Shapley Feature Utility

Ian C. Covert
University of Washington
Seattle, WA, USA
icovert@uw.edu

Scott Lundberg
Microsoft Research
Redmond, WA, USA
slund1@cs.washington.edu

Su-In Lee
University of Washington
Seattle, WA, USA
suinlee@uw.edu

Abstract

We propose *Shapley feature utility* (SFU) as a method for quantifying the global utility of features to an optimal model. Instead of explaining individual predictions, SFU describes a feature's importance through its impact on model performance. Our approach is built on the Shapley value from cooperative game theory and leads to an elegant interpretation in terms of mutual information. We propose a sampling-based approximation and demonstrate its application to three datasets.

1 Introduction

For all machine learning models, but particularly modern black-box models, it is difficult to understand which features are most informative. Recent research on model interpretability has focused on *local attribution*, i.e., explaining how features contribute to an individual prediction [7, 10, 12]. We consider the problem of *global feature importance*, and seek to allocate credit to features for their impact on the model's accuracy.

As a motivating example, consider the problem of disease subtype classification from gene expression data. Determining the genes that enable a model to make accurate predictions may lead to improved biological understanding and can guide future work on finding therapeutic targets. However, existing methods such as feature ablations, feature selection, and local explanations do not provide a sufficiently nuanced understanding of feature importance. We require a method that is mathematically principled and that accounts for complex feature interactions.

Here, we present *Shapley feature utility* (SFU), a principled approach to quantifying how much accuracy is derived from making each feature available to the model. Building on ideas from cooperative game theory, our approach is based on the Shapley value, a principled and widely used credit allocation method [9]. We show that SFU has an elegant information-theoretic interpretation and propose a model-agnostic, sampling-based approximation.

2 Shapley Feature Utility

2.1 Model Performance as a Cooperative Game

Consider a set of d features $\{X_1, X_2, \dots, X_d\} \in \mathcal{X}$ that can be used to predict a discrete target variable $Y \in \mathcal{Y} \equiv \{1, \dots, M\}$. We use $\mathcal{S} \subseteq D \equiv \{1, 2, \dots, d\}$ to represent a set of indices, $\bar{\mathcal{S}} \equiv D \setminus \mathcal{S}$ to denote the complement, and $X_{\mathcal{S}} \equiv \{X_i \ : \ i \in \mathcal{S}\}$ to indicate a subset of features.

We now derive an approach for quantifying each feature's contribution to an optimal model's performance. Given a subset of the features $X_{\mathcal{S}}$, a model that perfectly optimizes the loss will often incur non-zero population risk. It can be shown that the optimal classification model f^* trained with cross entropy loss and features $X_{\mathcal{S}}$ outputs the conditional distribution of the response variable, or $f^*(x_{\mathcal{S}}) = p(Y|X_{\mathcal{S}} = x_{\mathcal{S}})$. The model then has the following population risk, where H represents the discrete Shannon entropy and D_{KL} represents the Kullback-Leibler divergence [1]:

Machine Learning in Computational Biology (MLCB) 2019, Vancouver, Canada.

$$
\begin{aligned}
\min_f \; -\mathbb{E}\big[\ell(f(X_\mathcal{S}), Y)\big] &= \min_f \; -\mathbb{E}\big[\log f_Y(X_\mathcal{S})\big] \\
&= \min_f \; \mathbb{E}\big[D_{\text{KL}}\big(p(Y|X_\mathcal{S})||f(X_\mathcal{S})\big)\big] + H(Y|X_\mathcal{S}) \\
&= \mathbb{E}\big[D_{\text{KL}}\big(p(Y|X_\mathcal{S})||f^*(X_\mathcal{S})\big)\big] + H(Y|X_\mathcal{S}) \\
&= H(Y|X_\mathcal{S}). \qquad (1)
\end{aligned}
$$

We can replicate this derivation for any subset of features $X_\mathcal{S}$ and find that the population risk for an optimal model is always given by the conditional entropy $H(Y|X_\mathcal{S})$. It is natural to turn this process into a cooperative game, where a player's participation corresponds to a feature being made available to the model. To satisfy the property that the empty set $\mathcal{S} = \varnothing$ evaluates to zero (a common convention for cooperative games), we define the game v as the *reduction in risk* for the target variable Y compared to the situation where no features are available:

$$
\begin{aligned}
v(\mathcal{S}) &= H(Y) - \min_f \; \mathbb{E}\big[\ell(f(X_\mathcal{S}), Y)\big] \\
&= H(Y) - H(Y|X_\mathcal{S}) \\
&= I(Y; X_\mathcal{S}). \qquad (2)
\end{aligned}
$$

We now seek to provide global feature importance scores based on this game. This can be viewed as allocating credit to each of the players (features) and a compelling tool for doing so is the Shapley value, which provides the unique allocation strategy that fulfills a set of fairness axioms [9]. Shapley values form the basis of popular local attribution methods [2, 7, 11], but this provides an approach for global interpretability that generalizes the Shapley regression approach of Lipovetsky and Conklin [5].

The Shapley values $\phi_i(v)$ for the features $i = 1, 2, \ldots, d$ in our game are:

$$
\begin{aligned}
\phi_i(v) &= \sum_{\mathcal{S} \subseteq D \setminus \{i\}} \frac{|\mathcal{S}|!(d - |\mathcal{S}| - 1)!}{d!} \big(I(Y; X_{\mathcal{S} \cup \{i\}}) - I(Y; X_\mathcal{S})\big) \\
&= \sum_{\mathcal{S} \subseteq D \setminus \{i\}} \frac{|\mathcal{S}|!(d - |\mathcal{S}| - 1)!}{d!} I(Y; X_i | X_\mathcal{S}) \qquad (3)
\end{aligned}
$$

We refer to these Shapley values $\phi_i(v)$ as the *Shapley feature utility*. They are a weighted sum of conditional mutual information terms, which represent how much information is added by incorporating X_i when $X_\mathcal{S}$ is already known.

Following this derivation, we can apply a similar approach to regression tasks trained with MSE loss. We omit a derivation, but when the cooperative game is defined as the amount of *explained variance* in Y i.e., $v(\mathcal{S}) = \text{Var}(Y) - \mathbb{E}[\text{Var}(Y|X_\mathcal{S})]$, an interpretable identity arises from the law of total variance. In both the classification and regression cases, the Shapley feature utilities satisfy the properties of being non-negative ($\phi_i(v) \geq 0$ for $i = 1, \ldots, d$) and of summing to the total reduction in uncertainty ($\sum_{i=1}^d \phi_i(v) = I(Y; X)$ in the classification case, and $\sum_{i=1}^d \phi_i(v) = \text{Var}(\mathbb{E}[Y|X])$ in the regression case).

2.2 Sampling-Based Approximation

To calculate the Shapley feature utilities $\phi_i(v)$, it seems as though we require access to an exponential number of optimal models (one for each subset of features). That is clearly infeasible, so we proceed through a sampling-based approximation that relies on a *single model*. Our approach is similar to sampling methods proposed in prior work [3], and it relies on the observation that these values (Eq. 3) involve nested expectations over data-label pairs (x, y) and feature sets $\mathcal{S} \subseteq D$ drawn from a distribution induced by the Shapley value.

We propose a sampling procedure that works in two steps in order to share computation over all features. First, we use Algorithm 1 to calculate a large number of samples, which individually approximate the loss for a particular (x, y) pair and a particular subset of features $\mathcal{S} \subseteq D$. Second, we use Algorithm 2 to process the samples and approximate the Shapley values using an importance sampling re-weighting scheme.

Combining Algorithms 1-2 results in an estimator $\hat{\phi}_i(v)$ for each Shapley feature utility value. We omit a rigorous proof, but it can be shown that for an optimal model f^* and sampling of the missing features from their conditional distribution $p(X_{\bar{\mathcal{S}}}|X_{\mathcal{S}})$, the estimator $\hat{\phi}_i(v)$ converges to the correct value $\phi_i(v)$ in probability as $n \to \infty,\ m \to \infty$ (parameters in Algorithm 1).

Unfortunately, we rarely have access to an optimal model f^* and sampling from the exact conditional distribution is usually difficult. In practice we rely on a trained model f and require an approximation to the conditional distribution, such as sampling from the marginal distribution (an assumption of feature independence) or imputing with the mean (a further assumption of model linearity) [7]. The final result is therefore a biased estimate of the value $\phi_i(v)$, and may be interpreted as the utility of different features to a *particular model* f.

Algorithm 1: Accumulating samples

```
input : trained model $f : \mathcal{X} \mapsto \mathcal{Y}$, number of samples $n$, inner loop samples $m$
SubsetList ← []
LossList ← []
for $i = 1$ to $n$ do
    sample a random permutation $O \sim \pi(D)$
    sample a number of features $N \sim$ Uniform(0, $d$)
    determine subset of features $\mathcal{S} = O[:N]$
    sample $(x^i, y^i) \sim p(X, Y)$
    yList ← []
    for $k = 1$ to $m$ do
        sample $x_{\bar{\mathcal{S}}} \sim p(X_{\bar{\mathcal{S}}}|X_{\mathcal{S}} = x^i_{\mathcal{S}})$
        yList.append($f(x^i_{\mathcal{S}}, x_{\bar{\mathcal{S}}})$)
    end
    $\hat{y}$ ← mean(yList)
    $\hat{\ell} \leftarrow \ell(y^i, \hat{y})$
    SubsetList.append($\mathcal{S}$)
    LossList.append($\hat{\ell}$)
end
return SubsetList, LossList
```

Algorithm 2: Calculating Shapley feature utilities

```
input : SubsetList, LossList
for $i = 1$ to $d$ do
    LossIncluded ← $\left[\frac{\hat{\ell}(d+1)}{2|\mathcal{S}|}\right.$ for ($\mathcal{S}$, $\hat{\ell}$) in iterate(SubsetList, LossList) if $\left.i \in \mathcal{S}\right]$
    LossDiscluded ← $\left[\frac{\hat{\ell}(d+1)}{2(d-|\mathcal{S}|)}\right.$ for ($\mathcal{S}$, $\hat{\ell}$) in iterate(SubsetList, LossList) if $\left.i \notin \mathcal{S}\right]$
    $\hat{\phi}_i(v)$ ← mean(LossDiscluded) − mean(LossIncluded)
end
return $\hat{\phi}_1(v), \dots, \hat{\phi}_d(v)$
```

3 Experiments

In this section, we apply the Shapley feature utility method to several datasets. Our experiments are conducted on the MNIST digit recognition dataset [4], gene expression microarray data from the Cancer Genome Atlas (TCGA) [13], and mortality data from the NHANES survey [8]. For the gene expression data, which was collected from breast cancer patients, we confined our analysis to 50 genes. Several genes were chosen for known breast cancer associations, and the remaining genes were chosen at random. For the NHANES data, we omitted features that were frequently missing. A model was required for each dataset, so we trained a multi-layer perceptron (MLP) for

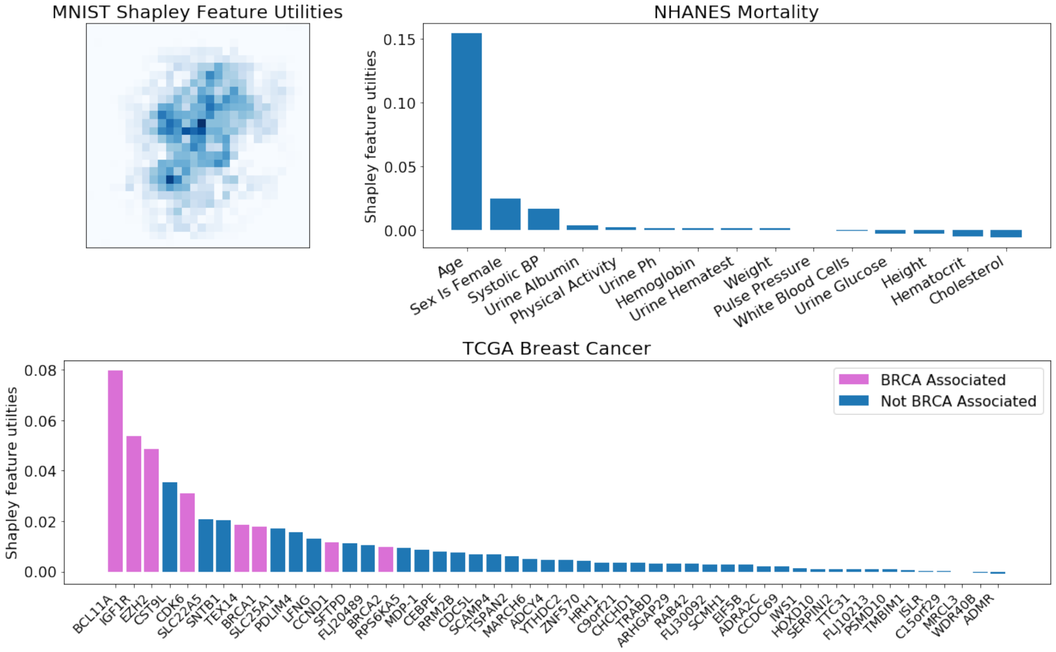

Figure 1: Shapley feature utility estimates for MNIST, NHANES and TCGA.

digit recognition (97.5% accuracy), a MLP to classify among four breast cancer subtypes for the gene expression data (79.3% accuracy), and a logistic regression model to predict whether NHANES participants lived beyond ten years (81.2% accuracy).

We then estimated Shapley feature utilities for each dataset using the proposed sampling method. When running Algorithm 1, we imputed missing features with their mean for MNIST and TCGA, and sampled from the marginal distribution for NHANES, due to its categorical features. The results are shown in Figure 1. The MNIST plot shows that, naturally, only pixels near the center have high utility to the model. The NHANES results are intuitive, showing that age contains the most signal for predicting mortality, followed by gender and blood pressure, which corroborrates findings from prior work [6]. For the TCGA data, we manually marked the genes with documented breast cancer associations, and we observe that these genes generally receive the highest Shapley feature utilities.

In a final experiment, we investigated the convergence properties of our sampling method. We ran the algorithm several times for MNIST and TCGA to examine the estimator's variance (Figure 2 right) and the Spearman correlation with the final estimates, based on one billion samples (Figure 2 left). The estimates converge with more samples, but more slowly for MNIST due to a larger number of features ($d = 784$ vs. $d = 50$). With MNIST, we observe that using the smaller validation set resulted in less estimator variance, but not faster convergence to the correct values. The required number of samples for a satisfactory result is best represented by the Spearman correlation plot: to achieve a rank correlation of $\rho = 0.9$, roughly 10^6 samples are required for TCGA, and 10^8 for MNIST. In our implementation, these require 2 minutes and 180 minutes, respectively.

4 Discussion

In this work, we proposed Shapley feature utility (SFU) as a method to quantify the global importance of each feature to a model. We presented a sampling-based approximation and applied it to three datasets, finding plausible results in each case. Future work will focus on developing faster estimation techniques and better approximations to the conditional distribution, providing a more thorough characterization of SFU's properties, comparing it quantitatively with alternative feature importance measures, and exploring high-impact applications.

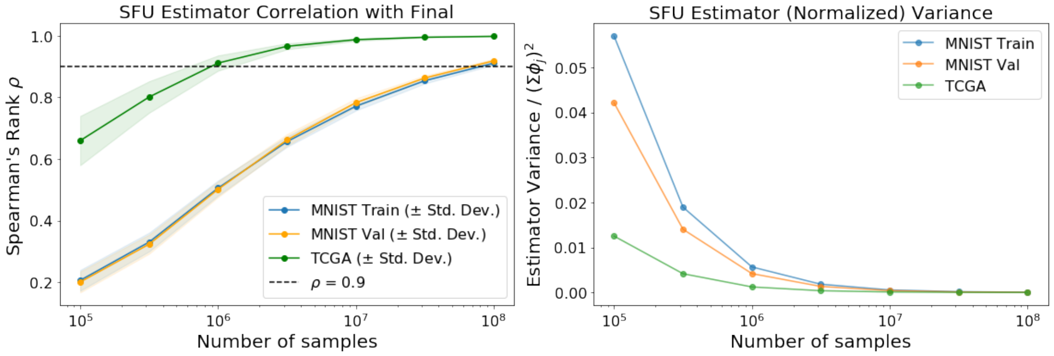

Figure 2: Convergence analysis of Shapley feature utility estimator.

References

[1] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[2] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617. IEEE, 2016.

[3] Igor Kononenko et al. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(Jan):1–18, 2010.

[4] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[5] Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

[6] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[7] 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.

[8] Henry W Miller. Plan and operation of the health and nutrition examination survey, united states, 1971-1973. *DHEW publication no.(PHS)-Dept. of Health, Education, and Welfare (USA)*, 1973.

[9] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[10] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[11] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

[12] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017.

[13] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.