XCS229ii: Project overview and proposal

March 2021

Stanford | ONLINE

Sebastian Hurubaru

Munich, Germany

hurubaru@stanford.edu

Agenda

Agenda

- Project overview

Agenda

- Project overview
- Project proposal

Agenda

- Project overview
- Project proposal
- Example of project proposal

Agenda

- Project overview
- Project proposal
- Example of project proposal
- Q&A

Project overview

Project overview

- Open-ended task

Project overview

- Open-ended task
- Bounded methods

Project overview

- Open-ended task
- Bounded methods
- Teams of up to four members

Project overview

- Open-ended task
- Bounded methods
- Teams of up to four members
- Cloud computing resources

Project overview

- Open-ended task
- Bounded methods
- Teams of up to four members
- Cloud computing resources
- Leverage the course staff

Project overview - milestones

Project milestone	Due date	Points
Proposal	Sunday, April 11	10

Project overview - milestones

Project milestone	Due date	Points
Proposal	Sunday, April 11	10
Literature review	Sunday, April 25	20
Slack post shareout (optional)	Sunday, April 25	3 (extra credit)

Project overview

Project milestone	Due date	Points
Proposal	Sunday, April 11	10
Literature review	Sunday, April 25	20
Slack post shareout (optional)	Sunday, April 25	3 (extra credit)
Experimental protocol	Sunday, May 9	20

Project overview

Project milestone	Due date	Points
Proposal	Sunday, April 11	10
Literature review	Sunday, April 25	20
Slack post shareout (optional)	Sunday, April 25	3 (extra credit)
Experimental protocol	Sunday, May 9	20
Final paper	Sunday, May 23	50
Project Draft Presentation or Video (optional)	Sunday, May 23	5 (extra credit)

Project proposal

Project proposal – guidelines

Project proposal – guidelines

- What should your proposed system do?

Project proposal – guidelines

- What should your proposed system do?
- What dataset(s) are you planning to use?

Project proposal – guidelines

- What should your proposed system do?
- What dataset(s) are you planning to use?
- What are the inputs and outputs of your system?

Project proposal – guidelines

- What should your proposed system do?
- What dataset(s) are you planning to use?
- What are the inputs and outputs of your system?
- How do you plan to assess the results of your system?

Project proposal – guidelines

- What should your proposed system do?
- What dataset(s) are you planning to use?
- What are the inputs and outputs of your system?
- How do you plan to assess the results of your system?
- What might be the challenges of building the system?

Project proposal – guidelines

- What should your proposed system do?
- What dataset(s) are you planning to use?
- What are the inputs and outputs of your system?
- How do you plan to assess the results of your system?
- What might be the challenges of building the system?
- Which topics might address these challenges?

Project proposal - Examples

Examples

- Predicting Mechanisms of Action for Drug Discovery

Examples

- Predicting Mechanisms of Action for Drug Discovery

- OTC Derivatives Trade Reporting Advanced Analytics

Example 1: What does the proposed system do?

Unlike the serendipitous discovery of drugs in the past, modern drug development has adopted a more targeted approach. The proteins influenced by a pathology are identified and molecules which could influence those proteins are developed in the lab. This process involves studying how a drug impacts gene expressions (genes activated by the drug) and cell viability (amount of healthy cells at the end of the experiment), and the consequent molecular effects they elicit. In literature, the biological activity of a molecule is also referred to as Mechanism of Action (MoA).

This project tries to learn the relation between the gene expressions and cell viability data to the MoA, by using data provided on approximately 5000 drugs and their corresponding gene expression and cell viability data. The dataset contains samples derived from lab experiments under two different conditions: dosage of the drug (high or low) and time of exposure to the drug. The gene expressions and the cell viability results of these experiments are then captured and provided in a normalized form in the data.

Example 2: What does the proposed system do?

Problem Background

Since the 2008 Global Financial Crisis, financial services institutions (FSIs) globally have increasingly been required to report their OTC financial derivative transactions to centralized trade repositories, to help regulators assess the risk of the next global financial crisis. Most of the industry effort since that time has been focused on the implementation of systems integration from FSIs' numerous source data platforms into the trade repositories and rudimentary data quality analysis by regulators, with relatively little effort invested toward more advanced analysis of the data. This project seeks to explore new approaches to making value-add use of the data through machine learning techniques.

Example 1: Dataset

Scientists from the Laboratory for Innovation Science at Harvard proposed this task as a kaggle competition and made a dataset to be used specifically for this purpose. Only 25% of this dataset is available to the participants for model development, the rest are reserved for testing. The accessible part of the dataset is split into train and test samples, train targets scored (targets that we have to predict) and train targets non-scored (labels that don't have to be predicted but that contain additional MoAs for the same experiments). Training data contains experiments for the 5000 drugs under a total of 7 different conditions (2 dosages, 3 ranges of time, a baseline without the drug). However, a part of the experiments was discarded by the scientists and the total amount of rows is about 24.000. Each experiment (row) comes also with two sets of features, one for each factor previously described.

Example 2: Dataset

We considered the use of proprietary data from one of the team member's employer (IHS Markit), which hosts a trade reporting platform. However, given the challenges of normalizing, anonymizing, and provisioning such data, we choose to make use of the publicly disseminated trade reporting data provided by one of the leading global trade repositories (DTCC), available at the following URL:

https://rtdata.dtcc.com/gtr/dashboard.do

This DTCC data set provides for real-time published data for the entire U.S. market for the most recent 30 days of trading activity. Transaction data counts across the 5 core financial asset classes, over the period 24-Sep-2020 to 24-Oct-2020, follows:

- **Equities**: 1,950,241
- **FX**: 402,817
- **Rates**: 194,365
- **Commodities**: 118,773
- **Credit**: 27,814

Example 1: What are the inputs and outputs?

Input/Output behavior of the system

From the data set, we partition into training, validation (and test sets). From the training set, we build an NN model. We will tune the model parameters [layer weights and activations] with the validation set. Once the model is trained and tuned, we can use this to predict the multi-label classifications on the test sets. After iterating the model offline we will submit the predictions to Kaggle MoA's[3] challenge so our predictions are scored with the private dataset of the competition

Example 2: What are the inputs and outputs?

System Description – Inputs, Outputs, and Phenomena to be Captured

We propose a multi-phase, progressive approach to building the system.

Phase 1: Supervised predictive analytics on transaction data

This phase will use supervised learning to seek to "predict" certain outcomes of future reporting data, at different levels:

Phase 2: Unsupervised feature augmentation

This phase will attempt to increase the predictive accuracy of the models developed in Phase 1 by use of unsupervised algorithms to augment the data set with additional engineered features.

Example 1: How will you assess the results?

The results will be evaluated running a k-fold cross validation. The principal monitored metric is the Binary Cross Entropy Loss, as it is the metric that Kaggle computes on the private test set. In order to get further terms of comparison on how the model performs, label ranking loss will be considered in addition to the afore mentioned one.

Example 2: How will you assess the results?

Measures of Success

We will treat the "MVP" of the project to be a well-grounded assessment of the presence of a predictable signal in the chosen data, for one or more of the measures outlined in Phase 1 above. Time permitting and based on the results of each prior phase, we may choose to attempt subsequent phases in turn.

Example 1: Challenges

The main challenge that the built model needs to address is learning the hidden patterns in the input features, which could be firing some of the MoAs simultaneously. One factor that can hinder the learning process is the fact that some MoAs fire quite infrequently and, as a result of this, the features influencing those MoAs might not be adequately learned. Now considering the fact that the same drug was employed to carry out several experiments with duration and dosage altered, if those are not responding differently to MoAs, that could make the corresponding experiments redundant, essentially limiting the number of independent experiments from which the model can learn.

Example 2: Challenges

Implementation Challenges

Unlike the proprietary data set that would otherwise be used with potentially hundreds of features, the publicly disseminated data set has far fewer features (44) due to its aggregated, anonymized nature. Therefore, the likelihood of predictive power in the data could be reduced. To address this challenge, we propose starting the implementation with simple models, and progressively working through more complex models, including unsupervised feature generation, to adjust the bias variance tradeoff. If this approach does not prove fruitful, then we will explore the possibility of acquiring additional data.

Another challenge will be for certain analytics (such as predicting cancel/correct actions), the data will be largely imbalanced, with only a small proportion of the total data set actually being subject to cancel/correct. To address this challenge, we propose exploring a number of possible remediations, including resampling, proper use of cross-validation, and clustering of the abundant class.

Example 1: What might address these challenges?

An important aspect that we plan to investigate is data preprocessing to get rid of the imbalances in the label representations. Techniques such as stratification [sampling proportional to the original data set, for the training data], and oversampling of the lower classes so that more information can be gained for lower represented classes. There are many methods available to oversample a dataset and that are used in a typical classification problem, such as Synthetic Minority Over-sampling Technique (SMOTE) and the Adaptive Synthetic sampling approach (ADASYN), which builds on the methodology of SMOTE. The end-result of oversampling is the creation of a balanced dataset that will allow neural networks to make more reliable predictions from being trained with balanced data. We plan to investigate some NN architectures that are "sample efficient" i.e. the neural networks may not require many examples of a new class to understand the hidden features present.

Example 2: What might address these challenges?

Implementation Challenges

Unlike the proprietary data set that would otherwise be used with potentially hundreds of features, the publicly disseminated data set has far fewer features (44) due to its aggregated, anonymized nature. Therefore, the likelihood of predictive power in the data could be reduced. To address this challenge, we propose starting the implementation with simple models, and progressively working through more complex models, including unsupervised feature generation, to adjust the bias variance tradeoff. If this approach does not prove fruitful, then we will explore the possibility of acquiring additional data.

Another challenge will be for certain analytics (such as predicting cancel/correct actions), the data will be largely imbalanced, with only a small proportion of the total data set actually being subject to cancel/correct. To address this challenge, we propose exploring a number of possible remediations, including resampling, proper use of cross-validation, and clustering of the abundant class.

Key takeaways

Key takeaways

- Keep it short and simple

Key takeaways

- Keep it short and simple
- Follow all the guidelines

Key takeaways

- Keep it short
- Free of choice format
- Team-work is fun

Key takeaways

- Keep it short
- Free of choice format
- Team-work is fun: Project Team/Idea Mixer - April 3rd, April 5th

Q&A