SHAP and Shapley Values

Quang Nguyen

Loyola University Chicago

Outline

- Flexibility-Interpretability Trade-Off
- Interpretable Machine Learning
- SHAP
- Shapley values
- Examples
- Resources

Model Interpretability

- Prediction accuracy - model interpretability trade-off is a big idea in ML.
- In general, interpretability decreases as flexibility increases (James et al., 2021)
- It is difficult to interpret more complex/black-box models (e.g. random forest, gradient boosted trees)

Model Interpretability

- The demand for model interpretability has increased in recent years.
- IML (interpretable machine learning) has emerged as a new area of research
- This has become an integral part of the machine learning pipeline
- More and more methods has been developed
  - LIME (Ribeiro et al., 2016)
  - **SHAP** (Lundberg & Lee, 2017)
  - BreakDown (Staniak & Biecek, 2018)
  - LEAF (Amparore et al., 2021)

SHAP

- **SH**apley **A**dditive ex**P**lanations
- Goal: explain the prediction of an observation x_{obs} by computing the contribution of each feature to the prediction

$$\hat{f}(x_{obs}) - \sum_{i=1}^{N} \hat{f}(x_i)$$

- Can be applied on a local level (a single row) and global level (aggregated into variable importance summaries)
- SHAP satistifies the following properties
  - Local accuracy
  - Missingness
  - Consistency

SHAP

- Based on **Shapley values** (Shapley, 1951)
  - Originated from game theory, named after Lloyd Shapley
  - Average marginal contribution of a player across all possible coalitions in a game

$$\phi_i(v) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{i\}} \frac{|S|!\,(p - |S| - 1)!}{p!} (v(S \cup \{i\}) - v(S))$$

- In a prediction setting
  - **"Game"**: prediction task for an x_{obs}
  - **"Players"**: feature values of x_{obs} that collaborate to receive gain/payout (i.e. predict a certain value)
  - **"Gain/Payout"**: difference between predicted x_{obs} and average prediction for all training observations

Disadvantages

- SHAP and Shapley values are computationally expensive
  - In most software/packages, approximations are used
- Shapley values can be misinterpreted

  × The difference of the predicted value after removing the feature from the model training

  ✓Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value.

Example

- **Random forest model on `titanic` dataset**
- `fastshap` (Greenwell, 2020) R package
- Data prep

```
library(tidyverse)
theme_set(theme_minimal())
```

```
titanic <- titanic::titanic_train %>%
  janitor::clean_names() %>%
  dplyr::select(survived, pclass, sex, age, embarked) %>%
  tidyr::drop_na() %>%
  dplyr::mutate(survived = as_factor(survived),
                across(where(is_character), as_factor))
```

Example

- RF fit

```
titanic_rf <- ranger::ranger(survived ~ .,
                             probability = TRUE,
                             data = titanic)
```

- Explain

```
surv_prob <- function(object, newdata) {
  predict(object, newdata)$predictions[, 2]
}
x_train <- dplyr::select(titanic, -survived)
```

```
titanic_explain <- titanic_rf %>%
  fastshap::explain(X = data.frame(x_train),
                    nsim = 100,
                    adjust = TRUE,
                    pred_wrapper = surv_prob)
```

Example

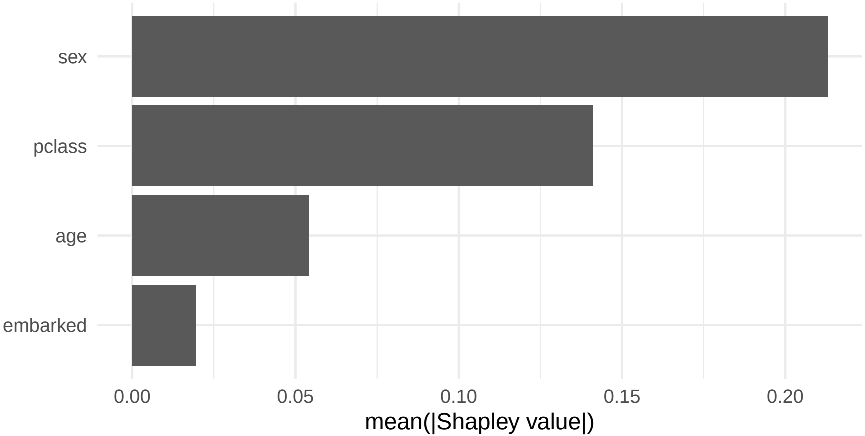
```
head(titanic_explain)
```

pclass	sex	age	embarked
-0.1049205	-0.1498677	-0.0302120	-0.0193185
0.2022373	0.3525334	-0.0075687	0.0315030
-0.1659573	0.2336614	0.0270597	-0.0300617
0.2110946	0.3145879	0.0454336	-0.0050744
-0.1171851	-0.1612892	-0.0391471	-0.0059721
0.1499265	-0.2048499	-0.1306206	-0.0143906

Example

- Variable importance, global level

```
autoplot(titanic_explain)
```



Example

- Explaining an individual observation: **Jack Dawson**

```r
jack <- data.frame(pclass = 3,
                   sex = "male",
                   age = 20,
                   embarked = "S")
```

```r
jack_prob <- surv_prob(titanic_rf, jack)
jack_prob
```

```
        1
0.1188755
```

```r
baseline_prob <- mean(surv_prob(titanic_rf, x_train))
baseline_prob
```

```
[1] 0.4067117
```

Example

```r
jack_explain <- fastshap::explain(titanic_rf,
                                  X = x_train,
                                  newdata = jack,
                                  nsim = 1000,
                                  adjust = TRUE,
                                  pred_wrapper = surv_prob)
jack_explain
```

pclass	sex	age	embarked
-0.1157425	-0.18642	0.0603141	-0.0459879

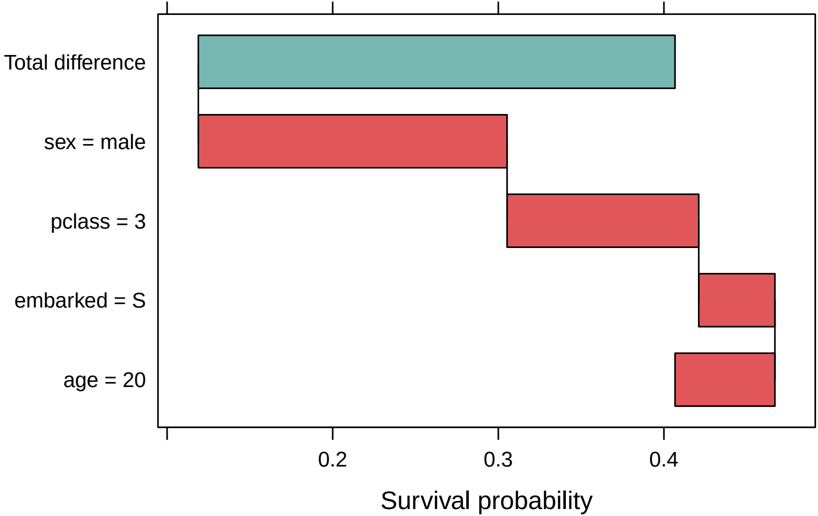
Example

- Waterfall chart

```r
feat_shap <- tibble(feature = str_c(names(jack), " = ", t(jack)),
                    shapley = t(jack_explain))

waterfall::waterfallchart(feature ~ shapley,
                          data = feat_shap,
                          origin = baseline_prob,
                          summaryname = "Total difference",
                          xlab = "Survival probability")
```

Example


Total difference
sex = male
pclass = 3
embarked = S
age = 20
0.2
0.3
0.4
Survival probability

Resources

- Book:
  - Interpretable Machine Learning (Molnar, 2019)
- Papers:
  - Explainable Artificial Intelligence: a Systematic Review (Vilone & Longo, 2020)
  - Landscape of R packages for eXplainable Artificial Intelligence (Maksymiuk et al., 2021)

References I

Amparore, E., Perotti, A., & Bajardi, P. (2021). To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods. *PeerJ. Computer Science*, *7*, e479–e479. https://doi.org/10.7717/peerj-cs.479

Greenwell, B. (2020). *Fastshap: Fast approximate shapley values*. https://CRAN.R-project.org/package=fastshap

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in r* (2nd ed.). Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777. https://dl.acm.org/doi/10.5555/3295222.3295230

Maksymiuk, S., Gosiewska, A., & Biecek, P. (2021). *Landscape of r packages for eXplainable artificial intelligence*. https://arxiv.org/abs/2009.13248

References II

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book

Ribeiro, M., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101. https://doi.org/10.18653/v1/N16-3020

Shapley, L. S. (1951). *Notes on the n-person game - I: Characteristic-point solutions of the four-person game*. RAND Corporation. https://doi.org/10.7249/RM0656

Staniak, M., & Biecek, P. (2018). Explanations of Model Predictions with live and breakDown Packages. *The R Journal*, *10*(2), 395–409. https://doi.org/10.32614/RJ-2018-072

Vilone, G., & Longo, L. (2020). *Explainable artificial intelligence: A systematic review*. https://arxiv.org/abs/2006.00093