



Methods: Permutation Feature Importance Model Agnostic Explainability

Lecturer (Assistant Professor) fani.deligianni@glasgow.ac.uk Dr. Fani Deligianni,

https://www.gla.ac.uk/schools/computing/staff/

Lead of the Computing Technologies for Healthcare Theme





Model Agnostic Approaches - Advantages

- Model Flexibility
- Explanation Flexibility
- Representation Flexibility



Model Agnostic Approaches

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

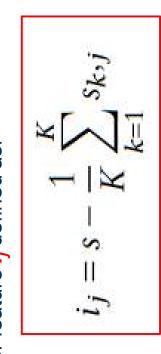
Permutation Feature Importance (PFI)

- 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.
- thus the drop in the model score is indicative of how much the model depends on This procedure breaks the relationship between the feature and the target, 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 R² 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 $\mathbf{s}_{k,j}$ of model m on corrupted data $\mathbf{D}_{k,j}$.
- Compute importance i, for feature f, defined as:





2010)

Permutation Feature Importance (PFI)

Algorithm 1

Algorithms for PermFIT

1: Randomly divide the data into K folds.

2: for k = 1 to K do.

3: 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 $\widehat{\mu}_k(\cdot)$.

 $\begin{aligned} & \textbf{for} \ j = 1 \ \textbf{to} \ \underset{ij}{\underline{p}} \ \overset{\textbf{do}}{(P,CV)} \\ & \text{Calculate} \ \overset{\textbf{M}}{M}_{ij}^{} \ \text{for subjects in } \mathcal{D}_k. \end{aligned}$

end for

8: end for

9: **for** j = 1 **to** p **do**10: Calculate $\widehat{M}_{j}^{(P,CV)}$ and estimate $\widehat{\operatorname{Var}}\Big[\widehat{M}_{j}^{(P,CV)}\Big]$.

11: end for

$$\widehat{M}_{ij}^{(P,CV)} = \sum_{k=1}^K \mathbf{I} \big(i \in V_k \big) \left[\left\{ Y_i - \widehat{\mu}_T \Big(X_i^{(j)} \Big) \right\}^2 - \left\{ Y_i - \widehat{\mu}_k (X_i.) \right\}^2 \right]$$



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