XCS229ii Experimental Protocol

April 28, 2021

Stanford | ONLINE

Agenda

Experimental Protocol - Agenda

- General information

Experimental Protocol - Agenda

- General information
- Sections

Experimental Protocol - Agenda

- General information
- Sections
- Examples

Experimental Protocol - Agenda

- General information
- Sections
- Examples
- Q&A

Experimental Protocol - General Information

Experimental protocol - Description

Experimental protocol - Description

- Designed to help you create your core experimental framework

Experimental protocol - Description

- Designed to help you create your core experimental framework
- Implement any potential extensions identified in the previous milestone

Experimental protocol - Description

- Designed to help you create your core experimental framework
- Implement any potential extensions identified in the previous milestone
- Hands-on practice with evaluating the models using ML theory

Experimental protocol - Info

Experimental protocol - Info

- Due on Sunday, May 9th

Experimental protocol - Info

- Due on Sunday, May 9th
- Worth 20 points, i.e. 20% out of the total project score

Experimental protocol - Info

- Due on Sunday, May 9th
- Worth 20 points, i.e. 20% out of the total project score
- 5 – 6 pages document in a free of choice format

Experimental protocol - Info

- Due on Sunday, May 9th
- Worth 20 points, i.e. 20% out of the total project score
- 5 – 6 pages document in a free of choice format
- Submit on Gradescope as a group

Experimental protocol - Info

- Due on Sunday, May 9th
- Worth 20 points, i.e. 20% out of the total project score
- 5 – 6 pages document in a free of choice format
- Submit on Gradescope as a group
- Cloud resources

Experimental protocol - Info

- Due on Sunday, May 9th
- Worth 20 points, i.e. 20% out of the total project score
- 5 – 6 pages document in a free of choice format
- Submit on Gradescope as a group
- Cloud resources
- Leverage the course staff

Experimental Protocol - Sections

1. Hypotheses

1. Hypotheses

- What is the problem to be addressed?

1. Hypotheses

- What is the problem to be addressed?
- Why is the problem important and why is it hard?

1. Hypotheses

- What is the problem to be addressed?
- Why is the problem important and why is it hard?
- What are the key limitations of prior work (with associated references)?

1. Hypotheses

- What is the problem to be addressed?
- Why is the problem important and why is it hard?
- What are the key limitations of prior work (with associated references)?
- What are opportunities to improve upon existing work and possible experiments to explore?

2. Data

2. Data

- Describe datasets used and/or procedure for collecting data (if applicable)

2. Data

- Describe datasets used and/or procedure for collecting data (if applicable)
- Typically good to spend short amount of time (1 ~ 3 days) gathering data

2. Data

- Describe datasets used and/or procedure for collecting data (if applicable)
- Typically good to spend short amount of time (1 ~ 3 days) gathering data
- Plot data distributions, input representations, feature space, labels, etc.

3. Metrics

3. Metrics

- Describe the metrics that form the baseline of evaluation

3. Metrics

- Describe the metrics that form the baseline of evaluation
- Touch at least 2 of the following ML Theory frameworks:
  - Error Analysis Diagnostics
  - Ablative Analysis
  - Approximation and Estimation Errors
  - Bias-Variance Diagnostic
  - Optimization Diagnostics

3. Metrics

- Describe the metrics that form the baseline of evaluation
- Touch at least 2 of the following ML Theory frameworks:
  - Error Analysis Diagnostics
  - Ablative Analysis
  - Approximation and Estimation Errors
  - Bias-Variance Diagnostic
  - Optimization Diagnostics
- For RL based projects, find a way to illustrate the performance of your system

4. Models

4. Models

- Describe your baseline model

4. Models

- Describe your baseline model
- Describe the model(s) you will be focusing on

4. Models

- Describe your baseline model
- Describe the model(s) you will be focusing on
- Preliminary description is sufficient

5. General Reasoning

5. General Reasoning

- Explain how the data and model(s) come together and inform your core hypothesis

5. General Reasoning

- Explain how the data and model(s) come together and inform your core hypothesis
- Analyse preliminary results

6. Summary of progress so far

6. Summary of progress so far

- What have you managed to do so far?

6. Summary of progress so far

- What have you managed to do so far?
- What still needs to be done?

6. Summary of progress so far

- What have you managed to do so far?
- What still needs to be done?
- Describe the obstacles/concerns

6. Summary of progress so far

- What have you managed to do so far?
- What still needs to be done?
- Describe the obstacles/concerns
- Report results for finished models

7. References

7. References

- List papers using a commonly accepted reference format

7. References

- List papers using a commonly accepted reference format
- Formats: https://www.bibguru.com/blog/citation-style-for-computer-science

Experimental Protocol - 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: Hypotheses

Based on the effectiveness that neural networks have already proved in the healthcare field [7][2], all the considered hypotheses will try to employ a suitable NN architecture given the data distribution and address two main recurring issues that affect the data available for learning purposes in the context of multi-label classification:

- Imbalance in the labels. When it is necessary to predict more than one label per sample in a classification problem where labels are imbalanced, the application of resampling techniques can be really tricky since data is affected by label co-occurrence. In order to deal with this problem it is necessary to find architectures or algorithms that can leverage on label co-occurrence to enhance the learning outcome.

- Presence of negative samples. Around 40% of our samples are baselines reading of the tool they use to measure the drug impact (gene expression and cell viability). We need to be able to find a way to take advantage of this group of samples.

Specifically, we have identified several tasks that could serve as our hypotheses to tackle each issue, as potential techniques that looked promising, some of which address single problems, others address multiple problems at a time:

K-Fold Stratification: To address the higher incidence of label co-occurrence, some changes to how the data is split into training samples which consider label co-occurrence were identified and implemented.

Example 2: Hypotheses

This project seeks to explore new approaches to making value-add use of publicly reported over-the-counter (OTC) derivatives regulatory trade reporting data through advanced analytics using machine learning (ML) techniques.

Our hypothesis is that we can use machine learning to predict, with high degrees of accuracy, whether a new trade report will be subsequently cancelled/corrected by the reporting counterparty, within a specified timeframe.

Example 1: Data

2.1 Summary:

Amount of labels: 207
Amount of features: 876
Dataset size: 23846 samples

Samples with zero positive zero labels: 9367 samples
Samples with one positive label: 12532 samples

Samples with multiple positive labels: 1915 samples

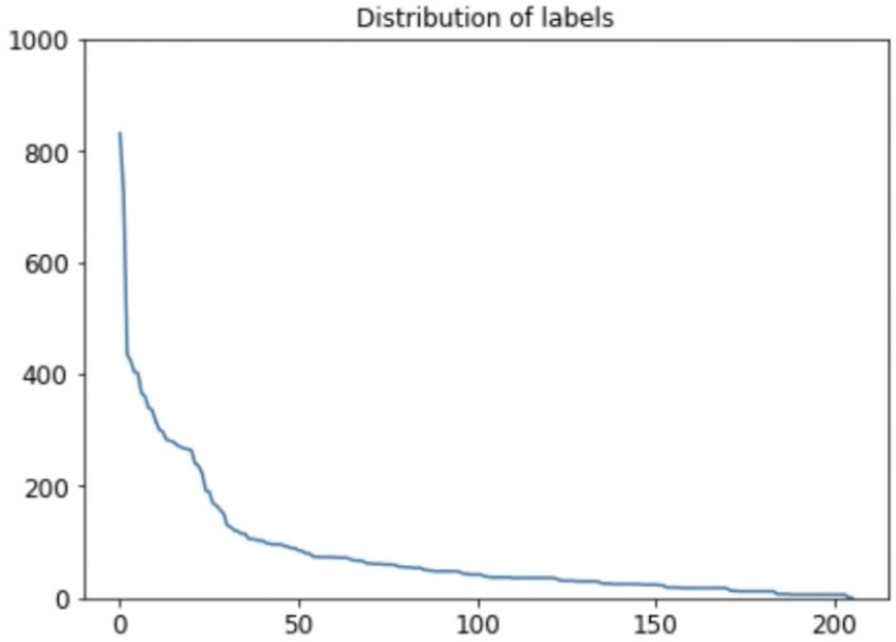

Figure 1: X axis - Label index, Y axis - Amount of samples for the given label

Samples which contain labels that have less than 7 samples: 126 samples (we aim to help the classification of this samples with few-shot learning)
Samples which contain labels that only occur in one combination: 12819 samples (we aim to help the classification of this samples with label co-occurrence techniques)
Unique combinations of labels: 328 combinations

Example 2: Data

The raw data contains 44 separate columns, with the columns, datatypes, and high-level descriptions as follows:

Name	Datatype	Description
DISSEMINATION_ID	int64	Unique identifier for the record.
ORIGINAL_DISSEMINATION_ID	Int64	Cross-reference to the unique identifier for the associated record, where the current record is a cancel or correct of a previous record.
ACTION	category	An indication that a record is a new trade, cancel, or correct record.
EXECUTION_TIMESTAMP	datetime64[ns]	The time and date of execution of the trade in UTC.

Given our hypothesis seeks to predict whether a new trade report will be subsequently cancelled/corrected within a specified timeframe, we consider the data across several different time horizons between 2 and 60 days. We observe the following volumes of data, and take particular note of the imbalanced nature of the positive class compared to the total number of records, which is consistently between 3% and 5% of the total volume:

Start Date	End Date	# Days	# Records	# New Canceled/Corrected	Pos Class %
19-Nov-2020	21-Nov-2020	2	20,321	703	3.5%
14-Nov-2020	21-Nov-2020	7	48,073	2,158	4.5%
7-Nov-2020	21-Nov-2020	14	90,967	4,436	4.9%
22-Oct-2020	21-Nov-2020	30	196,845	9,678	4.9%
22-Sep-2020	21-Nov-2020	60	384,073	18,718	4.9%

We also note that a large proportion of the columns contain blank values.

Example 1: Metrics - 1

3.1 Overall system Metrics

The metrics will be computed using a K-fold validation scheme. The main metric we will monitor will be Binary Cross Entropy Loss as it is the metric Kaggle computes with our predictions on the private test set. This metric will allow us to confirm that our validation scheme is adequate to evaluate how the model performs on completely unseen data (data that was not used to train in any of the folds).
Further metrics will be used to differentiate how the model performs. The complete set of metrics will be:

- **Binary cross entropy loss:** Loss function that will be used to train the models and it will be the metric reported by Kaggle in the test set.
- **Coverage error[3]:** It calculates how far down the ordered array of predictions are the actual correct labels
- **Label Ranking loss[3]:** Computes the average number of label pairs that are incorrectly ordered weighted by the number of labels in the correct label set.

Example 1: Metrics - 2

3.2 Siamese neural network

The siamese network will be trained with Triplet loss and it will be the main metric we will monitor during the training procedure. To be able to test all the different options for creating Triplets a validation set will be created that will be used to evaluate the different approaches to Triplet Mining.

Apart from the metrics and validation used to measure the training of the Siamese neural network. After training the embeddings will be evaluated by how good the relative distances determine which labels the embedding should have.

One important part of our dataset consists of negative samples. These are measurements of features that do not have any expected MoA and are supposed to help provide baseline measurements for when the drug is added to the experiment.

In our ablation studies we will explore training the model with and without this part of the dataset to learn the impact these samples have in the training process. Another point of our ablation studies will focus on the different approaches to feature engineering such as PCA, Triplet Embeddings, Normalization of inputs

Example 2: Metrics

We plan to use the following metrics, at a minimum, to compare the results of our different models:

- Ablative Analysis
  - Given the breadth of features, we plan to analyze how the successive removal of features impacts the performance of the models.
- Approximation and Estimation Errors
  - In particular, we plan to use accuracy, precision, and f1 scores as a metric of the relative performance of our models. Given the imbalanced nature of the data, we plan to pay particular attention to these metrics at the individual class level.
- Bias-Variance Diagnostic
  - We plan to analyze the tradeoff between bias and variance at each stage of model selection, to examine the tradeoffs between overfitting for more complex models, versus underfitting for less complex models. In particular, we plan to explore the use of various regularization features available in the chosen models.

Example 1: Models - 1

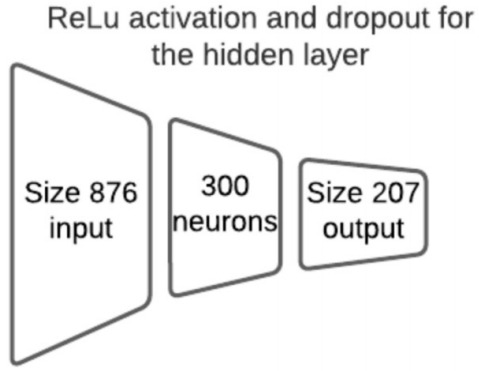

Figure 2: Baseline architecture

4.1 Baseline

The chosen model that will be used as a baseline is a simple feed forward neural net, with two layers. Input linear layer, with ReLU activation, between inputs and hidden nodes; an output layer, with sigmoid activation, between hidden nodes and the outputs. Given that the problem being addressed is a multi label problem, the loss used for the model was binary cross entropy loss. This model was used along with basic K-fold validation framework , along with learning rate decay and early stop based on validation loss metric.

Example 1: Models - 2

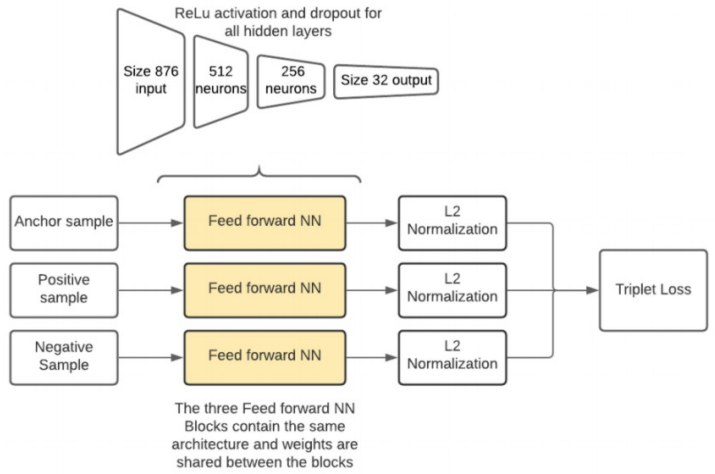

Figure 3: Architecture of neural network with Triplet loss.

4.2 Siamese Neural Networks with Triplet Loss

The chosen backbone for the Siamese Neural network is a simple 2-layer feed forward neural network with dropout and batch normalization. At the end of the backbone the vector is normalized with L2 norm to restrict the vectors to a module of 1[4].

After training the Siamese neural networks the embeddings will be fed to the Baseline network as extra features. [5]

There is still work to do in choosing the method to "mine triplets". There are several options in the literature on how to choose the (anchor, positive, negative) triplets to speed up the training such as offline triplet mining, online triplet mining (with variants such as batch-hard, batch-semi hard)

Example 2: Models

For our baseline model, we will use supervised learning, and will focus initially on logistic regression.

In order to assess the relative performance of models, we will then work through other models including SVM, neural networks, generative learning algorithms, decision tress, and neural networks.

Depending on the results from the supervised models and time permitting, we may attempt to supplement the models with additional cluster features using k-means or other unsupervised learning techniques.

Example 1: General reasoning

After initial analysis of the data and preliminary results on the model, few areas where an improvement can be achieved were tentatively identified:

- Techniques to improve neural network performance with high label co-occurrence
- Techniques to address the imbalance in the dataset, where samples for certain labels for some labels were much higher than others
- Techniques for identifying compact sets of features which could then be used to further improve the neural network performance.

Also evaluating the possibility of creating meaningful embeddings from the features to later feed them to the neural network has a dual objective. It may be a good approach to close the gap we have in the domain knowledge for the feature pre-processing and also help at the task of few-shot learning for the labels that have low amount of samples. In addition to the benefits that this approach may give us to perform the task of MoA prediction we think embeddings of biological features can be useful for other types of research.

Example 2: General reasoning

The data and the models should come together well to inform our core hypothesis. In particular:

- The data is not actually labeled for supervised learning of this particular problem, but reference ids enable records to be joined together to derive labels for learning
- The data is described as a set of features (columns) so therefore is well adapted to supervised learning algorithms
- Although the data is imbalanced, there are mechanisms to deal with the issue (i.e., over/under-sampling)

In addition to model choice and parameter tuning, will need to experiment with a variety of sample periods and features to find the optimal metrics.

Example 1: Summary of progress so far

We started by implementing the techniques onto the baseline model with a simpler set of metrics than the one proposed on the metrics section. After some experimentation we found the necessity to have a more complex set of metrics that provide fine-grained information for the different clusters of labels. Next steps include re-running the experiments with the new set of metrics to understand if the improvements impact specific sets of labels

K-fold Framework: For splitting the data into K folds, considering the data, we have identified a need for balancing the label co-occurrence pattern between train and validation sets, in addition to balancing the labels between the two sets. Since we have only about 334 unique patterns of label power sets, and

Label Co-occurrence Optimization: The implementation of this code, first involves studying the data and identifying all the co-occurring label patterns, and depending on the number of co-occurring labels a portion of the hidden nodes are classified as reserve nodes. During the initialization step, after the

Example 2: Summary of progress so far - 1

So far, we have:

- Developed python programs leveraging pandas to:
  - Read data and organize into appropriate data types
  - Label data using join mechanisms
  - Oversample the minority class to balance the data
  - For ease of initial implementation, reduce the initial set of columns to a set of categorical features only
  - One-hot encode the data

Additionally, we have managed to setup the system on Azure Labs. There, we have leveraged scikit-learn to test logistic regression, on a variety of sample periods, with the following results:

- 2 Days

	precision	recall	f1-score	support
False	0.93	0.77	0.84	4909
True	0.80	0.94	0.87	4900
accuracy			0.85	9809
macro avg	0.86	0.85	0.85	9809
weighted avg	0.86	0.85	0.85	9809

Example 2: Summary of progress so far - 2

We have noted that, with additional data over time, all other conditions held equal, the performance appears to degrade. This warrants further exploration; it may be an indication of data "drift" over time reducing the predictive power.

Left remaining to do, we have:

- Include additional floating and datetime features, with timeseries handling
- Add scaling
- Tune logistic regression parameters
- Explore additional supervised models
- Explore unsupervised models for feature generation

Example 1: References

References

[1] Sercan O. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning, 2020.

[2] A. Maxwell, R. Li, B. Yang, H. Weng, A. Ou, H. Hong, Z. Zhou, P. Gong, and C. Zhang. Deep learning architectures for multi-label classification of intelligent health risk prediction. *BMC Bioinformatics*, 18(Suppl 14):523, 12 2017.

[3] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

[4] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[5] Swati, G. Gupta, M. Yadav, M. Sharma, and L. Vig. Siamese networks for chromosome classification. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 72–81, 2017.

Example 2: References

References

Category	Authors	Title	Publication	Year
Imbalanced data	Ahmed, Mohiuddin Mahmood, Abdun Naser Islam, Md Rafiqul	A survey of anomaly detection techniques in financial domain	Future Generation Computer Systems	2016
Advanced analytics	Buehler, Hans Gonon, Lukas Teichmann, Josef Wood, Ben	Deep hedging	Quantitative Finance	2019
Data	DTCC	Real-Time Dissemination Dashboard	https://rtdata.dtcc.com/gtr/dashboard.do	2020
Imbalanced data	Haixiang, Guo Yijing, Li Shang, Jennifer Mingyun, Gu Yuanyue, Huang Bing, Gong	Learning from class-imbalanced data: Review of methods and applications	Expert Systems with Applications	2017
Data	Legal Information Institute	17 CFR Appendix A to Part 43 - Data Fields for Public Dissemination	https://www.law.cornell.edu/cfr/text/17/appendix-A_to_part_43	2012
Time series	Lu, Chi-Jie Lee, Tian-Shyug Chiu, Chih-Chou	Financial time series forecasting using independent component analysis and support vector regression	Decision support systems	2009
Advanced analytics	Ma, Xun Spinner, Sogee Venditti, Alex Li, Zhao Tang, Strong	Initial Margin Simulation with Deep Learning	Available at SSRN 3357626	2019
Data	National Archives	Real-Time Public Reporting of Swap Transaction Data	https://www.federalregister.gov/documents/2010/12/07/2010-29994/real-time-public-reporting-of-swap-transaction-data	2012
Imbalanced data	Yen, Show-Jane Lee, Yue-Shi	Cluster-based under-sampling approaches for imbalanced data distributions	Expert Systems with Applications	2009

Experimental Protocol - Summary

Experimental protocol - Key takeaways

Experimental protocol - Key takeaways

- Cover all sections

Experimental protocol - Key takeaways

- Cover all sections

- Start early

Experimental protocol - Key takeaways

- Cover all sections
- Start early
- Use the cloud resources mindfully

Experimental protocol - Key takeaways

- Cover all sections
- Start early
- Use the cloud resources mindfully
- Don’t hesitate to ask for help

Q&A