Machine Learning and Statistical Learning

Explainable machine learning model with SHAP

Ewen Gallic
ewen.gallic@gmail.com

MASTER in Economics - Track EBDS - 2nd Year

1. Introduction

Introduction

With supervised learning techniques, we build a model link some explanatory variables $\mathbf{x}$ to an output y we would like to be able to predict.

We have seen so far the problem can be written as follows:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i,$$

where $f(\cdot)$ is an unknown function and ε_i is a random term.

Interpretation

Some algorithms, such as linear regression, logistic regression, Lasso, ..., because of their **monotonicity constraints**, make it relatively easy to **interpret** the effects of a variation in an explanatory variable on the variable to be explained.

This monotonicity constraint ensures that the realtion between an explanatory variable and the variable to be explained **always goes in the same direction**, regardless of the value of $\mathbf{x}$.

It is therefore possible to explain the effects of a variation of $\mathbf{x}$ on y.

Interpretation

But the monotonicity constraint is not insured in many models (*e.g.*, in random forests, neural networks, ...).

The interpretation may become a very difficult task, if at all understandable to a human brain.

What is the goal of the estimation?

In some situations, the first **aim** may not be the interpretability:

- if I want to be able to recognize a human face on a picture, my first goal is not to be able to explain why my algorithm is able to correctly recognize the presence or absence of a face

But in some other situations, my focus may be on interpretability:

- if I want to predict whether an individual is likely to perform an action, or is likely to develop a disease of some kind, then I am more interested in the characteristics that help explain the result returned by my algorithm

Trade-off

If interpretability is a crucial point, it may thus be tempting to turn to **interpretable models**, such as linear regression or logistic regression.

The problem is that the **predictive capabilities** of such models are generally relatively lower than with more flexible models.

Usually, we face a **trade-off** between :

- a simple model with relatively low predictive abilities but which is interpretable
- a more flexible model with relatively high predictive abilities but which is hardly (if at all) interpretable

Black Box Models

Very flexible Machine Learning algorithms are usually a black box layer.

Even if one understands the principles and mathematics behind Machine Learning models, the prediction returned by the model can be **far too complex to be understood by a human brain**.

It is then possible to resort to methods of interpretability, which make it possible to explain the prediction provided by machine learning algorithms.

This chapter presents one of these methods: SHAP values.

References for this part of the course

- Interpretable machine learning. A Guide for Making Black Box Models Explainable" (Molnar 2019)
- A Unified Approach to Interpreting Model Predictions. (Scott M. Lundberg and Lee 2017)
- Explainable machine-learning predictions for the prevention of hypoxaemia during surgery (Scott M. Lundberg et al. 2018)
- Shapley Value (Hart 1989)
- Dr. Dataman (2019), Explain Your Model with the SHAP Values, towards data science.
- Cooperative Games and the Shapley Value, Youtube video by Vincent Knight

2. Shapley Values

Definition

> *The Shapley value is a solution concept in cooperative game theory. It was named in honor of Lloyd Shapley, who introduced it in 1951 and won the Nobel Prize in Economics for it in 2012. To each* ***cooperative game*** *it assigns* ***a unique distribution*** *(among the players) of a* ***total surplus*** *generated by the coalition of all players. The Shapley value is characterized by a collection of desirable properties*

Definition from Wikipedia.

In other words, Shapley values represent the marginal contribution of an agent, all coalitions having been considered.

Link with Machine Learning

What is the link between Shapley values and Machine Learning?

- The **game**: the prediction task made by the algorithms
- The **players**: the different explanatory variables
- The **total surplus**: the prediction made by the algorithm for a set of players with given characteristics.

The next section will introduce SHAP values, which aim at providing the **marginal contribution** of the value of an explanatory variable across all possible coalitions.

But before that, let us give more details about Shapley values.

Cooperative Games

In Game theory, there are two main branches of games:

- **non-cooperative games**: each player of the game tries to achieve their own goal (*e.g.*, increasing their utility, or reducing their costs)
  - those games can thus be seen as a competition between individual players
- **cooperative games**: the goal (*e.g.*, utiliy or costs) will not only depend on the strategy of a single player, but it will also be affected by the strategy of players within a coalition
  - those games can thus be seen as a competition **between coalitions of players**

Shapley Values: set up

Let us consider three person, Aldah, Bourgeot and Édouard who join on a Saturday afternoon for an escape game.

The escape game gives a reward depending on the amount of time spent in the room: the shorter it is, the larger the payout.

Aldah, Bourgeot and Édouard do not contribute the same to the game. How much of the payout should be given to them?

Shapley Values: set up

Let us suppose we know the following information:

1. Playing alone provides the following payout:
   - Aldah: 80
   - Bourgeot: 56
   - Édouard: 70
2. Playing with another people:
   - Aldah and Bourgeot: 80
   - Andres and Édouard: 85
   - Bourgeot and Édouard: 72
3. All three together:
   - Aldah, Bourgeot and Édouard: 90

Shapley Values: set up

Let us use some formal notations to describe the situation.

There are $N = 3$ players that play a game G given by a pair (N, v), where v: $2^{[N]} \to \mathbb{R}$ is a characteristic function that provides a payoff to every coalition of players.

Let us denote v the payout:

$$v(c) = \begin{cases} 80 & \text{if } c = \{A\} \\ 56 & \text{if } c = \{B\} \\ 70 & \text{if } c = \{E\} \\ 80 & \text{if } c = \{A, B\} \\ 85 & \text{if } c = \{A, E\} \\ 72 & \text{if } c = \{B, E\} \\ 90 & \text{if } c = \{A, B, E\} \end{cases}$$

Shapley Values

If we denote $\mathcal{S}$ a coalition of players, then $v(\mathcal{S})$ gives the payout for that coalition.

Note: $v(\cdot)$ is such that $v(\emptyset) = 0$.

We want to know how to distribute the payout to the players of the game (we assume they all collaborate). The Shapley value gives a way of doing so.

The amount player i should receive is given by:

$$\phi_i(v) = \sum_{\mathcal{S} \in N/\{i\}} \frac{\mid \mathcal{S} \mid !(n - \mid S \mid -1)!}{n!} \left(v(\mathcal{S} \cup \{i\}) - v(\mathcal{S})\right),$$

where n is the total number of players.

Shapley Values: alternative formula

An alternative formula for the Shapley value is given by:

$$\phi_i(v) = \frac{1}{N!} \sum_{\pi \in \Pi_n} \Delta_\pi^G(i)$$

it can therefore be interpreted as the **average of the marginal contribution of the players over all contributions of the players**.

How much should everyone pay?

Let us get back to our example. When they all play together, the payout is 90. If we want to know how much each of them should get from the reward:

- we can look at all the permutations in sequence
- for each permutation, we can compute the marginal payout of Aldah, Bourgeot, and Édouard

Example with a sequence:

- When Aldah plays alone, she gets 80
- When Bourgeot joins Aldah, they get 80 as well (no additional value for Bourgeot here)
- When Édouard joins, the reward rises to 90 -> additional payout for Édouard: 10
- The marginal payout value is therefore, for that sequence : $(80, 0, 10)$

How much should everyone pay?

Let us consider a second example:

- When Aldah plays alone, she gets 80
- When Édouard joins, they get 85 (there is an additional value of 5)
- When Bourgeot joins, they get 90 (there is an additional value of 5)
- The marginal payout value is therefore, for that sequence: $(80, 5, 5)$.

We just need to consider all the permuation of the players.

How much should everyone pay?

We therefore consider every permuation π of players and compute the cost of each

π	Δ_π^G
(A, B, E)	$(80, 0, 10)$
(A, E, B)	$(80, 5, 5)$
(B, A, E)	$(24, 56, 10)$
(B, E, A)	$(18, 56, 16)$
(E, A, B)	$(15, 5, 70)$
(E, B, A)	$(18, 2, 70)$

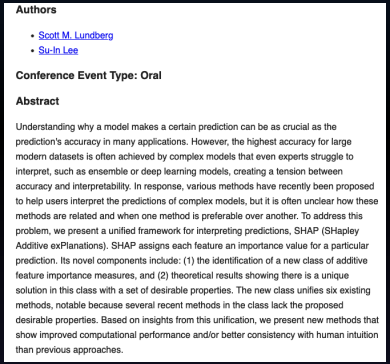
We just need to take the average of these marginal payouts: $(39.16667, 20.66667, 30.16667)$

These are the Shapley values. The reward of 90 should be divided, according to the Shapley value, as follows: Aldah gets 39.16, Bourgeot 20.68 and Édouard 30.16.

3. SHAP

SHAP: NIPS paper

Now, let us turn to how this game theory concept can be useful in Machine Learning.
In 2017, Scott Lundberg and Su-In Lee presented a paper called A Unified Approach to Interpreting Model Predictions during the annual Neural Information Processing Systems Conference (Scott M. Lundberg and Lee 2017). In this paper, they introduce a unified framework for interpreting predictions, called SHAP (SHapley Additive exPlanations).



Authors

- Scott M. Lundberg
- Su-In Lee

Conference Event Type: Oral

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between accuracy and interpretability. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

SHAP: intuition

Let us suppose that we have trained a model that predicts the probability of an insured having a car accident within a given year.

We can view each explanatory variable as a player in a cooperative game where the payout is the predicted probability of having a car accident.

SHAP: intuition

Let us say that our model accounts for four variables:

- the exposition (for example, how many kilometers the insured drives each month, on average) ;
- the age of the insured
- the type of the car they drive
- the region in which the insured lives.

SHAP: intuition

Let us consider a specific insured, with the following characteristics:

- exposition: 1,200km per month
- age: 32
- type of the car: SUV
- region: Provence Alpes Cote d'Azur

Let us say that the model predicts:

- a 10% probability of having a car accident within the year for that driver
- an average probability of having a car accident for all drivers within the sample of 7%.

How much does each variable contribute to the prediction of 10% compared to the average prediction of 7% ?

SHAP: intuition

How can we explain the $+3$ percentage points difference?

Maybe (let us imagine so), we face the following situation:

- the exposition contributed to 15%
- the age contributed to -13%
- type of the car contributed to 0%
- the region contributed 1%

The contributions, when summed, amount to 3%, *i.e.*, the final prediction for that individual (10%) minus the average predicted probability of having a car accident in the sample (7%).

SHAP: intuition

To sum up the situation:

- the game played here is the prediction task of the model for a single individual in the sample
- the players are the explanatory variables
- the payout is the prediction of the model for some given characteristics of the individual.

The shapley value for each variable

- will be equal to the difference between the prediction for a specific individual and the average prediction in the sample
- can be viewed as the contribution of each explanatory variable to move the prediction away from its expected value.

SHAP: intuition

Let us go back to the example, where the contributions could be:

- the exposition contributed to 15%
- the age contributed to -13%
- type of the car contributed to 0%
- the region contributed 1%

How can these contributions be calculated?

SHAP: intuition

Let us focus on the contribution of the region=PACA value, when it is added to a coalition of age=32 and type of car=SUV.

Let us focus on a coalitions of age, type of car and region (thus excluding exposition here)

- Step 1 (with the value region=PACA)
  - Another indivdual is drawn from the sample
    - the value for exposition is randomly drawn (let us say 1,000km per month)
    - age is set to 32
    - type of car is set to SUV
    - region is set to PACA
    - we compute the prediction for such an individual: 9%

SHAP: intuition

- Step 2 (possibly without the value region=PACA)
  - we remove region=PACA from the coalition and replace it with a random draw (it can be PACA or something else)
  - all other characteristics remain as in step 1
  - we compute the prediction for this new individual: 7%
- Step 3:
  - we compute the contribution of region=PACA as 9% - 7% = +2%

Then we repeat steps 1 to 3 to compute an average of the contributions.

SHAP: intuition

This iteration over these three steps considers just 1 coalition (here, age, type of car and region). All coalitions need to be considered, as the Shapley value corresponds to the average of all the marginal contributions to **all possible coalitions**:

- without any value for the explanatory variables
- exposition=1,200km
- age=32
- car=SUV
- exposition=1,200km + age=32
- exposition=1,200km + car=SUV
- age=32 + car=SUV (coalition previously illustrated)
- exposition=1,200km + age=32 + car=SUV

SHAP: intuition

The prediction with and without region=PACA is computed in each of the 8 different situations. The difference between the two gives the marginal contribution.

The **Shapley value** can then be calculated as the weighted average of marginal contributions.

SHAP: more formally

The usual SHAP value for one individual variable $\{i\}$ taken in a set of variables F is defined as follows:

$$\phi_i = \sum_{S \subseteq F \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right), \tag{1}$$

- f_S (resp., $f_{S \cup \{i\}}$): prediction function corresponding to the model trained using the variables contained in the subset S (resp., $S \cup \{i\}$)
- x_S (resp., $x_{S \cup \{i\}}$): observation of the values of variables contained in the subset S (resp., $S \cup \{i\}$)
- $\phi_i \in \mathbb{R}$: weighted average of the marginal contributions of variable $\{i\}$

SHAP: in application

From the previous formula, we understand that for a single individual, all possible coalitions of explanatory variables values need to be considered with and without the variable for which we want to compute the Shapley value: this is too costly in computational power!

The exact Shapley value can be estimated using Monte-Carlo sampling.

More details are providded in Scott M. Lundberg and Lee (2017) and Molnar (2019).

Remark: be careful with the interpretation: the Shapley values **do not give an estimation of a causal** effect.

SHAP: in application

See exercise on SHAP.

4. References

Hart, Sergiu. 1989. "Shapley Value." In *Game Theory*, edited by John Eatwell, Murray Milgate, and Peter Newman, 210–16. The New Palgrave. London: Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-20181-5_25.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Lundberg, Scott M., Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, et al. 2018. "Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia During Surgery." *Nature Biomedical Engineering* 2 (10): 749–60. https://doi.org/10.1038/s41551-018-0304-0.

Molnar, Christoph. 2019. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.