“Hands-on Machine Learning Lab Report”

Jeheon Park, Life Science Informatics in B-it.

The lab report is written by Jeheon Park, but the lab experiments collaborate with Daria, Nicolas in the Same Master program. Therefore, most of the results are shared.

Introduction

This report has two results from the lab. The first result is about sentiment analysis with the BERT model on Drug review datasets.[1] We built a model to predict the sentiment of the drug review. The second result is about survival analysis with RSF and ANN on the breast cancer datasets from the University of Wisconsin.[2] We built a model to predict time till the events, breast cancer occurs in our case, with other features.

Theoretical Background

- Text Mining

Text mining is to automatically extract the information from the bulk of structured or unstructured text datasets. It catches the information patterns in the datasets. It considers frequency counts of words, length of the sentence, and the presence/absence of certain words to extract the information. Therefore, it needs text cleaning(HTML, irrelevant number, and stop words filtering, etc.), tokenization, lemmatization or stemming, vectorization (Word2Vec, TFIDF, etc.), and building a proper model to process vectorized data. Text Mining is often a precursor of the NLP. NLP considers what meaning the text conveys, for example, sentiment, keywords, and the datasets for NLP can be speeches or text. In our case, we use text mining as a precursor to analyze the sentiment of the review.

- Sentiment Analysis

Sentiment analysis is a machine learning technique that detects polarity (e.g. a positive or negative opinion) within the text, whether a whole document, paragraph, sentence, or clause.[3] Sentiment analysis is widely applied to healthcare materials for applications that range from marketing to customer service to clinical medicine.[4]

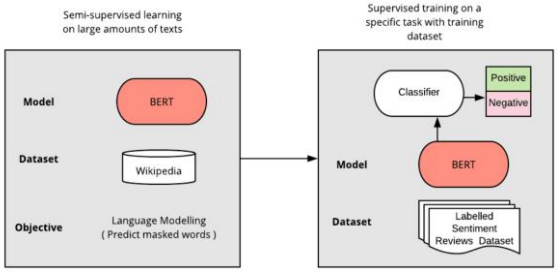

Fig.1 This is the model of sentiment analysis we used in this lab course.[5]

Sentiment analysis uses various Natural Language Processing (NLP) methods and algorithms. The main types of algorithms are rule-based, automatic, and hybrid systems.

- **Rule-based systems** that perform sentiment analysis based on a set of manually crafted rules.
- **Automatic systems** that rely on machine learning techniques to learn from data.
- **Hybrid systems** that combine both rule-based and automatic approaches.

We use a pre-trained BERT model with classifier layers and the datasets have ratings and we transform the ration into positive, neutral, and negative.

- F1-score

F1-score is the harmonic mean of precision and recall.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

Precision represents the proportion of items – in this case, sentiment – that the accurately correct system returns.

$$Precision = \frac{|true\ positive|}{|true\ positive| + |false\ positive|}$$

Recall indicates how much of all items that should have been found, were found.[6]

$$Recall = \frac{|true\ positive|}{|true\ positive| + |false\ positive|}$$

We can set the threshold in the favor of precision or recall, it is called pay-off of precision and recall. Therefore, it is not good measurements for evaluations. On the other hand, F1-score is widely used to evaluate the model because it guarantees the middle point of precision and recall pay-off region.

- Survival Analysis & time to event modeling

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as a death in biological organisms and failure in mechanical systems.[7] Time to events modeling relaxes this assumption. It considers when the events occur or do not occur(non-censored or censored).

The typical example of survival analysis in cancer analysis. Our analysis is also about breast cancer but we don't use death as events. In cancer, another important measure is the time between response to treatment and recurrence or relapse-free survival time (also called disease-free survival time). Most survival analyses in cancer journals use some or all of Kaplan–Meier (KM) plots, log-rank tests, and Cox (proportional hazards) regression.[8]

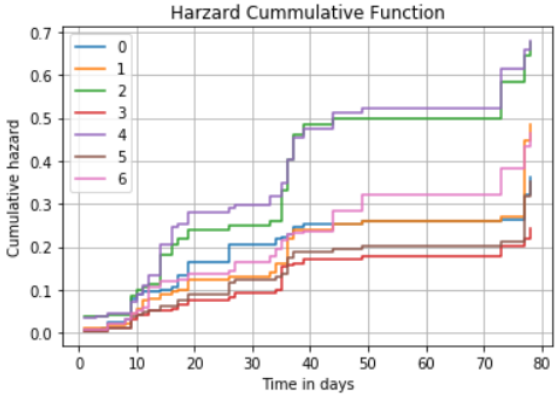

Fig.2 This is the cumulative hazard from the RSF mode and it predicts test datasets.

We use Cox regression. Cox regression is to estimate the hazard at the time with other features. The hazard is usually denoted by h(t) or l(t) and is the probability that an individual who is under observation at a time t has an event at that time. The model approximates the linear regression in the power of exponential. Fig.2 shows the cumulative hazard from our RSF model.

$$h(t; z_i) = h_0(t)\exp(z_i\beta)$$

- Concordance Index

One of the most popular performance measures for assessing learned models in survival analysis is the Concordance Index (CI). It does not measure the actual differences in values. It evaluates the order of pairs. It can be interpreted as the fraction of all pairs of subjects whose predicted survival times are correctly ordered among all subjects that can be ordered.[9] Therefore, even if the predicted values are all wrong, the concordance index can be 1, the maximum value. Casting survival analysis as a ranking problem is an elegant way of dealing not only with the typically skewed distributions of survival times but also with the censoring of the data.

Data

Drug Review Dataset (Drugs.com) Data Set[1] is used for Sentimental Analysis in this report. The data is from Drugs.com. The dataset provides patient reviews on specific drugs along with related conditions and a 10-star patient rating reflecting overall patient satisfaction. The attributes of the datasets are the name of the drug, name of the condition, patient review, 10-star patient rating, date of review entry, and the number of users who found the review useful.

Breast Cancer Wisconsin (Prognostic) Data Set[2] is used for survival analysis. Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Dr. Wolberg since 1984 and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. The attributes are ID number, Outcome (R = recur, N = nonrecur), Time (recurrence time if outcome = R, disease-free time if outcome = N), Ten real-valued features are computed for each cell nucleus: radius (mean of distances from the center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" - 1)

Machine Learning Methods

BERT

BERT is a method of pre-training language representations, meaning that we train a general-purpose "language understanding" model on a large text corpus (like Wikipedia).[10] BERT uses Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text.[11] The encoder is only needed in the Transformer and it is bidirectional to understand the contextual relations. BERT can be used for a wide variety of language tasks. In this report, we used [CLS] tokens and add a classification layer on the top of the output of the pre-trained model because the sentiment analysis is a classification problem.

RSF

Random Survival Forest is an ensemble tree method for the analysis of right-censored survival data.[12] The random survival forests (RSF) methodology extends Breiman's random forests (RF)[13] method. A random survival forest consists of random survival trees. Using independent bootstrap samples, each tree is grown by randomly selecting a subset of variables at each node and then splitting the node using a survival criterion involving survival time and censoring status information.

Cox Regression

$$h(t; z_i) = h_0(t)\exp(z_i\beta)$$

We also make a linear regression model with an elastic net as a regularizer. This model approximates the parameter of β in cox-regression.

ANN

Artificial Neural Network is a multi-layer perceptron(MLP). It consists of input, hidden, and output layers. Each

neuron has an activation function that defines the output. Learning, updates the parameters, consists of propagation and backpropagation. It is evaluated by the loss function. Our components of ANN are presented in **Fig.3**. This is also just calculating parameters β in Cox Regression.

Methods	Name
Tuning Hyperparameter	Grid Search, Optuna
Regularizer	Elastic Net
Activation	Selu
Intiailizer	Lecun_Normal
Loss_Function	Custom Function
Optimizer	Nadam

Fig.3 ANN setting of this report.

Results

This report has four machine learning models for two purposes, sentiment analysis and survival analysis. BERT pre-trained model is for the sentiment analysis and RSF, Cox Regression, and ANN are for the survival analysis. This report will compare the results of the models.

1. **Sentiment Analysis of Drug Review Dataset (Drugs.com) Data Set**

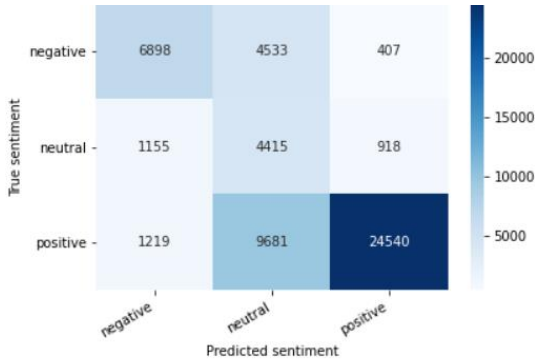

Fig.4 Confusion Matrix of the sentiment analysis.

Our model has 0.60, f1-score(macro avg.). It is not a satisfying score. We can check the details of the score in the confusion matrix(**Fig.4**) above and the score table(**Fig.5**). Its overall precision is 0.64 and its overall recall is 0.65. The precision and recall of each sentiment status explain the limitation of the model. The significant differences between the scores are from the neutral precision and other classes precision. Precision means how accurately the model's positive prediction is made. Accordingly, the model frequently predicts actual positive and negative reviews as neutral reviews. The confusion matrix elaborates this problem, 9,681 actual positive reviews, 4,533 actual negative reviews are predicted into neutral reviews. Approximately, 27.3% of positive reviews are misclassified as neutral reviews and approximately, 38.2% of negative reviews are also misclassified as neutral reviews. The proportion of predicted neutral is 24.3%, 23.7%, and 52.0%, actual negative, actual neutral, and actual positive respectively. The number of actual positive reviews in the test dataset is three times bigger than actual negative reviews in the test dataset. Therefore, the most severe misclassification of the model is that it cannot distinguish between neutral and negative. The distinguishing between positive and neutral is also a problem, but the former problem is more significant according to the ratio of the confusion matrix.

Sentiment	Score			
	Precision	Recall	F1-score	Support
Negative	0.74	0.58	0.65	11838
Neutral	0.24	0.68	0.35	6488
Positive	0.95	0.69	0.80	35440
Accuracy			0.67	53766
Macro avg	0.64	0.65	0.60	53766
Weighted avg	0.82	0.67	0.71	53766

Fig.4 Score table, represents precision, recall, f1-score, and support and it is divided into negative, neutral, and positive.

2. Survival Analysis of Breast Cancer Wisconsin (Prognostic) Data Set

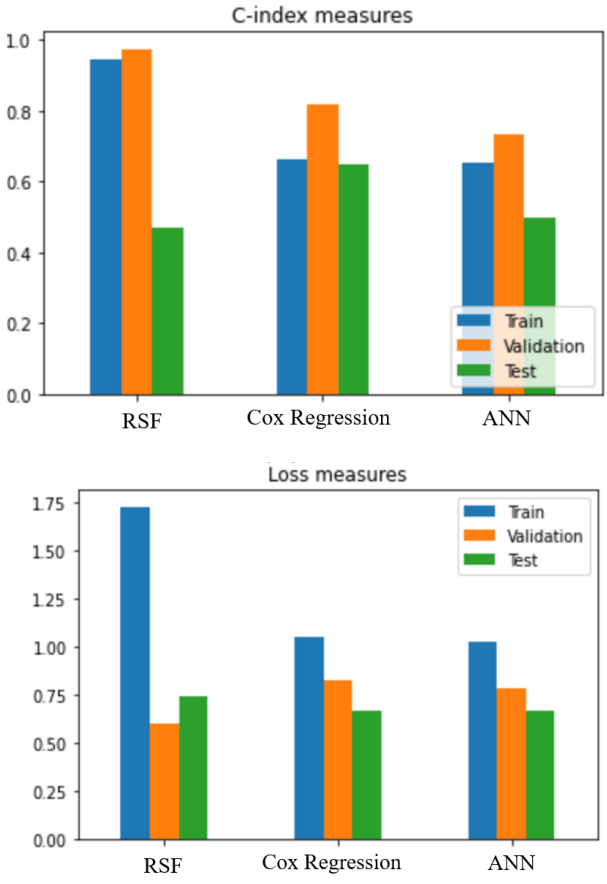

Fig.5 Bar graph of c-index and loss measures of each model for survival analysis.

In survival analysis, there are 3 models and the metric for evaluation of the model is the concordance index. The highest c-index on the test dataset is cox regression penalized by elasticnet. It was an unexpected result because we expected the relationship with the cox regression parameter and features is somehow non-linear. Therefore, we expected RSF and ANN have higher c-index than just linear regression. It was true in the training dataset and validation dataset. We can easily deduce RSF and ANN are overfitting on the datasets, especially in RSF. Loss measure, we use given custom function to calculate the loss[14], can confirm this. The loss function only considers the difference between the true value and the predicted value. However, the loss measure of RSF is higher than

other models on the training dataset. During the training, the RSF model learns the order of the data but it did not learn the actual values, this result should be studied in the future works. Therefore, it is not overfitted. There is also no significant difference in loss measure in the other datasets. We can consider it is caused by the sparsity of the time-to-event dataset, it is hard to learn the event occurs because the breast cancer recurrence is a rare case, 47 cases out of 198 cases, and the lack of data if the expectation is correct. So, we need to check this limitation in future works.

However, we don't have any evidence to conclude our result is wrong. Our conclusion is the best model of survival analysis is the Cox regression penalized by elasticnet based on our metric score(**Fig.6**).

C-index results of the trained model			
	train	**validation**	**test**
RSF	0.944	0.974	0.470
Cox Regression	0.664	0.820	0.647
ANN	0.653	0.734	0.500

Fig.6 Table of c-index of each model for survival analysis.

Conclusion

- Sentiment Analysis

We used BERT pre-trained model and add the classification layer on the model for sentiment analysis. The results tell us it is hard to distinguish neutral sentiment and positive or negative sentiment. We need to consider the new structure to predict the neutral sentiment or more data of neutral sentiment.

Limitation

It is problem for the model to distinguish neutral and negative reviews. This is from the structural problem of the model and the data itself also. First of all, we used the hyperparameter of the recommendation from BERT paper[10] without any logical inference, we should have thought about our data and the characteristics of the model. Secondly, time and computation power were limited and we had to compromise the time and resources, we truncated the data randomly and we cut the token size shorter than our analysis result suggests, and also we changed the batch size for the computation speed.

Our data is imbalanced because the number of neutral reviews is significantly small. We tried to cut the value above the 8 but our evaluation standards are above the 7. Moreover, We need to think about the limitation of ratings in reviews. People can write a review with neutral words and give the rating 10 points or vice versa. In this case, even if we made a perfect model, we cannot distinguish people's behavior. We can imagine the default setting of the rating was 10 points to get the favor of other customers and people to write the review and just simply click the send button. Companies are using this trick a lot in the online shop.

We didn't use the statistical methods to evaluate the model and it will give us more solid results with a confidence interval of the f1-score. Therefore, we don't know how much we can believe the result.

Strength

The model can relatively distinguish negative and positive well. It can be explained by the abundance of the data and the word itself can be differentiated more easily. BERT pre-trained model is a powerful model for fine-tuning. BERT helps us to optimize the model and build our model efficiently.

Future Works

If we build other models for this purpose in the future and we have proper time and computation power, the model will be constructed as follows:

1. Formulate the null hypothesis or define the target metrics with a specific confidence interval.
2. Exploratory Data Analysis: check the data by clustering, distribution, simple regression, and correlation with vectorized data.
3. The data is imbalanced. Therefore, we make data be balanced (sample more data, resample the data, or SMOTE) or I would like to use GPT-3 to produce the neutral sentiment text.
4. We need to search for different pre-trained models concerning the purpose.
5. With small-batch, we test our hyperparameters properly and choose it based on the results.
6. We train the model with multi GPUs.
7. Analysis of the data with f1 score and confusion matrix and compare this with the previous one. We need to build a proper statistical testing model to compare results or get the estimation of the actual parameter with a proper confidence interval.

In this way, we can add more neutral reviews in the datasets and we can pick hyperparameter properly, not blindly pick or compromised value. We can also compare the results statistically.

- Survival Analysis

We used three models to implement the survival analysis on the breast cancer datasets. The models are Random Survival Forest, Cox Regression with elasticnet, and ANN. The result indicates the Cox regression is the best model. Therefore, we realize the coefficients in the Cox regression follow linear regression.

Limitation

The most significant limitation is we don't know the differences of metric is meaningful in the scope of statistics. We did not use the p-value or confidence interval to confirm the differences. Consequently, we cannot say the result is valid.

The small dataset size is also a problem to train the data. We can interpret this with Bayesian Inference.

$$P(H|DX) = P(H|X) * \frac{P(D|HX)}{P(D|X)}$$

D: given data, X: Every background information, H: Hypothesis

We updated the posterior by changing the likelihood of the hypothesis. The likelihood will be narrower and narrower when we get more and more data. In our experiment, the data is small. Thus, the distribution of the likelihood is spread-out space and we cannot infer the correct hypothesis.

ANN setting also has limitations because of the bottleneck in the output layer. The bottleneck is used for the representation of the previous data. However, our layer design is 32-16-16-16-1 and we used Lecun Normal Initialization and Selu as an activation function. It means our coefficient is pretty much fixed in the range of normalization. It helps fast optimization but I think it can make trouble in a bottleneck because the coefficient can't move a wide range of value. 16 values have to be represented 1 value with an almost fixed coefficient.

Strength

The learning is relatively easy. We build three models. Thus, we can build other weak models and stack them together. The combination of a weak model can be a powerful model[15]. The next step to improve this model is stacking the models.

Future Works

The methods are almost the same as the previous section:

1. Formulate the null hypothesis or define the target metrics with a specific confidence interval.
2. Exploratory Data Analysis: check the data by clustering, distribution, simple regression, and correlation with vectorized data.
3. The data have a null value. We will fill the value with SMOTE or drop the data.
4. We train several weak models including RSF, Cox Regression, and small ANN, and train XGboost for comparison.
5. We stack the different models and pick final prediction models, linear and non-linear models. We have three final models, two stacking with linear and non-linear final model and XGboost.
6. We need to build proper statistical testing, ANOVA test, model to compare results or get the estimation of the actual parameter with a proper confidence interval.

References

1. Gräßer F, Kallumadi S, Malberg H, Zaunseder S. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In: *Proceedings of the 2018 International Conference on Digital Health*. ACM; 2018:121-125. doi:10.1145/3194658.3194677

2. UCI Machine Learning Repository: Breast Cancer Wisconsin (Prognostic) Data Set. Accessed August 30, 2020. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic)

3. Sentiment Analysis. MonkeyLearn. Published June 20, 2018. Accessed August 30, 2020. https://monkeylearn.com/sentiment-analysis

4. Sentiment analysis. In: *Wikipedia*. ; 2020. Accessed August 30, 2020. https://en.wikipedia.org/w/index.php?title=Sentiment_analysis&oldid=975607751

5. [1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Accessed August 16, 2020. https://arxiv.org/abs/1810.04805

6. Derczynski L. Complementarity, F-score, and NLP Evaluation. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA); 2016:261–266. Accessed August 30, 2020. https://www.aclweb.org/anthology/L16-1040

7. Survival analysis. In: *Wikipedia*. ; 2020. Accessed August 30, 2020. https://en.wikipedia.org/w/index.php?title=Survival_analysis&oldid=972914464

8. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. *Br J Cancer*. 2003;89(2):232-238. doi:10.1038/sj.bjc.6601118

9. Steck H, Krishnapuram B, Dehing-oberije C, Lambin P, Raykar VC. On Ranking in Survival Analysis: Bounds on the Concordance Index. In: Platt JC, Koller D, Singer Y, Roweis ST, eds. *Advances in Neural*

Information Processing Systems 20. Curran Associates, Inc.; 2008:1209–1216. Accessed September 5, 2020. http://papers.nips.cc/paper/3375-on-ranking-in-survival-analysis-bounds-on-the-concordance-index.pdf

10. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv181004805 Cs*. Published online May 24, 2019. Accessed August 16, 2020. http://arxiv.org/abs/1810.04805

11. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. *ArXiv170603762 Cs*. Published online December 5, 2017. Accessed September 6, 2020. http://arxiv.org/abs/1706.03762

12. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-860. doi:10.1214/08-AOAS169

13. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

14. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. 2018;18(1):24. doi:10.1186/s12874-018-0482-1

15. Džeroski S, Ženko B. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Mach Learn*. 2004;54(3):255-273. doi:10.1023/B:MACH.0000015881.36452.6e