Demonstrator on Counterfactual Explanations for Differentially Private Support Vector Machines*

Rami Mochaourab[1] ✉, Sugandh Sinha[1], Stanley Greenstein[2], and Panagiotis Papapetrou[3]

[1] Digital Systems Division, RISE Research Institutes of Sweden, Stockholm, Sweden
{rami.mochaourab,sugandh.sinha}@ri.se
[2] Department of Law, Stockholm University, Sweden
stanley.greenstein@juridicum.su.se
[3] Department of Computer and Systems Sciences, Stockholm University, Sweden
panagiotis@dsv.su.se

Abstract. We demonstrate the construction of robust counterfactual explanations for support vector machines (SVM), where the privacy mechanism that publicly releases the classifier guarantees differential privacy. Privacy preservation is essential when dealing with sensitive data, such as in applications within the health domain. In addition, providing explanations for machine learning predictions is an important requirement within so-called high risk applications, as referred to in the EU AI Act. Thus, the innovative aspects of this work correspond to studying the interaction between three desired aspects: accuracy, privacy, and explainability. The SVM classification accuracy is affected by the privacy mechanism through the introduced perturbations in the classifier weights. Consequently, we need to consider a trade-off between accuracy and privacy. In addition, counterfactual explanations, which quantify the smallest changes to selected data instances in order to change their classification, may become not credible when we have data privacy guarantees. Hence, robustness for counterfactual explanations is needed in order to create confidence about the credibility of the explanations. Our demonstrator provides an interactive environment to show the interplay between the considered aspects of accuracy, privacy, and explainability.

Keywords: Counterfactual Explanations · Support Vector Machines · Differential Privacy.

1 Motivation

Machine learning algorithms have proven to be powerful for learning from data and making decisions with high accuracy. In particular, they are able to outperform humans on many specific tasks. However, such data-driven technologies are seldom value-neutral to the extent that they include social and ethical values.

* Demonstrator video is available under:
https://rami-mochaourab.github.io/papers/2022-ECML/demo-video.mp4

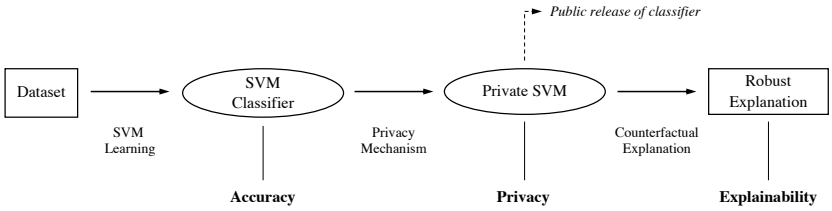

Fig. 1. The considered relationship between accuracy, privacy, and explainability.

Even when such values are integrated into the models they may be mandated by regulatory frameworks, such as traditional laws or policy documents. The goal of our work, reported in [3, 2], is to demonstrate in a technical context the link between three social and ethical values advocated by the General Data Protection Regulation (GDPR), namely, *explainability*, *privacy*, and *accuracy*.

Fig. 1 gives an overview on how the three mentioned social values are related within this work: **Accuracy** is targeted when learning an SVM classifier. **Privacy** is guaranteed using a differentially private mechanism for the classifier [4]. Afterwards, the private SVM version is made publicly available. **Explainability** for private SVM is done by designing counterfactual explanations [5] which take into account the characteristics of the classifier and privacy mechanism [3].

The innovative aspects of this work correspond to the simultaneous analysis of these three desired aspects, namely, accuracy, privacy, and explainability. The application domains of our work include those with sensitive data, such as within health, as well as within high risk applications as referred to by the EU AI Act, where explainability for data driven predictions is needed. To the best of our knowledge, there does not exist other work that studies explainability for privacy-preserving machine learning models.

The target users of our work are both machine learning researchers, working on explainable AI, as well as AI regulatory bodies interested in understanding the interplay between machine learning based decision-making, privacy guarantees, and explainability of machine learning predictions.

2 Demonstrator

Our demonstrator provides an interactive environment to understand the effects of privacy guarantees on the classification accuracy and counterfactual explanations. We use two datasets for this purpose as is shown in the snapshots from the demo in Fig. 2 and Fig. 3.

Fig. 2 shows the optimal linear SVM (solid line) and its private version (dashed line). The first sliding bar corresponds to the differential privacy parameter [1] which affects the extent of privacy guarantees. A low value means larger privacy. Consequently, larger perturbations on the classifier weights are performed when constructing the private SVM. Counterfactual explanations are

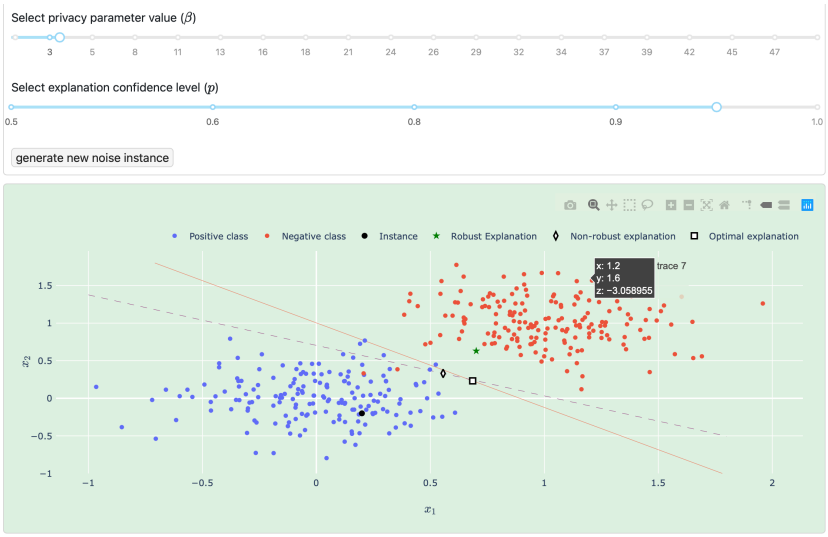

Fig. 2. Demo snapshot for explanability of linear SVM classifications on data generated from two bivariate Guassian distributions.

the closest points to the selected instance (•) that lie on the decision boundaries. Non-robust explanation (◇) may have the same class as the instance with respect to the optimal (unknown) SVM, as is shown the the screenshot. Hence, non-robust explanations are not credible and therefore we construct robust explanations (⋆) that provide confidence in explanation credibility.

The second sliding bar at the top corresponds to the confidence in the credibility of the counterfactual explanations. A large confidence means that we are more certain that the explanation has a different classification compared to that of the instance. However, a larger confidence level comes at a cost in terms of a larger distance between the explanation and the instance we want to explain. In other words, we have a tradeoff between the explanation credibility and the smallest changes needed to alter the classifier decision from the instance.

Fig. 3 demonstrates similar functionality as above but on the publicly available UCI Breast Cancer Wisconsin (Diagnostic) dataset. Here, we use a feature mapping generated using a Radial Basis Function (RBF) kernel approximation (see details in [3]). Due to the high number of features, the demo allows to visualize in two dimensions by selecting pairs of features through a drop-down menu. In order to identify the classifier errors, we mark the false positives and false negatives for both optimal and private SVM. In this way, we can see the extent of errors for different privacy parameter values. In addition, at the top-right corner we show the classification of both the selected instance and the

Fig. 3. Demo snapshot for explainability of kernel SVM classifications on the UCI Breast Cancer Wisconsin (Diagnostic) dataset.

explanation. This, highlights the diverse miss-classification possibilities inherent in the machine learning models.

The calculation of robust counterfactual explanations for kernel SVM is based on the bisection method aided by prototypes, as is detailed in [3]. A prototype for a specific data class is a typical case for that class known by the domain expert. By increasing the explanation confidence level, we can visualize how the explanations move towards the prototype at the center of the desired data class.

Acknowledgements The authors would like to thank Luis Quintero and Zhendong Wang for their help in developing the demonstrator. This work has been supported by the Digital Futures center (https://www.digitalfutures.kth.se) within the project "EXTREMUM: Explainable and Ethical Machine Learning for Knowledge Discovery from Medical Data Sources".

References

1. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (Aug 2014)
2. Greenstein, S., Papapetrou, P., Mochaourab, R.: Embedding human values into artificial intelligence. De lege 2021: Law, AI and Digitalisation pp. 91 – 115 (2022)
3. Mochaourab, R., Sinha, S., Greenstein, S., Papapetrou, P.: Robust counterfactual explanations for privacy-preserving SVMs. In: International Conference on Machine Learning (ICML 2021), Workshop on Socially Responsible Machine Learning (2021)

4. Rubinstein, B.I.P., Bartlett, P.L., Huang, L., Taft, N.: Learning in a large function space: Privacy-preserving mechanisms for SVM learning. Journal of Privacy and Confidentiality **4**(1) (Jul 2012)
5. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, forthcoming **31**(2) (2018)