University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

Model Agnostic Explainability Methods: Permutation Feature Importance

Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

WORLD CHANGING GLASGOW

Model Agnostic Approaches - Advantages

- Model Flexibility
- Explanation Flexibility
- Representation Flexibility

Model Agnostic Approaches

- **Permutation Feature Importance**
- Local Interpretable Model-agnostic Explanations
- Shapley Additive Explanations

Permutation Feature Importance (PFI)

- **Permutation feature importance (PFI)** is a model inspection technique that can be used for any fitted estimator.
- This is especially useful for **non-linear or black-box estimators**.
- The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled.
- This procedure breaks the **relationship between the feature and the target**, thus the drop in the model score is indicative of how much the model depends on the feature.

Permutation Feature Importance (PFI)

- The PFI algorithm is outlined as followed:
  - Inputs: Fitted predictive model ***m*** and dataset ***D***.
  - Compute the reference score ***s*** of the model ***m*** on data ***D*** (for instance the accuracy for a classifier or the R^2 for a regressor).

  - For each feature ***j*** and for each repetition ***k*** in **1,...,K** :
    - Randomly shuffle column ***j*** of dataset ***D*** to generate a corrupted version of the data named $D_{k,j}$.
    - Compute the score $s_{k,j}$ of model m on corrupted data $D_{k,j}$.
    - Compute importance i_j for feature f_j defined as:

$$\boxed{i_j = s - \frac{1}{K} \sum_{k=1}^{K} s_{k,j}}$$

Permutation Feature Importance (PFI)

Algorithm 1

Algorithms for PermFIT

```
Randomly divide the data into K folds.
for k = 1 to K do.
    Denote the data in k^th fold as V_k and the rest of the data as \overline{V}_k.
    Build the machine learning model with \overline{V}_k, denoted as \hat{μ}_k(·).
    for j = 1 to p do
        Calculate \hat{M}_ij^(P,CV) for subjects in D_k.
    end for
end for
for j = 1 to p do
    Calculate \hat{M}_j^(P,CV) and estimate \widehat{Var}[\hat{M}_j^(P,CV)].
end for
```

$$\boxed{\widehat{M}_{ij}^{(P,CV)} = \sum_{k=1}^{K} \mathrm{I}(i \in V_k)\left[\left\{Y_i - \widehat{\mu}_T\left(X_{i\cdot}^{(j)}\right)\right\}^2 - \left\{Y_i - \widehat{\mu}_k(X_{i\cdot})\right\}^2\right]}$$

Mi et al. ‘Permutation-based identification of important biomarkers for complex diseases via machine learning models’, Nature Communications 2021

PFI - Disadvantages

- An in-depth understanding of the model decision is not possible
- The interaction between features via the original model is not taken into consideration
- Exact/local explanations may be required due to legal or ethical reasons

Summary

- Conceptually simple, yet powerful global ‘explainability’ method.
- PFI explains the complete dataset and not individual samples.
- It can provide a score of how important an input variable is to the prediction
- It depends on reshuffling features, adding randomness to the data measurements.

References

- Ribeiro et al. ‘Model-Agnostic Interpretability of Machine Learning’, ICML Workshop on Human Interpretability in Machine Learning, 2016.
- Mi et al. ‘Permutation-based identification of important biomarkers for complex diseases via machine learning models’, Nature Communications, 2021.