Learning Groupwise Explanations for Black-Box Models

Jingyue Gao[1] , **Xiting Wang**[2] , **Yasha Wang**[1*] , **Yulan Yan**[3] , **Xing Xie**[2]
[1]Peking University [2]Microsoft Research Asia [3]Microsoft
{gaojingyue1997, wangyasha}@pku.edu.cn, {xitwan, yulanyan, xing.xie}@microsoft.com

Abstract

We study two user demands that are important during the exploitation of explanations in practice: 1) understanding the overall model behavior faithfully with limited cognitive load and 2) predicting the model behavior accurately on unseen instances. We illustrate that the two user demands correspond to two major sub-processes in the human cognitive process and propose a unified framework to fulfill them simultaneously. Given a local explanation method, our framework jointly 1) learns a limited number of groupwise explanations that interpret the model behavior on most instances with high fidelity and 2) specifies the region where each explanation applies. Experiments on six datasets demonstrate the effectiveness of our method.

1 Introduction

The decision process of state-of-the-art machine learning models such as deep neural networks is often obfuscated by their intricate architectures. Despite the impressive prediction accuracy, lack of interpretability inevitably hinders the adoption of these models, especially when users need to understand the model behavior and ensure the models are correct and ethical [Ribeiro *et al.*, 2016; Chu *et al.*, 2018].

In recent years, providing explanations for black-box models has attracted increasing attention in the research community. Substantial efforts have been devoted to explaining the model prediction of an individual instance with high fidelity [Ribeiro *et al.*, 2016; Elenberg *et al.*, 2017; Lundberg and Lee, 2017; Dhurandhar *et al.*, 2018; Guidotti *et al.*, 2018; Plumb *et al.*, 2018]. These methods have achieved great success in providing explanations that are both succinct and faithful. However, for machine learning practitioners, there is still a gap between understanding each instance well and gaining a clear and comprehensive understanding of the overall model behavior on most instances. We observe that during the subsequent exploitation of explanations in practice, fulfilling the following user demands are essential.

D1: Obtaining a faithful understanding of the overall model behavior with limited cognitive load. While instance-level explanations are insightful, it is infeasible to examine explanations for all instances to gain a overall understanding of the model. Existing methods solve this issue either by generating one global explanation over the entire input space [Ghorbani *et al.*, 2019; Kim *et al.*, 2018] or selecting representative local explanations [Ribeiro *et al.*, 2016; Ramamurthy *et al.*, 2020]. While global explanations often fail to provide a succinct or faithful explanation when the target model is complex [Ribeiro *et al.*, 2018], selecting representative local explanations depends on heuristic assumptions about which explanations are the most representative. The latter methods lack a mechanism to directly optimize the fidelity of the selected explanations on all instances. As a result, there is no guarantee that the representative explanations they select are highly faithful over the entire input space.

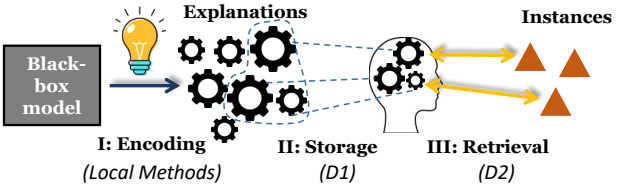

Figure 1: Human cognitive process for model interpretation.

D2: Making accurate predictions about the model behavior on unseen instances. Researchers have found that to achieve high *human precision* [Ribeiro *et al.*, 2018] or *generalized fidelity* [Ramamurthy *et al.*, 2020], it is essential that we clearly define the region where an explanation applies. Otherwise, users may easily make mistaken predictions about the model behavior on unseen (test) instances by employing an incorrect explanation. Existing methods typically assume that explanations can be applied to instances that are similar according to a certain feature space [Plumb *et al.*, 2018; Ramamurthy *et al.*, 2020]. This biased assumption often yield sub-optimal generalized fidelity [Ribeiro *et al.*, 2018]. To alleviate this issue, [Ribeiro *et al.*, 2018] define the region an explanation applies by using association rules. However, this method fails to provide insights on the relative importance of each feature, and there is no guarantee that the association rules can be applied to a large percent of instances.

Satisfying the two user demands is importance as they enable us to take a more complete view of the human cognitive process when interpreting models (Fig. 1). According to cognitive psychology, the major sub-processes of human's

*Yasha Wang is the corresponding author

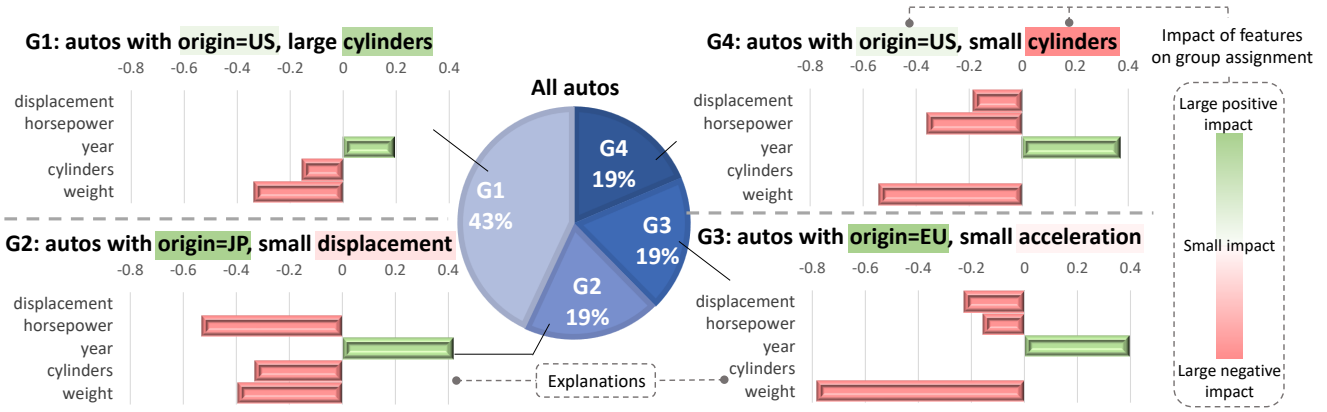

Figure 2: Groupwise explanations generated by GIME on the AutoMPG dataset. The task is to predict miles per gallon based on attributes of automobiles. GIME divides autos into four groups (G1-G4) with noticeable difference in model behaviors and clearly defines each group. Users can quickly identify that the model behaves differently on autos with different origins and that the US autos can be further divided based on the cylinder (G1 and G4). The fuel consumption of autos in G1 is the least sensitive to the input attributes. This is consistent with the fact that the prediction score for these autos varies the least among all groups (standard deviation is the smallest), although G1 contains the largest number of autos. The cylinder impacts the fuel consumption of autos in G1 and G2, but hardly influences that in G3 and G4. This is reasonable considering that only for G3 and G4, the Pearson correlation between the cylinder and the prediction score is the smallest among all attributes.

information-processing model are *encoding*, *storage*, and *retrieval* [Lang, 2000]. While local explanation methods help interpret and encode a single piece of information about one instance faithfully (*encoding*), users still need to distinguish important pieces of encoded information so that they can be stored properly with limited cognitive resource (D1, *storage*). Then, users need to accurately reactivate the piece of relevant information for decision making (D2, *retrieval*).

In this paper, we propose a principled way to simultaneously fulfill D1 and D2. Given a post-hoc local explanation method, we study how *groupwise* explanations that faithfully reveal model behavior on multiple instances can be learned. In particular, our contributions are three folds.

First, we propose a unified **G**roupw**I**se **M**odel-agnostic **E**xplanation (**GIME**[1]) framework, which jointly 1) learns a limited number of groupwise explanations that interpret the overall model behavior with high fidelity (D1) and 2) specifies the region each explanation applies (D2). Different from existing methods, which handle D1 or D2 separately with different heuristic assumptions, our framework treats the extraction of groupwise explanations as a learning task and fulfills the two inherently interconnected user demands simultaneously by directly optimizing the overall fidelity. As shown in Fig. 2, our method can automatically divide instances into groups with noticeable difference in model behaviors, clearly defines each group, and provides interesting insights.

Second, we show how to effectively formulate the region an explanation applies. In particular, we discuss the desirable properties of the region formulations and why straightforward solutions may be problematic. Based on the discussion, we introduce the von Mises–Fisher (vMF) distribution [Banerjee *et al.*, 2005], which is an example distribution that satisfies the desirable properties and leads to good empirical results.

Finally, we conduct both quantitative experiments and experiments with real users to demonstrate the effectiveness of our method. Codes are provided in the supplementary material to facilitate reproduction of the experimental results.

[1]Source code: https://github.com/jygao97/GIME

2 Problem Formulation

We formulate our problem as follows.

Input. The input of our framework includes a dataset $\mathbf{X}$, a target model f to be explained, and a cognitive budget K.

- The **dataset** $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ consists of N instances, each denoted by a feature vector $\mathbf{x}_i \in \mathbf{R}^d$.
- The **target model** is treated as a black-box function $f : \mathbf{R}^d \to \mathbf{R}$. For a classification model, $f(\mathbf{x}_i)$ denotes the predicted probability that $\mathbf{x}_i$ belongs to certain class.
- The **cognitive budget** K is the maximum number of explanations that the user can examine.

Output. We find K **groupwise explanations** and the regions where they apply. The explanations can be formulated by following a given local explanation method. We use LIME [Ribeiro *et al.*, 2016] as a guiding example and define an explanation as an interpretable surrogate model $g_k(\mathbf{x}) = \boldsymbol{\theta}_k^T \mathbf{x}$, where $\boldsymbol{\theta}_k \in \mathbf{R}^d$ reveals the feature importance.

3 Background

Our groupwise explanation framework is designed by extending existing local explanation methods, which interpret the model prediction of a single instance with high-fidelity. Given an instance $\mathbf{x}$ and a black-box model f, identifying important features that contribute to $f(\mathbf{x})$ is usually achieved by finding casual structures in the model's response to input features. A common practice is to perturb the input and observe how $f(\mathbf{x})$ changes accordingly. Following this spirit, LIME [Ribeiro *et al.*, 2016] constructs the neighborhood $\mathcal{N}(\mathbf{x})$ of instance $\mathbf{x}$ by perturbing $\mathbf{x}$ and derives the explanation $\boldsymbol{\theta}$ by approximating f in the neighborhood:

$$\min_{\boldsymbol{\theta}} \sum_{\mathbf{x}' \in \mathcal{N}(\mathbf{x})} \mathcal{L}(f(\mathbf{x}'), \boldsymbol{\theta}^\top \mathbf{x}') + \lambda \Omega(\boldsymbol{\theta}), \tag{1}$$

where $\mathcal{L}(f(\mathbf{x}'), \boldsymbol{\theta}_k^\top \mathbf{x}') = \omega(x, x')(f(\mathbf{x}') - \boldsymbol{\theta}_k^\top \mathbf{x}')^2$ is a loss function that measures how well the surrogate model approx-

imates the target model, $\omega(x, x') = exp(-D(x, x')^2/\sigma^2)$, $D(x, x')$ is the distance function (cosine distance for text and L2 distance for image or tabular data), and σ is the bandwidth. $\Omega(\cdot)$ is a regularization term that punishes complex $\boldsymbol{\theta}$ to make it sparse and ensure it can be easily understood. λ controls the trade-off between fidelity and interpretability.

4 Groupwise Explanation Framework

In this section, we introduce our GIME framework.

Vanilla group assignment. When the number of instances N is much larger than the cognitive budget K, it is infeasible for users to check local explanations for all instances. In this case, we need to learn groupwise explanations that can be applied to multiple instances. A straightforward method is to introduce a group assignment matrix $\mathbf{O} \in \{0, 1\}^{N \times K}$, where o_{ik} is 1 if $\mathbf{x}_i$ is assigned to the k-th group and is 0 otherwise. Then, Eq. (1) can be extended to a groupwise version:

$$\min_{\boldsymbol{\Theta}, \mathbf{O}} \sum_{k=1}^{K} \sum_{i=1}^{N} o_{i,k} \sum_{\mathbf{x}' \in \mathcal{N}(\mathbf{x}_i)} \mathcal{L}(f(\mathbf{x}'), \boldsymbol{\theta}_k^\top \mathbf{x}') + \lambda \sum_{k=1}^{K} \Omega(\boldsymbol{\theta}_k), \quad s.t., \sum_{k=1}^{K} o_{ik} = 1. \tag{2}$$

When $K = N$, solving Eq. (2) is equivalent to generating local explanations via Eq. (1) for each instance. When $N > K$, optimizing Eq. (2) captures K larger patterns by directly optimizing the explanation fidelity over the entire dataset.

Although Eq. (2) provides a mechanism to generate groupwise explanations (D1), it fails to help users make accurate predictions about the model behavior on unseen instances (D2). Since each explanation is characterized by a potentially large number of training instances instead of a well-defined and easy to understand closed-form expression, it is difficult for users to determine which explanation could be used for an unseen instance: users either need to make heuristic assumptions (e.g., instances with certain common features share one explanation) or need to optimize Eq. (2) again. The former method leads to low human precision [Ribeiro *et al.*, 2018] or decreased generalized fidelity [Ramamurthy *et al.*, 2020] due to the potentially problematic assumptions. The latter method is not only computationally and cognitively expensive, but also has data leakage issues and contradicts the goal of improving *generalized* fidelity, since the explanation assignment can only be determined *after* we observe the model behavior (e.g., $f(\mathbf{x})$) on the unseen instance.

Bayesian assignment. We tackle the above problem by adopting a Bayesian framework [Bishop, 2006]. Let us consider the closed-form expression for the applied region of an explanation. Without loss of generality, we define the expression by using a probabilistic distribution $p(\mathbf{x}|\boldsymbol{\psi}_k)$, where $\boldsymbol{\psi}_k$ denotes the region parameters for the k-th groupwise explanation, and $p(\mathbf{x}|\boldsymbol{\psi}_k)$ refers to the probability that the k-th explanation can be applied to a potentially unseen instance $\mathbf{x}$. Choosing different types of distribution allows us to adjust how "hard" the region assignment is. If we use a distribution like Gaussian, then whether an instance belongs to a region is probabilistic ("soft"). When we adopt distributions like the Gumbel-Softmax [Jang *et al.*, 2016], we can clearly determine whether an instance belongs to a region or not ("hard").

Let $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_K]$ denote all region parameters. Following the topic assignment probability in topic models [Blei *et al.*, 2003], we rewrite $p(\mathbf{x}_i|\boldsymbol{\psi}_k)$ as

$$p(\mathbf{x}_i|\boldsymbol{\psi}_k) = p(\mathbf{x}_i|o_{ik} = 1, \boldsymbol{\Psi}). \tag{3}$$

Then the Bayes' Rule can be used to compute the posterior probability of o_{ik}:

$$p(o_{ik} = 1|\mathbf{x}_i, \boldsymbol{\Psi}) = \frac{p(\mathbf{x}_i|o_{ik} = 1, \boldsymbol{\Psi}) p(o_{ik} = 1)}{\sum_{k'} p(\mathbf{x}_i|o_{ik'} = 1, \boldsymbol{\Psi}) p(o_{ik'} = 1)}, \tag{4}$$

where $p(o_{ik} = 1)$ is the prior that an instance be assigned to the k-th explanation. In this paper, we assume that there is no prior knowledge for preferred regions, i.e., $p(o_{ik} = 1) = p(o_{ik'} = 1)$ for $\forall k, k'$. In this scenario, Eq. (4) simply chooses explanations for an instance through normalization. When there exists prior knowledge about the preferred regions (e.g., when larger regions are considered better), we can also easily incorporate these priors with Eq. (4).

Then, we can rewrite Eq. (2) to consider the expected fidelity in all instances:

$$\min_{\boldsymbol{\Theta}, \boldsymbol{\Psi}} J(\boldsymbol{\Theta}, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \sum_{i=1}^{N} p(o_{ik}{=}1|\mathbf{x}_i, \boldsymbol{\Psi}) \sum_{\mathbf{x}' \in \mathcal{N}_i} \mathcal{L}(f(\mathbf{x}'), \boldsymbol{\theta}_k^\top \mathbf{x}') + \lambda_1 \sum_{k=1}^{K} \Omega(\boldsymbol{\theta_k}) + \lambda_2 \Omega(\boldsymbol{\Psi}). \tag{5}$$

Eq. (5) joint learns the groupwise explanations ($\boldsymbol{\Theta}$) and the regions in which they apply ($\boldsymbol{\Psi}$). Here, minimizing the first term in $J(\boldsymbol{\Theta}, \boldsymbol{\Psi})$ is equal to maximizing the expected fidelity in all instances. Note that the weight for the infidelity $\sum_{\mathbf{x}' \in \mathcal{N}_i} \mathcal{L}(f(\mathbf{x}'), \boldsymbol{\theta}_k^\top \mathbf{x}')$ is $p(o_{ik} = 1|\mathbf{x}_i, \boldsymbol{\Psi})$, which satisfies $\sum_{k=1}^{K} p(o_{ik} = 1|\mathbf{x}_i, \boldsymbol{\Psi}) = 1$ for $\forall i$. This means that the infidelity in terms of any single instance is punished. Moreover, we add $\Omega(\boldsymbol{\Psi})$, which is a regularization term that measures how easy it is for humans to understand the applied regions.

Optimization. $J(\boldsymbol{\Theta}, \boldsymbol{\Psi})$ can be optimized by using an alternative minimization algorithm. The key idea is to minimize the objective function with respect to either $\boldsymbol{\Theta}$ or $\boldsymbol{\Psi}$, while fixing the other parameter. We find that this alternative optimization method stably convergences to better solutions than directly applying gradient descent on both parameters. More specifically, we first initialize the instance groups by using K-means. We then initialize $\boldsymbol{\Theta}$ by learning an explanation for each group. Next, we iteratively apply two steps. The first step is **learning applied regions**, in which we fix $\boldsymbol{\Theta}$ and minimize $J(\boldsymbol{\Theta}, \boldsymbol{\Psi})$ with respect to $\boldsymbol{\Psi}$ by using gradient descent. The second step is **groupwise explanation generation**, in which we fix $\boldsymbol{\Psi}$ and minimize the objective with respect to $\boldsymbol{\Theta}$. Following [Ribeiro *et al.*, 2016], we set $\Omega(\boldsymbol{\theta}_k)$ to the L0 norm $||\boldsymbol{\theta}_k||_0$ and approximately optimize the objective by using the K-LASSO algorithm [Ribeiro *et al.*, 2016].

5 Region Formulation

In this section, we introduce how to formulate the region that an explanation applies, i.e., how to determine the mathemati-

cal form of distribution $p(\mathbf{x}_i|\boldsymbol{\psi}_k)$. We start by discussing the desirable properties of the regions and why intuitive solutions like Gaussian may be problematic. Then, we introduce an example distribution that satisfies the desirable properties: the vMF distribution [Banerjee *et al.*, 2005].

Desirable properties. Similar with the surrogate model g_k, to achieve high human precision, we need to define the regions to ensure both interpretability and accuracy (fidelity):

- **Interpretability** requires that it is easy for humans to understand the regions. Thus, the formulation of the region cannot be too complicated (e.g., consists of multiple nonlinear neural layers). Moreover, the number of features involved should be limited and it is more desirable if the formulation allows for modeling of sparse features.
- **Accuracy** demands that the formulation of the region is expressive enough to discriminate different regions and accurately represent the covered instances of an explanation. For example, compared with a formulation that assumes independence between features, it is more desirable if relations between features can be modeled.

Issues of the Gaussian distribution. A straightforward choice of $p(\mathbf{x}_i|\boldsymbol{\psi}_k)$ is the Gaussian distribution, in which $\boldsymbol{\Psi} = [(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), ..., (\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)]$ and

$$\begin{aligned} p(\mathbf{x}_i|\boldsymbol{\psi}_k) &= p(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &= \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_k|}} \exp(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)). \end{aligned} \tag{6}$$

Eq. (6) means that k-th explanation can be applied to instances with mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\Sigma}_k$ in the feature space. However, characterizing explanation region with Gaussian distribution has two issues. First, the interpretability of the regions is limited, because the Gaussian distribution cannot effectively constrain the number of involved features by modeling sparse features. When only part of the features are involved, $\boldsymbol{\Sigma}_k$ is not full rank, i.e., $|\boldsymbol{\Sigma}_k| = 0$, and the Gaussian distribution degenerates and does not have a density. Second, the accuracy of the regions may not be desirable, since it is difficult for the Gaussian distribution to model relationships between features. If we consider $\boldsymbol{\Sigma}_k$ as a diagonal matrix with d parameters, then Eq. (6) fails to model how correlated features compose a pattern. Learning feature correlations requires that we learn the d^2 parameters of $\boldsymbol{\Sigma}_k$, which is quite computationally expensive, especially for natural language processing tasks with tens of thousands of words (features).

Modeling with the vMF distribution. To address these issues, we introduce the von Mises-Fisher (vMF) distribution, which 1) is an example distribution that satisfies the two desirable properties and 2) achieves good empirical results. The vMF distribution is a probability distribution on a unit hypersphere in $\mathcal{R}^d$. In particular, vMF is parameterized by a direction vector $\boldsymbol{\phi} \in \mathcal{R}^d$ with $||\boldsymbol{\phi}|| = 1$ and a factor $\tau > 0$ that determines the concentration of the distribution:

$$p(\mathbf{x}_i|\boldsymbol{\psi}_k) = C(\tau) \exp(\tau \boldsymbol{\phi}_k^T \frac{\mathbf{x}_i}{||\mathbf{x}_i||}). \tag{7}$$

$C(\tau) = (2\pi)^{-d/2} \frac{\tau^{d/2-1}}{I_{d/2-1}(\tau)}$ is a normalization factor and $I_{d/2-1}(\tau)$ is the modified Bessel function of the first kind.

After normalization ($\frac{\mathbf{x}_i}{||\mathbf{x}_i||}$), all instances lie on the surface of a unit hypersphere. $\boldsymbol{\phi}_k$ denotes the mean direction vector of the instances and τ represents how concentrate the distribution is. The sign of ϕ_{kj} illustrates whether the j-th feature should be positive or negative for instances in the region, and a large $|\phi_{kj}|$ means that the instances in region k should have a large absolute value in terms of the j-th feature and that the j-feature is important for determining whether an instance belongs to region k. τ is learned together with $\boldsymbol{\phi}_k$ and larger τ denotes a more concentrated (narrow) distribution.

We can easily check that Eq. (7) is well-defined when some elements of $\boldsymbol{\phi}_k$ are zeros. This means that vMF can model sparse features and leads to better region **interpretability** compared with the Gaussian distribution. Moreover, vMF is often used by topic models [Song *et al.*, 2015] to represent the distribution of co-occurred words (features) in a topic, which illustrates its capability in modeling relationships between features and representing regions with good **accuracy**.

Based on Eqs. (3)(4)(7), we have:

$$p(o_{ik} = 1|\mathbf{x}_i, \boldsymbol{\Psi}) = \frac{\exp(h_k(\mathbf{x}_i))}{\sum_{k'} \exp(h_{k'}(\mathbf{x}_i))}, h_k(\mathbf{x}_i) = \tau \boldsymbol{\phi}_k^T \frac{\mathbf{x}_i}{||\mathbf{x}_i||}. \tag{8}$$

Eq. (8) shows that $p(o_{ik} = 1|\mathbf{x}_i, \boldsymbol{\Psi})$ can be computed without considering the complicated normalization factor $C(\tau)$.

6 Experiment

6.1 Experimental Settings

Table 1: Statistics of datasets. (TE) and (TA) denote textual and tabular datasets, respectively.

	Train	Valid/Test	Features
Polarity (TE)	7,000	1,500	43,548
Subjectivity (TE)	7,000	1,500	17,478
20 Newsgroup (TE)	1,079	358	17,980
AutoMPG (TA)	274	59	7
Wine Quality (TA)	1,119	240	11
Communities (TA)	1,395	299	100

Datasets. We use six real-world benchmark datasets. The first three are textual datasets and the last three are tabular datasets from the UCI repository [Dua and Graff, 2017]. Specifically, **Polarity** [Maas *et al.*, 2011] contains highly polar movie reviews and the task is to classify whether the sentiment of a review is positive or negative. **Subjectivity** [Pang and Lee, 2004] includes processed sentences that are labeled as either subjective or objective. **20 Newsgroup**[2] is a collection of news articles. Following [Ribeiro *et al.*, 2016], we focus on the task of determining whether an article is about Christianity or Atheism. **AutoMPG** concerns predicting fuel consumption based on attributes of cars. **Wine Quality** predicts wine quality based on physicochemical tests. **Communities** enables predicting community crimes based on socioeconomic data. Statistics of datasets are shown in Table 1.

6.2 Generalized Fidelity

Baselines. We compare GIME with five baselines. The first two baselines, **LIME** [Ribeiro *et al.*, 2016] and

[2]http://qwone.com/~jason/20Newsgroups/

Table 2: Comparison of RMSE and accuracy when explaining BERT on three textual datasets (classification task). Best results are highlighted in bold. The symbol * means that the improvements over all baselines are significant according to t-test (p-value < 0.01).

Dataset	Polarity		Subjectivity		20 Newsgroup	
	RMSE	Accuracy	RMSE	Accuracy	RMSE	Accuracy
LIME	0.349 ± 0.003	0.666 ± 0.002	0.466 ± 0.001	0.664 ± 0.005	0.411 ± 0.004	0.672 ± 0.004
MAPLE	0.318 ± 0.001	0.718 ± 0.002	0.450 ± 0.004	0.634 ± 0.003	0.328 ± 0.003	0.777 ± 0.008
ETC	0.307 ± 0.003	0.739 ± 0.003	0.370 ± 0.001	0.759 ± 0.006	0.406 ± 0.002	0.770 ± 0.006
CTE	0.349 ± 0.003	0.630 ± 0.002	0.408 ± 0.002	0.655 ± 0.004	0.388 ± 0.002	0.686 ± 0.012
GIME	$\mathbf{0.271 \pm 0.002}$*	$\mathbf{0.774 \pm 0.003}$*	$\mathbf{0.365 \pm 0.001}$*	$\mathbf{0.768 \pm 0.006}$*	$\mathbf{0.318 \pm 0.002}$*	$\mathbf{0.808 \pm 0.010}$*
Impv.	+11.7%	+4.7%	+1.4%	+1.2%	+3.0%	+4.0%

Table 3: Comparison of RMSE when explaining SVR on three tabular datasets (regression task). Best results are highlighted in bold. The symbol * means that the improvements over all baselines are significant according to t-test (p-value < 0.01).

Dataset	AutoMPG	Wine Quality	Communities
LIME	0.415 ± 0.002	0.372 ± 0.001	0.532 ± 0.001
MAPLE	0.175 ± 0.002	0.371 ± 0.001	0.282 ± 0.001
MAME	0.212 ± 0.002	0.356 ± 0.001	0.493 ± 0.001
ETC	0.368 ± 0.001	0.346 ± 0.001	0.515 ± 0.001
CTE	0.157 ± 0.002	0.283 ± 0.001	0.288 ± 0.001
GIME	$\mathbf{0.127 \pm 0.001}$*	$\mathbf{0.236 \pm 0.002}$*	$\mathbf{0.240 \pm 0.001}$*
Impv.	+19.1%	+16.6%	+14.9%

MAPLE [Plumb *et al.*, 2018], are widely-used local explanation methods. We use the submodular method in [Ribeiro *et al.*, 2016] to pick representative explanations for them. The third baseline, **MAME** [Ramamurthy *et al.*, 2020], selects multi-level representative explanations from pre-computed instance groups. We also design two other methods as baselines for groupwise explanations: **CTE** (Cluster-Then-Explain) that first determines instance groups by clustering with K-means and then learns an explanation for each group by optimizing Eq. (2), and **ETC** (Explain-Then-Cluster) that first learns local explanations and then clusters the local explanations to form instance groups. The groupwise explanations of ETC are average local explanations in each group.

Evaluation Metrics. Following [Plumb *et al.*, 2018], we evaluate the fidelity of explanations by:

$$\sum_{\mathbf{x}} \sum_{\mathbf{x}' \in \mathcal{N}(\mathbf{x})} \omega(\mathbf{x}, \mathbf{x}') \mathcal{M}(f(\mathbf{x}'), g(\mathbf{x}')), \tag{9}$$

where $\mathbf{x}$ is a test instance, the neighborhood size $|\mathcal{N}(\mathbf{x})|$ is set to 10, and the neighborhood weight $\omega(\mathbf{x}, \mathbf{x}')$ is the same as that in Eq. (1). To measure the difference between f and g, two types of $\mathcal{M}(\cdot, \cdot)$ are considered: the root mean square error (**RMSE**) and the classification **Accuracy** that evaluates whether $f(\mathbf{x}')$ and $g(\mathbf{x}')$ give the same label. Lower RMSE and higher accuracy indicate better performance.

Implementation details. We train f and learn explanations on the training set, tune hyperparameters by using the validation set, and evaluate explanations on the test set. The hyperparameters of the baselines are initialized by following the corresponding paper and tuned to achieve the optimal performance. For fair comparison, all explanation methods use the same interpretable surrogate model as in [Plumb *et al.*, 2018; Ribeiro *et al.*, 2016], i.e., $g_k(\mathbf{x}) = \boldsymbol{\theta}_k^\top \mathbf{x}$. If not specifically mentioned, K is set to 20 for large datasets (Polarity and Subjectivity), 10 for middle-sized datasets (20 Newsgroup, Wine Quality, Communities), and 4 for small datasets (AutoMPG). We ensure that all explanations have the same number of non-zero features (5 for tabular data and 50 for textual datasets). Each experiment is repeated five times and we report the average and standard deviation of RMSE and Accuracy.

Following [Ramamurthy *et al.*, 2020], we measure generalized fidelity by simulating the process that users apply explanations on unseen instances. Given an unseen instance $\mathbf{x}_i$ and explanations $g_1, ..., g_K$, simulated users leverage the explanation with the maximum probability p_{ik} to predict model behaviors on $\mathbf{x}_i$. Though p_{ik} is well defined for our method ($p_{ik} = p(o_{ik} = 1 | \mathbf{x}_i, \boldsymbol{\Psi})$) and MAPLE ($p_{ik}$ is the local training distributions), other baselines lack the mechanism for determining whether an explanation can be applied to a test instance. For these baselines, we try three formulations of p_{ik} and use the one that results in the highest validation fidelity. The first and second formulations measure how many features in $\mathbf{x}_i$ are "activated" by g_k: $p_{ik} \propto e^{s_{ik}}$, s_{ik} is $||\theta_k \odot \mathbf{x}_i||_0$ and $||\theta_k \odot \mathbf{x}_i||$ respectively, where $\odot$ denotes elementwise product. In the third formulation, s_{ik} is the negative L2 distance between $\mathbf{x}_i$ and $\mathbf{x}_k$ (local methods) or the center of instance groups (others). We find that LIME performs the best with the first formulation and others with the third one.

Overall Performance. We train **BERT** [Devlin *et al.*, 2018] and **SVR** [Awad and Khanna, 2015] on textual and tabular datasets, respectively. These models are considered as black-box models to be explained. Table 2 and 3 show the explanation fidelity on unseen instances. The results of MAME on textual datasets are omitted due to memory issues (mentioned also in [Ramamurthy *et al.*, 2020]): it cannot successfully run on a 64-bit server with 64G of memory for textual datasets.

We observe that GIME significantly outperforms other baselines in terms of fidelity on different datasets and tasks. This indicates that GIME allows users to get more faithful understanding of the model within limited cognitive budget. We also find that MAPLE and MAME generally perform better than LIME. This is because they consider more instances when generating explanations, thus their explanations may generalize to a larger region. CTE and ETC achieve the second best performance on some datasets, which illustrates the effectiveness of groupwise explanations. However, these

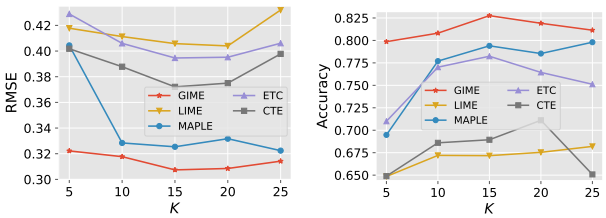

Figure 3: Fidelity at different values of K on 20 Newsgroups.

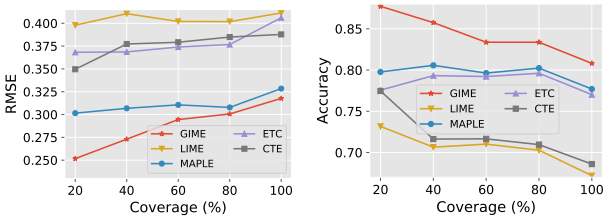

Figure 4: Fidelity vs. coverage on 20 Newsgroups.

two methods heuristically determine instance assignment via clustering, thus yielding lower fidelity compared with GIME, which jointly learns groupwise explanations and the regions that they apply in a unified framework.

Parameter Sensitivity Analysis. Fig. 3 shows how RMSE and accuracy change with the user budget K on 20 Newsgroups. We find that GIME stably performs better than baselines at different values of K, which demonstrates the robustness of our method. Moreover, the fidelity of GIME increases with increasing K at first, and then slowly decreases after $K > 15$. This shows that too many or too few explanations may hurt generalized fidelity, validating the necessity of learning a limited number of groupwise explanations.

We also evaluate how fidelity changes with different levels of coverage. An instance $\mathbf{x}_i$ is marked as covered if $\exists k, p_{ik} \geq \tau$. We vary τ to obtain the results when the percentage of unseen instances that are covered is 20%, 40%, .., and 100%. As shown in Fig. 4, GIME consistently achieves better fidelity than other methods at different levels of coverage. Moreover, the fidelity of some baselines like MAPLE does not always increase with increasing p_{ik} (decreasing coverage), which indicates that the way they determine whether an explanation can be applied is not always reliable. This may due to the fact that these methods heuristically determine p_{ik} instead of learning it jointly with the explanations. In comparison, the fidelity of GIME stably increases with larger p_{ik}, indicating that our region formulation is effective in selecting the instances that can be best interpreted by the explanation.

6.3 User Study

This experiment evaluates whether the explanations are useful in helping real users predict the model behavior on unseen instances. Specifically, we hire five native English speakers through a vendor company. For each explanation method, users are asked to first investigate the provided explanations and the information about when the explanations may be applied (ϕ_k for GIME, instance center for CTE and MAME, and the examplar instance for MAPLE). Then, the users are given five instances from the test set and asked to 1) match each instance with an explanation and 2) predict the model output for the instance. For the regression task on AutoMPG, users are only required to choose the appropriate range: [<0.5, -0.5 to 0, 0 to 0.5, and >0.5]. The explanations are provided to the users in random order and the users do not know which explanation method they are labeling. Explanations generated by GIME are shown in Table 5 and Fig. 2.

Table 4: Results of the user study. We report the results of ETC on Polarity and the results of MAME on AutoMPG.

	Polarity		AutoMPG	
	M-Acc	P-Acc	M-Acc	P-Acc
MAPLE	0.72	0.76	0.80	0.80
ETC/MAME	0.76	0.68	0.80	0.76
CTE	0.68	0.64	0.84	0.76
GIME	**0.92**	**0.84**	**0.96**	**0.92**
Impv.	+21.1%	+10.5%	+14.3%	+15.0%

Table 5: Groupwise explanations for BERT on Polarity. We show the two largest groups with top important features in characterizing applied regions, and positive/negative features in the explanations.

(a) Group 1: plots-related descriptions

Features	plot, funniest, intersting, ridiculous, horrible
Instances	"the plot is ridiculous and the characters are horrible." "attention to the details of ... in an interesting manner"

(b) Group 2: general descriptions

Features	movie, love, best, worst, terrible
Instances	"I love movies and ... turned out the best product ..." "... probably the single worst piece of trash ..."

Table. 4 compares GIME with the three most competitive baselines. **M-Acc** denotes the portion of instances that are matched to the correct explanations and **P-Acc** represents the portion of instances that are predicted correctly. We can see that GIME achieves higher M-Acc than other methods on the two datasets. It demonstrates that modeling the region with an intuitive closed-form expression helps users better understand the applied regions of explanations, which is a prerequisite for successful prediction of model behaviors on unseen instances. GIME also achieves the highest P-Acc. This validates the effectiveness of our framework in faithfully interpreting the black-box model and enabling users to make accurate predictions with a limited number of explanations.

7 Conclusion

We study two user demands that are important for understanding black-box models: 1) obtaining a faithful overall understanding of the model with limited cognitive load and 2) making accurate predictions about the model on unseen instances. To fulfill the two demands, we propose a unified Groupwise Model-agnostic Explanation framework, which learns a limited number of groupwise explanations with high fidelity as well as the region where each explanation apply. Experiments demonstrate the effectiveness of our method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61772045) and the Project 2019BD005 supported by PKU-Baidu fund.

References

[Awad and Khanna, 2015] Mariette Awad and Rahul Khanna. Support vector regression. In *Efficient learning machines*, pages 67–80. Springer, 2015.

[Banerjee *et al.*, 2005] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345–1382, 2005.

[Bishop, 2006] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[Chu *et al.*, 2018] Lingyang Chu, Xia Hu, Juhua Hu, Lanjun Wang, and Jian Pei. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1244–1253, 2018.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dhurandhar *et al.*, 2018] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in neural information processing systems*, pages 592–603, 2018.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. Uci machine learning repository. 2017.

[Elenberg *et al.*, 2017] Ethan Elenberg, Alexandros G Dimakis, Moran Feldman, and Amin Karbasi. Streaming weak submodularity: Interpreting neural networks on the fly. In *Advances in neural information processing systems*, pages 4044–4054, 2017.

[Ghorbani *et al.*, 2019] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.

[Jang *et al.*, 2016] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *ICLR*, 2016.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2018.

[Lang, 2000] Annie Lang. The limited capacity model of mediated message processing. *Journal of communication*, 50(1):46–70, 2000.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[Maas *et al.*, 2011] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, pages 142–150, 2011.

[Pang and Lee, 2004] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*, page 271, 2004.

[Plumb *et al.*, 2018] Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. Model agnostic supervised local explanations. In *Advances in neural information processing systems*, pages 2515–2524, 2018.

[Ramamurthy *et al.*, 2020] Karthikeyan Natesan Ramamurthy, Bhanukiran Vinzamuri, Yunfeng Zhang, and Amit Dhurandhar. Model agnostic multilevel explanations. In *Advances in neural information processing systems*, 2020.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Ribeiro *et al.*, 2018] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, volume 18, pages 1527–1535, 2018.

[Song *et al.*, 2015] Yangqiu Song, Shixia Liu, Xueqing Liu, and Haixun Wang. Automatic taxonomy construction from keywords via scalable bayesian rose trees. *IEEE Transactions on Knowledge and Data Engineering*, 27(7):1861–1874, 2015.