Exercise 1: Permutation feature importance

Permutation Feature Importance is one of the oldest and most widely used IML techniques. It is defined as

$$\widehat{PFI}_S = \frac{1}{m}\sum_{k=1}^{m} \mathcal{R}_{\text{emp}}(\hat{f}, \tilde{\mathcal{D}}^S_{(k)}) - \mathcal{R}_{\text{emp}}(\hat{f}, \mathcal{D})$$

where $\tilde{\mathcal{D}}^S_{(k)}$ is the dataset where features S were replaced with a perturbed version that preserves the variables marginal distribution $P(X_S)$. We can approximate sampling from the marginal distribution by random permutation of the original feature's observations.

(a) PFI has been criticized to evaluate the model on unrealistic observations. Describe in a few words why this extrapolation happens, e.g. using an illustrative example.

(b) Under a (seldomly realistic) assumption PFI does not suffer from the extrapolation issue. What is that assumption? Briefly explain why.

(c) Download the `extrapolation.csv` dataset. Fit an unregularized ordinary least squares linear regression model without interactions to the data. Do not look at the model's coefficients or perform an exploratory analysis of the data yet. Assess the MSE of the model on test data.

(d) Implement Permutation Feature Importance. Apply Permutation Feature Importance to the model (on test data) and plot the results using a barplot with an error bar indicating the standard deviation. In order to make your code reusable for the upcoming exercises, break down the implementation into three functions:

  (i) `pfi_fname` which returns the PFI for a feature `fname`

  (ii) `fi_naive` a function that computes the importances for all features using `fi_fname`

  (iii) `n_times` a function that repeats the computation n times and returns mean and standard deviation of the importance values

(e) Interpret the PFI result. What insight into model and data do we gain?

  (i) Which features are (mechanistically) used by the model for it's prediction?

  (ii) Which features are (in)dependent with Y?

  (iii) Which features are (in)dependent with its covariates?

  (iv) Which features are dependent with Y, given all covariates?

(f) Perform an exploratory analysis of the data (correlation structure between features and with y) and print the model's coefficient and intercept. Compare your PFI interpretation with the ground truth.

(g) What additional insight into the relationship of the features with y do we gain by looking at the correlation structure of the covariates in addition to the PFI (assuming that all dependencies are linear)?

(h) Demonstrate the extrapolation problem on a dataset of your choice, e.g. on the extrapolation.csv dataset. *Hint:* For the extrapolation dataset all dependencies can be assumed to be pairwise. In order to assess the data distribution before and after perturbation, you can therefore do pairwise density or scatterplots before and after perturbing the features of interest.

Exercise 2: Conditional sampling based feature importance techniques

Conditional Feature Importance and conditional SAGE value functions have been suggested as an alternative to Permutation Feature Importance.

(a) Implement a linear Gaussian conditional sampler. For conditional sage values the sampler must be able to learn Gaussian conditionals with multivariate conditioning set and multivariate target, whereas for conditional feature importance the target can be assume to be univariate. *Advice:* For multivariate Gaussian data, the conditional distributions can be derived analytically from mean vector and covariance matrix, see here.

(i) First, learn a function that returns the conditional mean and covariance structure given specific values for the conditioning set. Given the decomposition of the covariance matrix as

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}, \tag{1}$$

the distribution of X_1 conditional on $X_2 = a$ is the multivariante normal $N(\bar{\boldsymbol{\mu}}, \overline{\mathbf{\Sigma}})$

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{a} - \boldsymbol{\mu}_2) \tag{2}$$

$$\overline{\mathbf{\Sigma}} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}. \tag{3}$$

(ii) Then write a function, that takes the conditional mean and covariate structure and allows to sample from the respective (multivariate) Gaussian.

(b) Using the sampler, write a function that computes CFI (You can assume that the data is multivariate Gaussian).

(c) Apply CFI to the dataset and model from Exercise 1. Interpret the result (insights into model and data) and compare the result to PFI.

(d) Write a function that computes conditional SAGE values.

(e) Apply the conditional SAGE values with respect to an empty coalition and with respect to all remaining variables to the dataset and model from Exercise 1. Interpret the result (insights into model and data) and compare it to CFI and PFI.

Exercise 3: Refitting based importance

We can also assess the importance of a feature by refitting the model with and without access to the feature of interest and compare the respective predictive performances. The method is also referred to as so-called leave-one-covariate-out (LOCO) importance.

(a) Implement LOCO.

(b) Apply LOCO to the dataset from Exercise 1 (use an unregularized OLS model again).

(c) Interpret the result (insight into model and data). Compare the result to PFI, CFI and conditional SAGE value functions.