which variable the missing values depend on. Please justify your answers by explaining what method(s) you used to determine the types of missing values. Hint: heatmaps and plots may help visualize relationships among columns.
2. [**2 points**] Explain what might have caused the values to be missing.
3. [**4 points**] Explain in detail how you would handle the missing values? For example, if you use imputation, describe the value you'll impute the missing values with and why. Note that there may be more than one right answer so it's important to justify your answer.

For at least one column with missing values, do the following:

4. [**Extra credit - 5 points**] Experiment with at least two different ways of handling missing values. For each way, train a machine learning model on the train set to predict survival on a held-out test set. How do these different methods of handling missing values influence your model performance? Is that what you have expected?

It's possible to write code to answer question 4 of this problem. If you write code, screenshot the code and add the screenshots as images to your PDF.