Explainable Empirical Risk Minimization

Yifan Zhu
yifan.zhu@aalto.fi

Tutor: Alexander Jung

Abstract

The demand for explainability in machine learning (ML) predictions has been increasing with the wider application of ML techniques and the significance of the decisions from predictions. An ideal ML model should have both high predictive accuracy and explainability, whereas there is a trade-off between these two targets. As a potential solution, Explainable Empirical Risk Minimization (EERM) has been proposed to learn a predictive model by balancing its empirical risk and user-specific explainability. This paper presents the details of EERM, evaluates its performance by implementing EERM in linear hypothesis space, and compares experiment results with linear regression, ridge regression and LASSO regression. The result preliminarily shows the utility of EERM in linear hypothesis space.

KEYWORDS: empirical risk minimization, explainable machine learning, regularized loss minimization

1 Introduction

With the development of Machine Learning (ML) in recent years, ML techniques have been applied to much broader areas, including drug design [1], healthcare [2], and finance [3]. Many decisions of these applications, which might be affected by ML predictions, are crucial in that wrong decisions can cause severe outcomes [3]. Furthermore, not knowing why and how the prediction is generated makes it difficult to adjust the ML model. Thus, the demand for explainability of ML predictions is rapidly increasing, which even limits the further broader application of ML techniques.

Both predictive performance and explainability are significant indicators of ML model performance. However, for most models, there is a trade-off between these two factors [4]. With the improvement of predictive performance, the model can become more complex, then it is more difficult to interpret the model. As a result, Explainable Empirical Risk Minimization (EERM) [5] has been proposed to learn an ML model such that both predictive performance and explainability are considered. In EERM, a user's knowledge for the data is quantified by an information-theoretic measure as a "user summary" [6], and then the explainability of the model can be measured by the distance between the prediction and user summary. As an instance of Regularized Loss Minimization (RLM), EERM uses the explainability quantification as the regularization function to learn the optimal ML model, which is to have low empirical risk and to be clear for the user to understand.

To the best of our knowledge, EERM has only been formulated, and there is little further evaluation and analysis of it. This paper reviews the main features of EERM, evaluates its performance with experiments, and analyses its utility together with other RLM algorithms.

This paper is organized as follows. Section 2 introduces the basic knowledge and prerequisites of EERM. Section 3 presents the fundamentals and details of EERM. Section 4 contains the experiment details, analyses the experiment results and discusses the utility of EERM. Finally, Section 5 provides conclusions of EERM.

2 Prerequisites

In this section, we present some prerequisites for understanding EERM in detail. Section 2.1 introduces the basic concepts of ML and learning principles related to EERM. Section 2.2 elaborates on the definition and methods of explainable machine learning. Finally, Section 2.3 presents the basic knowledge of information-theoretic measures used in EERM.

2.1 Learning Principles in ML

Machine Learning is a technique that detects important patterns from existing knowledge to obtain new information [7]. For example, in the application of weather prediction, taking the weather data of the past year as input, ML algorithms can extract features, such as the correlation between weather and seasons, to predict the future weather.

Key Principle

The key principle of most ML algorithms is to train a model which builds a prediction map from the input data to the output prediction. Let $\mathcal{X} = \{\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_n\}$ denote the input space, $\mathrm{x}_i = (\mathrm{x}_i^1, \mathrm{x}_i^2, ..., \mathrm{x}_i^m)$ denote each instance in $\mathcal{X}$, which is also a feature vector , $\mathcal{Y}$ denote the output space, which is the label set for $\mathcal{X}$. Then, to learn a model is to learn a prediction map h

$$\begin{gathered} y = h(\mathrm{x}) \\ h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y} \end{gathered} \tag{1}$$

where x is a feature vector in $\mathcal{X}$, and y is the predicted label for x. The prediction map h is referred to as a hypothesis in the hypothesis space $\mathcal{H}$ as well as a predictor.

Generalization

One of the most significant goals of ML algorithms is to learn a predictor with a strong performance in generalization. Generalization is an ability that the predictor can make predictions on unseen instances with satisfying accuracy based on the training process. This ability can be measured by a loss function $\mathcal{L}(y, h(\mathrm{x}))$, which calculates a score of error based on the distance between the true label y and the prediction $h(\mathrm{x})$. High value given by $\mathcal{L}(y, h(\mathrm{x}))$ indicates that there are many differences between the prediction result and the true label, and then the predictor has poor prediction performance. Hence, the goal of ML can be quantified as finding the optimal hypothesis h^* such that

$$h^* = \arg\min_{h \in \mathcal{H}} R(h) \tag{2}$$

where $R(h)$ denotes the generalization error of hypothesis h.

Empirical Risk Minimization

Since the true label for the unseen data is unknown, we cannot directly measure the generalization error. Then, Empirical Risk Minimization

(ERM) [8] was proposed as a learning principle to approximate generalization performance by computing the average error over the training set Γ

$$\widetilde{R}(h) = \frac{1}{|\Gamma|} \sum_{(\mathrm{x},y)\in\Gamma} \mathcal{L}(h(\mathrm{x}), y) \tag{3}$$

$\widetilde{R}(h)$ is referred to as the empirical risk of hypothesis h. With ERM, we may select the hypothesis h^* with the minimum empirical risk as the optimal predictor

$$h^* = \arg\min_{h\in\mathcal{H}} \widetilde{R}(h) \tag{4}$$

However, ERM may arise the problem of overfitting. Overfitting is a phenomenon such that the predictor with small empirical risk has high risk in the unseen data, which indicates that the predictor cannot generalize on the unseen data. This phenomenon is mainly caused by the gap between the size of training set and the complexity of hypotheses. Since the training set size is often limited, constraining complexity of hypotheses may avoid overfitting.

Regularized Loss Minimization

The learning principle Regularized Loss Minimization (RLM) was proposed to balance the empirical risk and the predictor complexity by minimizing the sum of $\widetilde{R}(h)$ and a regularization function $\mathcal{R}(h)$

$$h^* = \arg\min_{h\in\mathcal{H}} \widetilde{R}(h) + \lambda\mathcal{R}(h) \tag{5}$$

where λ is a hyperparameter. $\mathcal{R}(h)$ measures the complexity of hypothesis h. For example, Lasso regression adopts L1 norm $||\mathrm{w}||_1 = \sum_{i=1}^{n} |w_i|$ as the regularization function to constrain the number of model parameters as well as the complexity in RLM; thus we can obtain a hypothesis with both satisfying empirical risk and complexity.

2.2 Explainable Machine Learning

The growing demands for the explainability have brought increasing popularity and importance for Explainable Machine Learning. Explainable ML is defined by Murdoch [9] as extracting the relations from the learned ML model, which are either between data features and labels or learned by the predictor.

The main categories for Explainable ML methods are intrinsic and post hoc [10]. Intrinsic methods refer to ML models that are intrinsically explainable due to their simple structure or relatively low complexity, while methods belonging to post hoc analyse the information extracted from the learned model to obtain explanations. The key issue of Intrinsic methods is the trade-off between explainability and prediction accuracy, since model with high complexity may lead to low explainability and high accuracy. EERM is then proposed by Jung [5] as an intrinsic method to learn a predictor with balanced accuracy and explainability.

2.3 Differential entropy

Entropy is a basic concept in Information theory, which quantifies the uncertainty of a random variable [11]. The uncertainty can be interpreted as the amount of information contained in the random variable. Specifically, the occurrence of a rare event can convey a large amount of information as well as high uncertainty.

Differential entropy is the entropy for continuous random variables. Conditional entropy $H(\mathrm{Y}|\mathrm{X})$ is an extension of entropy in that it quantifies the uncertainty of random variable Y give the information of X. Let Y be a continuous random variable, $\mathcal{S}_Y$ be the corresponding domain set, then $H(\mathrm{Y}|\mathrm{X})$ can be defined by

$$\begin{aligned} H(\mathrm{Y}|\mathrm{X}) &= -\int_{\mathcal{S}_X,\mathcal{S}_Y} f(x,y)\log f(y|x)dxdy \\ &= -\mathrm{E}[\log f(y|x)] \end{aligned} \tag{6}$$

where $f(x,y)$ is the joint probability density function of X and Y, $f(y|x)$ is the conditional probability distribution of X and Y.

3 Regularized Loss Minimization algorithms

In the following, we present the details of EERM algorithm, which is an instance of RLM, and classical RLM algorithms, namely ridge regression and LASSO regression. Then, in Section 4, we conduct experiments on these algorithms to compare their performances and then discuss the utility of EERM.

3.1 Explainable Empirical Risk Minimization

Explainable Empirical Risk Minimization is proposed by Jung [5] to search for the optimal hypothesis with low empirical risk and high explainability in the hypothesis space, which is intrinsically explainable. The essence of EERM is the combination of RLM and user background. EERM models users' background, the understanding of data to be specific, as a user summary, which quantifies users' expected value for the prediction results. Then, the explainability of the predictor is measured by the conditional differential entropy between prediction results and the user summary. To learn the optimal model in the form of RLM, the empirical risk is used to quantify the model generalization performance, and the explainability quantification is used as the regularization function. The optimal model obtained from training is an optimized result of the trade-off between explainability and predictive performance, which has satisfying generalization performance and is explainable to the user.

User Summary

In Explainable ML, it is significant to notice that explainability is measured based on the context [12], which is the user's background for, for example, the application and ML techniques. Users with different backgrounds can have different level of understanding and different concerns for the predicted results. For example, a user who is familiar with linear regression can interpret the predicted results directly by the weight of each feature, which represents the significance of the feature, while a user with little knowledge of the algorithm may need a more detailed explanation for understanding the result.

To process the varying user background, EERM proposed user summary to model user's understanding for the data and features. User summary is defined as user's expected value for the predicted result of given data, which is largely based on user's intuition for the correlation between data features and labels [5]. For example, in weather prediction, the user can provide own prediction for future weather based on the observations and intuitions, such as the correlation between weather and seasons. Thus, by providing the data and user summary to the training process, users with different backgrounds can obtain a personalized explainable ML model.

Explainability Quantification

With user summary $\hat{u}$ and predicted result $\hat{y}$ for data x, the explainability of the predictor can be modeled as conditional differential entropy $H(\hat{y}|\hat{u})$, which quantifies the uncertainty of observing the predicted result while given user summary. Since user summary contains user's expectation for the prediction, high uncertainty indicates that the predicted result is largely different from user's expectation, which suggest that the predictor is out of user's understanding; thus the predictor has poor explainability.

Furthermore, if we narrow down the range of hypothesis space to linear predictors and assume that the data and its user summary follow multivariate normal distribution with zero mean, the explainability quantification can be further derived as $E[(\hat{y} - \alpha\hat{u})^2]$ [5] with the entropy of normal distribution, where α is a constant. This quantification directly measures the distance between predicted results and user summary, which is easier to understand as well as compute.

Regularization

The objective function in EERM can be obtained by combining ERM and the explainability quantification [5]

$$\begin{aligned} h^* &= \arg\min_{h \in \mathcal{H}} \widetilde{R}(h) \\ s.t.\ & H(\hat{y}|\hat{u}) \leq \eta \end{aligned} \tag{7}$$

where the entropy is constrained to be smaller than η.

In practice, if we keep the assumptions above and quantify explainability as $E[(\hat{y} - \alpha\hat{u})^2]$, then we can write (7) as [5]

$$h^* = \arg\min_{h \in \mathcal{H}} \widetilde{R}(h) + \lambda E[(\mathrm{w}^T \mathrm{x} - \alpha\hat{u})^2] \tag{8}$$

where w is the weight parameters in hypothesis h.

It is noticeable that (8) is in the form of RLM with explainability quantification as the regularization term. Clearly, during the training process, EERM algorithm will seek to find the optimal hypothesis with balanced generalization performance and explainability based on the user's specific background.

3.2 Classical Regularization Algorithms

Both ridge regression and LASSO regression are classical regularization algorithms. They share the similarity of being performed on linear hypotheses, and the major difference between them is the choice of regularization functions, which leads to the difference in their preferences for training the weight parameters.

Ridge Regression

Ridge regression is a regularization algorithm performed on linear regression with l_2 norm as the regularization function

$$h^* = \arg\min_{h \in \mathcal{H}} \lambda ||\mathbf{w}||_2^2 + \sum_{i=1}^{n} (h(\mathbf{x}_i) - y_i)^2 \tag{9}$$

where λ is the regularization parameter that controls the penalty on the parameters; $||\mathrm{w}||_2$ is the l_2 norm, which equals to $\sqrt{\sum_{i=1}^{m} w_i^2}$.

Utilizing l_2 norm as the regularization function, ridge regression tends to regularize the weight parameters to approach zero with a large value of λ; thus, it is capable of decreasing model complexity.

LASSO Regression

LASSO stands for least absolute shrinkage and selection operator. Unlike ridge regression, LASSO regression uses l_1 norm as the regularization function

$$h^* = \arg\min_{h \in \mathcal{H}} \lambda||\mathbf{w}||_1 + \sum_{i=1}^{n}(h(\mathbf{x}_i) - y_i)^2 \tag{10}$$

where $||\mathbf{w}||_1$ equals to $\sum_{i=1}^{m} |w_i|$.

From (10) we can observe that if the regularization parameter λ is chosen be a large value, then LASSO regression will penalize weight parameters of insignificant features to be equal to zero. In other words, LASSO regression is capable of performing feature selection, which leads to the similar effects of ridge regression.

4 Experiments and Results

To evaluate the performance of EERM and compare it with ridge regression as well as LASSO regression, we implemented the EERM algorithm specifically for a linear regression task; conducted experiments of these three algorithms as well as the baseline linear regression algorithm on the same dataset; presented their scores as well as learned weight parameters respectively as the experiment result. Codes for the experiments can be found in GitHub.

4.1 Dataset

The dataset we used is the California Housing dataset. It contains the housing information of 20640 census location blocks from the 1990 California census. Ten attributes of census blocks are included in the dataset, which are *longitude*, *latitude*, *the median age of housing*, *total amount of rooms*, *total amount of bedrooms*, *population*, *total amount of households*, *median income*, *distance to the ocean* and *the median house value*. The first nine attributes are used as data features, and the median house value is used as the target.

4.2 Experiment details

To ensure that the dataset is usable, we cleaned the data in advance by processing missing value and categorical value. Furthermore, to avoid data overflow in the implementation of EERM, the data is scaled by standardization. The dataset is then randomly split into training set and test set for training and evaluating the performance of the algorithms.

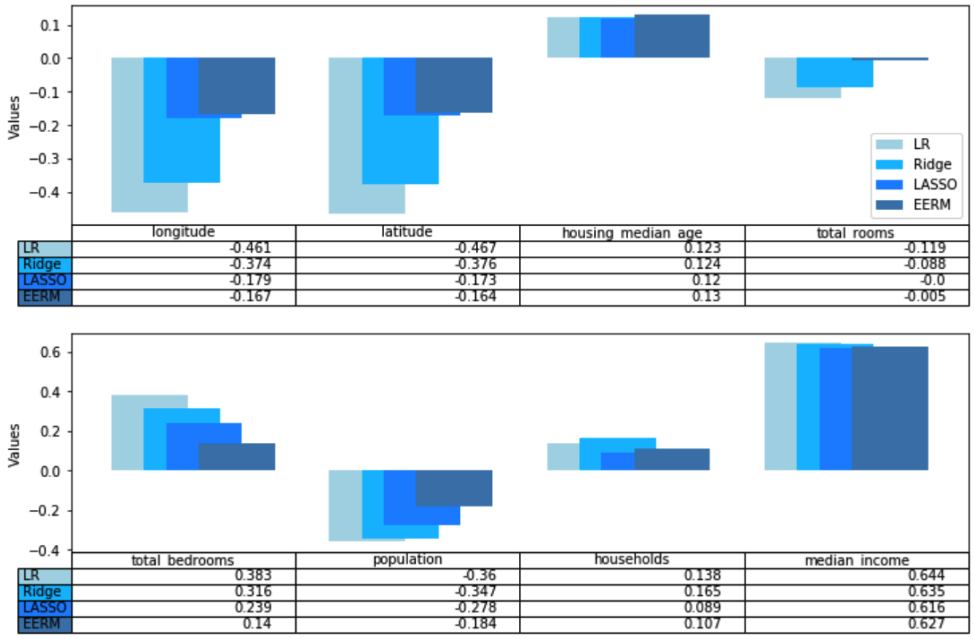
The models of Linear regression, ridge regression and LASSO regression are directly fetched from scikit-learn library, and the model of EERM is implemented according to (8). In particular, to obtain the user summary for EERM, we first plot the correlation between all ten attributes to simulate the intuition for the data. Then, we remove features that are not strongly correlated to the target, which are *longitude*, *total amount of bedrooms*, *population* and *total amount of households*, to form a new dataset and fit it with another linear regression. The prediction of the new dataset is used as the user summary. Coefficient of determination regression (R^2) score is used to measure the model generalization performance. Score approaching to 1 indicates that the model can fit unseen data well, and negative score may suggest that the model is worse than a random model.

4.3 Experiment results

R^2 score for linear regression, ridge regression, LASSO regression and EERM are 0.461, 0.450, 0.408 and 0.413 respectively. Figure 1 presents the learned weight parameters of these models.

From R^2 scores we can observe that models of linear regression and ridge regression have a similar performance, which is better than the other two; the EERM model performs slightly better than the LASSO regression model. These observations may result from the fact that the dataset includes few features compared to its large amount of data samples, which indicates that regularization is hardly needed here to constrain model complexity, especially the feature selection from LASSO regression.

From Figure 1 we can clearly observe the effects of regularization algorithm compared to the baseline model. The weights from regularization algorithms are smaller than the one from the baseline model, which suggests that models from regularization have lower complexity. Furthermore, the features which are deemed to be insignificant in the user summary have the smallest learned weights from EERM, which indicates that



	longitude	latitude	housing median age	total rooms
LR	-0.461	-0.467	0.123	-0.119
Ridge	-0.374	-0.376	0.124	-0.088
LASSO	-0.179	-0.173	0.12	-0.0
EERM	-0.167	-0.164	0.13	-0.005

	total bedrooms	population	households	median income
LR	0.383	-0.36	0.138	0.644
Ridge	0.316	-0.347	0.165	0.635
LASSO	0.239	-0.278	0.089	0.616
EERM	0.14	-0.184	0.107	0.627

Figure 1. 8 learned weight coefficients of linear regression, ridge regression, LASSO regression and EERM.

the model from EERM is more consistent with the user's intuition for the data, and thus it is more explainable to the user.

The experiment shows the essence of EERM, which is the trade-off between empirical risk and explainability. The model of EERM has acceptable empirical risk and better user-specific explainability compared to other models. Nevertheless, more circumstances need to be considered for a comprehensive evaluation of EERM, such as giving a user summary that contains wrong intuition for the data and applying EERM to non-linear hypothesis space.

5 Conclusion

This paper introduces the essence and details of Explainable Empirical Risk Minimization. For evaluating the performance of EERM preliminarily, we implement EERM in linear hypothesis space and then conduct experiments together with algorithms of linear regression, ridge regression and LASSO regression. The result shows that in linear hypothesis space, with appropriate user summary, EERM is capable of learning a model with satisfying empirical risk and better user-specific explainability. Further work needs to be done to evaluate EERM comprehensively, including implementations in non-linear hypothesis space and experiments with user summary containing wrong data intuition.

References

[1] Katsunori Sasahara, Masakazu Shibata, Hiroyuki Sasabe, Tomoki Suzuki, Kenji Takeuchi, Ken Umehara, and Eiji Kashiyama. Feature importance of machine learning prediction models shows structurally active part and important physicochemical features in drug design. *Drug metabolism and pharmacokinetics*, 39:100401, 2021.

[2] K. Shailaja, B. Seetharamulu, and M. A. Jabbar. Machine learning in healthcare: A review. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 910–914, 2018.

[3] Wei-Yang Lin, Ya-Han Hu, and Chih-Fong Tsai. Machine learning in financial crisis prediction: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):421–436, 2012.

[4] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013.

[5] Alexander Jung. Explainable empirical risk minimization. *CoRR*, abs/2009.01492, 2020.

[6] Alexander Jung. A gentle introduction to supervised machine learning. *CoRR*, abs/1805.05052, 2018.

[7] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014.

[8] V. Vapnik. Principles of risk minimization for learning theory. In J. Moody, S. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991.

[9] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, Oct 2019.

[10] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[11] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[12] Katharina Beckh, Sebastian Müller, Matthias Jakobs, Vanessa Toborek, Hanxiao Tan, Raphael Fischer, Pascal Welke, Sebastian Houben, and Laura von Rüden. Explainable machine learning with prior knowledge: An overview. *CoRR*, abs/2105.10172, 2021.