

feature importance

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Synonyms

variable importance

Abstract

A feature importance score quantifies, for each feature, how much it contributes to the predictions of a learned hypothesis or to their quality. The permutation importance of a feature is the increase of the average loss on a dataset when the values of that feature are randomly permuted across the data points. A decision tree additionally offers an internal score by summing the improvement of its splitting criterion over the splits that use the feature, and averaged SHapley Additive exPlanations (SHAP) contributions yield another global score. Importance scores are properties of the learned hypothesis and the scoring dataset: correlated features share their credit, and an implausibly large importance is a symptom of data leakage.

Definition

A random forest serves as a classifier that predicts whether a day will be wet from a weather station’s morning measurements. Whether its predictions deserve trust depends on which of the measurements they rest on — and on whether a column that should carry no information receives a large share of the credit.

A feature importance score quantifies, for each feature, how much that feature contributes to the predictions of a learned hypothesis $\hat{h}$, or to their quality. The permutation importance of the j -th feature is the increase of the average loss on a dataset when the values of that feature are randomly permuted across the data points, which breaks their association with the label while leaving the feature’s marginal distribution intact (Breiman, 2001; Fisher et al., 2019). A feature whose permutation leaves the average loss unchanged played no measurable role for $\hat{h}$ on that dataset (Fig. 1). Permutation importance treats $\hat{h}$ as a black box; a decision tree additionally offers an internal score, obtained by summing the improvement of its splitting criterion over all splits that use the feature (Hastie et al., 2009). SHapley Additive exPlanations (SHAP) assigns each feature a contribution to one single prediction (Lundberg and Lee, 2017); averaging the magnitudes of these contributions over a dataset yields another global importance score.

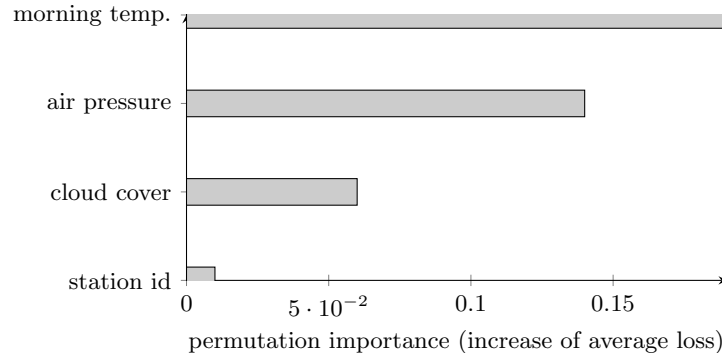


Figure 1: Permutation importances of four features for a hypothetical rain classifier (schematic values). Permuting the morning temperature or the air pressure raises the average loss noticeably, so the predictions rest on these two features; the station id contributes nothing, as it should

Importance scores are properties of the learned $\hat{h}$ and the dataset used for scoring, not of the world: two strongly correlated features share their credit, so permuting either one alone understates how much the pair carries (Fisher et al., 2019). The scores also serve as an audit: a feature that could not plausibly matter but receives a large importance points to data leakage, and a feature whose removal leaves the importance ranking and the average loss unchanged was not needed. As a global summary of which features matter, feature importance is one of the tools of explainability for otherwise opaque methods such as a random forest.

Cross References

feature, random forest, decision tree, SHAP, explainability, data leakage, loss

References

1. Breiman (2001). Random Forests. *Machine Learning*, 45(1), pp. 5–32.
2. Fisher et al. (2019). All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177), pp. 1–81.
3. Lundberg and Lee (2017). A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems* 30, pp. 4765–4774.
4. Hastie et al. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Springer Science+Business Media, New York, NY, USA.