Investigating the Contribution of Privileged Information in Knowledge Transfer LUPI by Explainable Machine Learning

Niharika Gauraha and Henrik Boström

COPA 2023

Outline

- Background:
  - Learning Under Privileged Information
  - Knowledge Transfer LUPI
  - SHAP
- Model Explainability using SHAP Values
- Split Knowledge Transfer LUPI
- Model Explainability of KT-LUPI using SHAP Values
- Experiments
- Conclusion and Future Work

Learning Under Privileged Information

- The classical machine learning paradigm assumes, training examples in the form of iid pair:

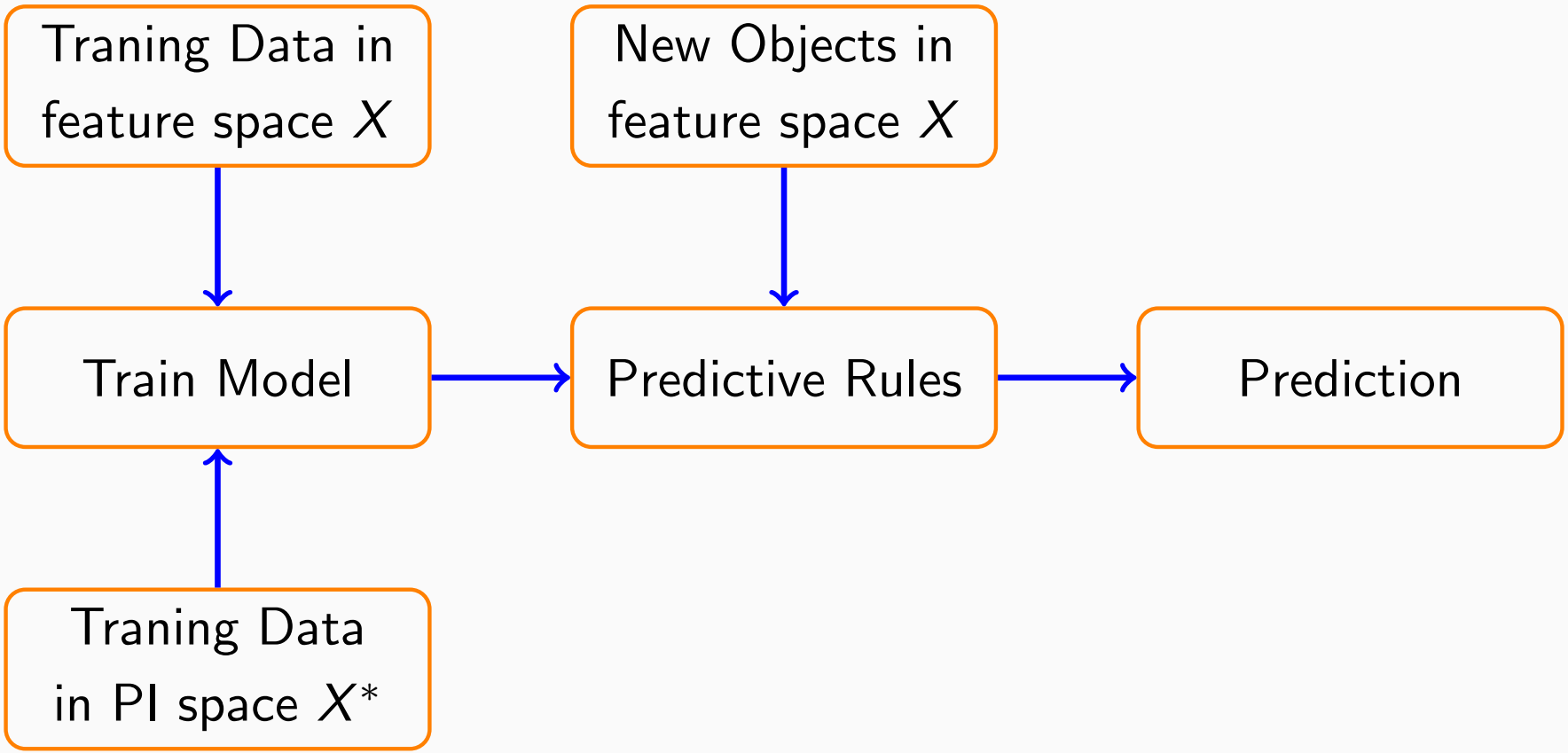
$$(x_1, y_1), ..., (x_l, y_l), \quad x_i \in X, \quad y_i \in \{-1, +1\}$$

- LUPI paradigm assumes, training examples in the form of iid triplets:

$$(x_1, x_1^*, y_1), ..., (x_l, x_l^*, y_l), \quad x_i \in X, \quad x_i^* \in X^*, \quad y_i \in \{-1, +1\}$$

- $x_i^* \in X^*$ is Privileged Information (PI) or additional data available for each training example, but not available for testing examples.

LUPI contd...


Traning Data in feature space X
New Objects in feature space X
Train Model
Predictive Rules
Prediction
Traning Data in PI space X^*

Knowledge Transfer LUPI

Given a set of IID triplets $\{x_i, x_i^*, y_i\}_{i=1}^{l}$, where $x_i \in \mathcal{X} = \mathcal{R}^p$, $x_i^* \in \mathcal{X}^* = \mathcal{R}^m$, $y_i \in \{-1, +1\}$, generated according to a fixed but unknown probability measure $P(x, x^*, y)$.

For knowledge transfer from space $\mathcal{X}^* -> \mathcal{X}$, for each privileged feature a regression function is learned by using p decision features in $\mathcal{X}$ as explanatory variables and privileged feature vectors $\mathcal{X}^{*(i)}$, for $i = 1, \ldots, m$, as response variables. In total m regression functions $\phi_j(x), j = 1, \ldots, m$ are learned.

Knowledge Transfer LUPI

We can augment the training data set as follows:

$$\begin{pmatrix} y_1 & x_1 & \phi_1(x_1) & \dots & \phi_m(x_1) \\ \dots & \dots & \dots & \dots & \dots \\ y_l & x_l & \phi_1(x_l) & \dots & \phi_m(x_l) \end{pmatrix}$$

Then, we apply some standard algorithm to this modified training data and construct an $p + m$-dimensional decision rule $f(x, \Phi(x))$, where

$$\Phi(x) = (\phi_1(x)\ \phi_2(x)\ \dots\ \phi_m(x))$$

Knowledge Transfer LUPI: Contd...

Various regression techniques can be used for approximating privileged features. For example, multiple linear Regression and Kernelized Ridge Regression (KRR) with RBF kernel etc.

SHAP

- SHAP (SHapley Additive exPlanations) is based on a game-theoretic approach, which offers a way to fairly distribute the payoffs to the individual players in a coalition.
- Given a specific data object, SHAP outputs a value for each feature that represents the contribution of that feature to the final prediction.

Model Explainability using SHAP Values

- Let us denote the decision function for a machine learning method with $f(x)$. Let the player set be a single data object $x = \{x_i | 1 \leq i \leq p\}$. Then the payoff of a coalition $s \subseteq x$ is the scalar value prediction $f(s)$ calculated from the subset of feature values.
- Since the decision function takes the input in feature space $x \in \mathcal{X}$, for computing $f(s)$, the missing input feature values are imputed with reference values, e.g., the mean computed from multiple instances.

Model Explainability of KT-LUPI using SHAP Values

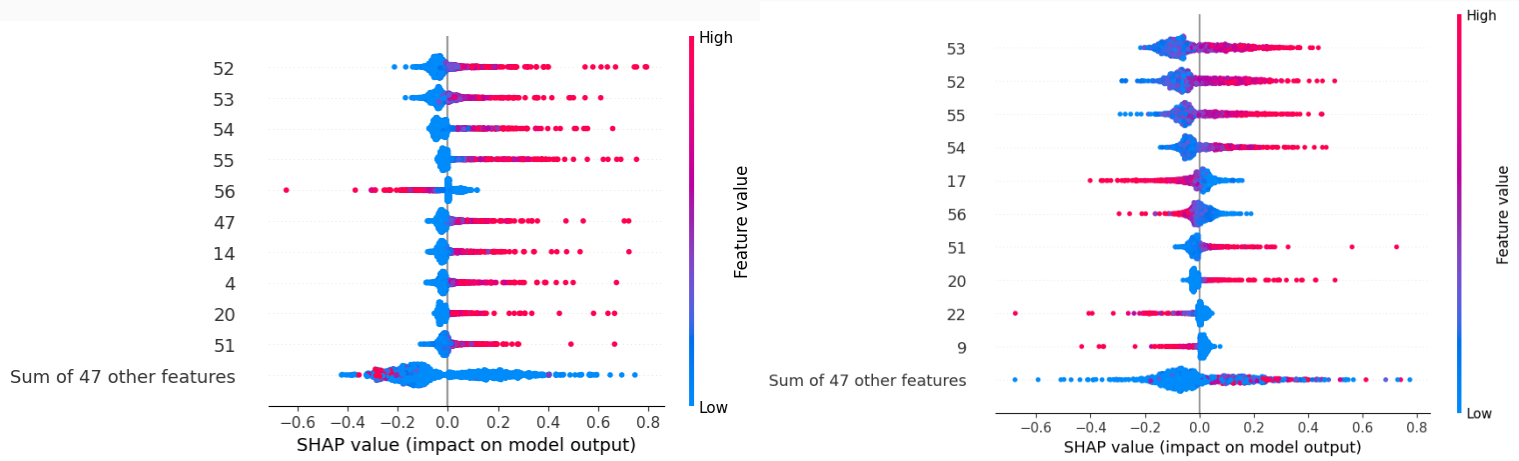
- In general, for LUPI methods, the input space for the decision rule is limited to the standard feature space, and it is not possible to compute SHAP values for the Privileged Features (PFs).
- However, for KT-LUPI, the input for the decision function is both standard features and transformed PFs. Hence, it is possible to compute the importance of PFs (though in the transformed space).

Experiment 1: Model performance and explainability for classification datasets

Table 1: Experiment 1: The first column presents the name of the datasets. The AUC scores for SVM on standard features, KT-LUPI and SVM on all features are reported in the three following columns. The last column shows the sum of absolute difference between the SHAP values for the privileged features, when computed using KT-LUPI and SVM on all features, repectively.

Dataset	SVM on standard features	KT-LUPI	SVM on all features	SAD
Spambase	0.869	0.871	0.899	0.137
Breast Cancer	0.975	0.975	0.979	0.027
Phishing Websites	0.727	0.727	0.926	0.374
Australian	0.754	0.782	0.884	0.374
Parkinsons	0.757	0.769	0.808	0.099

Experiment 1: Contd...



(a) **(b)**

Figure 1: Results from experiment 1 for the Spambase dataset; (a) SVM on all features $(X + X^*)$ (b) KT-LUPI using SVM with X as standard features and X^* as PI

Experiment 2: Model performance and explainability for regression datasets

Table 2: Experiment 2: The first column presents the name of the datasets. The RMSE scores for SVR on standard features, KT-LUPI and SVR on all features are reported in the three following columns. The last column shows the sum of absolute difference between the privileged feature SHAP values computed by KT-LUPI and SVR on all features.

Dataset	SVR on standard features	KT-LUPI	SVR on all features	SAD
Wine	1.011	1.010	0.478	1.576
Energy Efficiency	4.651	4.236	2.937	1.296
Concrete	14.323	11.918	11.594	2.252

Experiment 3: Varying numbers of privileged features

Table 3: Experiment 3: The first column presents the number of privileged features selected. The AUC scores for SVM on standard features, KT-LUPI and SVM on all features are reported in the three following columns. The last column shows the sum of absolute difference between the PFs SHAP values computed by KT-LUPI and SVM on all features.

# PFs	SVM on standard features	KT-LUPI	SVM on all features	SAD
5	0.869	0.871	0.899	0.13868
10	0.834	0.846	0.899	0.29582
15	0.781	0.794	0.899	0.37658
20	0.762	0.784	0.899	0.52306

Experiment 4: Varying the sample size

Table 4: Experiment 4: The first column presents the number of sample size. The AUC scores for SVM on standard features, KT-LUPI and SVM on all features are reported in the following three columns. The last column shows the sum of absolute difference between the PF SHAP values computed by KT-LUPI and SVM on all features.

Sample size	SVM on standard features	KT-LUPI	SVM on all features	SAD
100	0.833	0.766	0.866	0.198
200	0.802	0.802	0.867	0.105
500	0.780	0.784	0.855	0.086
1000	0.850	0.864	0.866	0.072
2000	0.908	0.921	0.933	0.064

Conclusion

- As reported previously, KT-LUPI was observed to have higher predictive performance than SVM (or SVR) using standard features only. However, in terms of explainability, KT-LUPI was comparable with SVM (or SVR) using all features.
- As the number of PFs increased, the predictive performance of both KT-LUPI and SVM using standard features were observed to decrease, but KT-LUPI still outperformed the latter.
- It was also noticed that with an increase of the number of PFs, the SAD score increased.
- As expected, when increasing the sample size, the SAD score was observed to decrease.

Future Direction

- One direction for future research is to analyze the contribution of privileged features for split KT-LUPI.
- Other directions for future research include investigating scenarios in which there is a cost associated with obtaining privileged information and explore strategies to exploit the outcome from analyzing the contribution of such information.

Thank you