sensitive attribute

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Synonyms

protected attribute, protected characteristic

Abstract

A sensitive attribute of a data point is an attribute, such as gender or ethnic origin, that must not influence the prediction a machine learning system (ML system) delivers for that data point, whether or not it belongs to the feature vector. The word carries two senses that are kept apart: a regulatory one, in which data protection law fixes which categories of personal data are special, and a machine learning (ML) one, in which the application names the attribute s that a fairness criterion compares predictions across. Dropping s from the feature vector does not remove it from the data, because the remaining features can be dependent on it. Learned features, such as those principal component analysis (PCA) delivers, are no exception: the criterion that learns them never mentions s, and they can be predictive of it all the same. How much of s they carry is measured by the recovery accuracy, the accuracy with which a linear classifier predicts s from them on data points it was not learned from. For the weather at Krems an der Donau (Austria), two learned features recover the half of the year a day belongs to with a recovery accuracy of 0.94, against a majority-class rate of 0.51. Projection of the learned features onto the orthogonal complement of the direction the linear classifier uses lowers that accuracy without bringing it to the majority-class rate, so a sensitive attribute is easier to measure than to remove.

Definition

Consider a bank that decides on loan applications from the income, the occupation, and the address of an applicant. A supervisor asks whether the decisions depend on the gender of the applicant. The bank answers that the gender was never recorded as a feature: the feature vector $\mathbf{x}$ that the learned hypothesis $\widehat{h}$ reads holds those three numbers and nothing else. That answer settles nothing. The occupation and the address of an applicant can be predictive of the gender, and a prediction computed from $\mathbf{x}$ then depends on the gender all the same.

A sensitive attribute of a data point is an attribute that must not influence the prediction delivered for that data point, whether or not it belongs to the

feature vector. A data point then carries three parts: the feature vector $\mathbf{x} \in \mathcal{X}$, the label y, and the sensitive attribute s, which typically takes few values. The gender of a loan applicant is written $s \in \{0, 1\}$ below.

Which attribute counts as sensitive is settled in two different ways, and the two senses are kept apart. In the regulatory sense, the answer is given by law. The general data protection regulation (GDPR) prohibits, with exceptions, the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and of genetic data, biometric data used to identify a person, health data, and data concerning sex life or sexual orientation (European Parliament and Council of the European Union, 2016, Art. 9(1)). The EU AI Act adopts that list under the name special categories of personal data (European Parliament and Council of the European Union, 2024, Art. 3(37)). In the machine learning (ML) sense, the application decides, and its choice need not belong to that list: the gender of the loan applicant is a sensitive attribute of the example above and is not a special category under the GDPR. A fairness criterion compares the predictions of $\widehat{h}$ across the values of s (Barocas et al., 2023), while privacy protection asks instead that the output of the machine learning system (ML system) not allow inference of s, an inference measured by privacy leakage.

The obvious mitigation is the one the bank applied: leave s out of the feature vector. It removes s from the feature vector and not from the data. The data point is modeled as a realization of random variables (RVs), with the feature vector and the sensitive attribute drawn jointly. The attribute is still present whenever the two are dependent,

$$\mathbb{P}\left(s = 1 \mid \mathbf{x} \in \mathcal{A}\right) \neq \mathbb{P}\left(s = 1\right) \text{ for some } \mathcal{A} \subseteq \mathcal{X}. \tag{1}$$

A feature for which the dependence (1) holds is a proxy for s. Such redundant encodings are what lets a decision rule that ignores s discriminate all the same (Dwork et al., 2012).

Feature learning does not undo the dependence (1). It replaces the raw features by learned ones, $\mathbf{z} = \mathbf{\Phi}(\mathbf{x}) \in \mathbb{R}^{d'}$, delivered by a feature transformation $\mathbf{\Phi}$ that is itself learned from a dataset. Principal component analysis (PCA) learns a linear $\mathbf{\Phi}$ whose rows are the d' eigenvectors of the sample covariance matrix with the largest eigenvalues, a criterion in which s never appears. The learned features can be predictive of s nonetheless, and how much of it they carry is a number that can be measured. A linear classifier $\widehat{h}^{(s)}$ that predicts s from $\mathbf{z}$ is learned by least squares on a training set $\mathcal{D}^{(\text{train})}$, and its recovery accuracy is its accuracy on a test set $\mathcal{D}^{(\text{test})}$ of data points outside $\mathcal{D}^{(\text{train})}$,

$$\operatorname{acc}\left(\widehat{h}^{(s)} \mid \mathcal{D}^{(\text{test})}\right) := 1 - \frac{1}{\left|\mathcal{D}^{(\text{test})}\right|} \sum_{(\mathbf{z},s) \in \mathcal{D}^{(\text{test})}} L^{(0/1)}\left((\mathbf{z}, s), \widehat{h}^{(s)}\right), \tag{2}$$

with the 0/1 loss, which is 1 for a wrong prediction and 0 otherwise. The recovery accuracy (2) is read against the majority-class rate, the accuracy of predicting the more frequent value of s for every data point.

The features that PCA learns from the weather at Krems an der Donau (Austria) show what this measurement reports. Each day from January 6 to December 31 of 2024 is a data point, and its raw features are the eight daily measurements of each of the five previous days, forty numbers. Its attribute s says whether the day lies in the cold half of the year, October through March; 178 of the 361 days do, so the majority-class rate is 0.51. The half-year is not among the forty raw features. Each raw feature is scaled to zero sample mean and unit sample variance first, because temperatures, precipitation, and pressure carry different units. Two learned features recover it with a recovery accuracy (2) of 0.94, averaged over twenty random splits of the days into two halves, $\mathcal{D}^{(\text{train})}$ and $\mathcal{D}^{(\text{test})}$, and a single learned feature already gives 0.94. The decision boundary of one such linear classifier is a line in the plane of the two learned features (Fig. 1(a)). Dropping every temperature measurement from the raw features, the obvious proxy for the season, leaves twenty-five numbers per day. The recovery accuracy is then 0.82 with two learned features and 0.89 with five. The two curves of Fig. 1(b) compare the two sets of raw features for $d' = 1, \ldots, 10$. The half-year of a day is not a sensitive attribute. It stands in for one here, and the computation does not depend on what the attribute means.

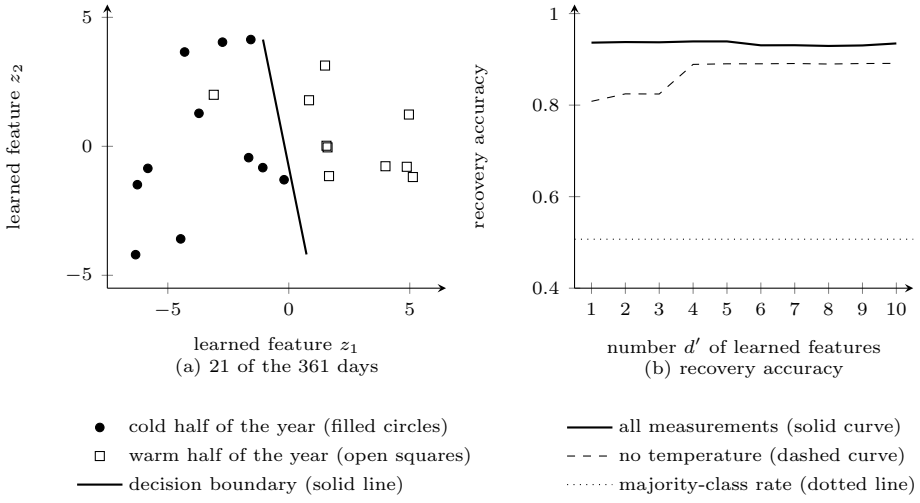


Figure 1: An attribute that is no feature, recovered from learned features of the weather at Krems an der Donau. (a) Every 18th of the 361 days of 2024 in calendar order, 21 days, placed by two learned features and marked by the half of the year they belong to, the cold one (filled circles) or the warm one (open squares), with the decision boundary (solid line) of the linear classifier learned from 180 of the 361 days. (b) The recovery accuracy (2) against the number d' of learned features, for all eight daily measurements (solid curve) and with every temperature measurement dropped (dashed curve), against the majority-class rate (dotted line). Data generated by `pythondemos/sensattr.py`

A sensitive attribute is easier to measure than to remove. The linear classifier $\widehat{h}^{(s)}$ uses a single direction of the learned feature space, the one given by its learned model parameters $\widehat{\mathbf{w}}$. The projection of the learned features onto the orthogonal complement of that direction, the nullspace of the linear map $\mathbf{z} \mapsto \widehat{\mathbf{w}}^\top \mathbf{z}$, removes it and no more. For ten learned features of the days at Krems, a linear classifier learned afresh on the projected features still reaches 0.84 on one of the twenty splits, against 0.93 before the projection, so the other directions carry the attribute too. Learning a feature transformation under a penalty term for the recovery of s attacks the same problem where the features are learned (Zemel et al., 2013). Every one of these steps needs s to be recorded, which is why the EU AI Act lets a provider of a high-risk artificial intelligence system (high-risk AI system) process special categories of personal data when that is strictly necessary for detecting and correcting bias, under safeguards (European Parliament and Council of the European Union, 2024, Art. 10(5)).

Cross References

fairness, trustworthy AI, privacy protection, privacy leakage, feature learning, PCA, linear classifier, GDPR, EU AI Act, personal data, nullspace, accuracy

References

1. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

2. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union, L 119/1, May 4. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

3. Barocas et al. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press.

4. Dwork et al. (2012). Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214–226.

5. Zemel et al. (2013). Learning fair representations. In: International Conference on Machine Learning, pp. 325–333.