Machine Learning

Trace, explain and interpret

S. Nõmm

[1]Department of Software Science, Tallinn University of Technology

03.05.2022

Motivation

- In many areas where ML methods are applied it is required to provide human readable explanation.
- Examples: medicine, finance, and low.
- Case of supervised learning is considered.
- What is human readable explanation?
- The area of tracing, explaining and interpreting classifier predictions is closly related to the adversarial machine learning.
- Sometimes the notion of Decision Interpretability is used.
- This area is relatively new, gaining popularity and there is no general agreement in using different terms.

Problems

- There is no agreement within ML community about the necessity to explain and interpret ML and AI decisions.
- There is no agreement with ML community about notions and definitions.
- Existing literature is poor with respect to formal definitions.
- There is no unified approach to evaluate explanations and interpretations.
- There is a tendency to develop techniques for particular cases.

References

The present lecture is based on the following sources: (corresponding pdf files (articles only) are shared through the ained.ttu.ee between the course participants).

- " Why should i trust you?" Explaining the predictions of any classifier by MT Ribeiro, S Singh, C Guestrin
- Comprehensible Classification Models a position paper by Alex A. Freitas at al.
- How to Explain Individual Classification Decisions by David Baehrens at al.
- A Survey of Methods for Explaining Black Box Models by RICCARDO GUIDOTTI at al.
- Interpretable Machine Learning A Guide for Making Black Box Models Explainable Christoph Molnar.
- https://towardsdatascience.com/idea-behind-lime-and-shap-b603d35d34eb
- https://gilberttanner.com/blog/local-model-interpretation-an-introduction
- A Unified Approach to Interpreting Model Predictions by Scott M. Lundberg at al.

Definitions

Consider two class classification problem. Let $D = D_{train} \cup D_{valid} \subset S = \mathbb{R}^m$ be an m dimensional data set of interest. Denote X the feature set used to train classifier C and $\hat{y} = C(x_i)$ class label estimation given by C to the point $x_i \in \Pi_X(D_{valid})$ (here Π denotes the projection of the set D_{valid} in to the space spanned by feature set X).

Definition

ε- decision boundary of the classifier C is the set of points in S such that the ε- neighborhood of each point contains at least one point of each class.

Such setting implies that decision boundary depends on the classifier.

Tracing

Definition

Prediction (class label estimation) trace (or trace of the $\hat{y}_i = C(x_i)$) is the set of at least m inequalities explaining position of x_i with respect to the decision boundary in coordinate system defined by X. Whereas each inequality may be complemented by the weigh describing its importance in making label prediction.

- Decision trees naturally produce the set of inequalities. The number of inequalities may in this case may differ from m. If the number of inequalities is smaller than m and some features are left unused then the set may be complemented by the inequalities of the form $-\infty < x_i < \infty$. The ordering of the conditions provide natural way of describing their importance for the decision making.
- SVMs may explicitly provide linear or nonlinear inequality of the decision boundary in more than one variable. in this case one may convert it into a system of parametric inequities.
- k- nearest neighbors does not produce any inequalities naturally.

Explanation

Definition

Prediction explanation is the set of less than m inequalities explaining position of x_i with respect to the decision boundary in coordinate system defined by $\Pi_{S_r}(X)$, where $\dim(\mathrm{S_r}) < \mathrm{m}$ (projection of X in to the subspace of lower dimension) . Whereas each inequality may be complemented by the weight describing its importance in making label prediction.

An explanation of the prediction may be seen as an attempt to reduce dimesionality of the prediction trace.

Interpretation

The idea of the interpretation is to explain class prediction given by the classifier C in terms of feature set which differs from X. This implies necessity of the mapping $M_I : \Pi_X(D) \to \tilde{D}$ where $\tilde{D}$ is spanned by the feature set to be used in the explanation. An easy example is the features spanning S except X. More formally, denote X_S the feature set spanning space S, then one may attempt to interpret classification results given by C in terms of features $X_S \setminus X$.

Definition

Prediction interpretation is the set of inequalities explaining position of $M_I(x_i)$ with respect to the image of the decision boundary in coordinate system defined by $M_I(X)$. Whereas each inequality may be complemented by the weigh describing its importance in making label prediction.

One may easily see that explanation is one particular case of interpretation, where mapping M_I is a projection.

Interpretation

- Tracing provides one with the explicit information describing positioning of the point with respect to the decision boundary.
- Explanation is the attempt to reduce the amount of information needed to explain positioning of the point with respect to the decision boundary.
- Does interpretation make any sense? Among these three notions interpretation (as defined here) was not extensively studied. Possible example is: Let D be a data set in the six dimensional space. Let three even features will be uncorrelated, but each of them will be tightly linearly correlated to the corresponding odd feature. For the classifier C trained only on even features features one may try to explain classification results using odd features instead.
- Applicability of such approach may be questioned but formally such procedure is correct.

Human readable?

- How to determine if given set of inequalities is human readable?
- May be explaining labels by stating the labels of nearest neighbors is easier to understand?
- This question is tightly connected to the psychology and on the present stage left out of the frameworks of the present course.
- Most probably the term *understandability* should be defied?

Let us now deviate from the questions discussed above.

Most popular techniques

- SHAP - SHapley Additive exPlanations.
- LIME - Local Interpretable Model agnostic Explanations by M.T. Ribeiro.

Few more notions

Let us suppose that the space of our interest is $\mathbb{R}^n$

- Global: the property or relation is global is it is valid everywhere in $\mathbb{R}^n$
- Generic: the property or relation is generic if it valid everywhere in $\mathbb{R}^n$ except the finite number of elements.
- Local: the property or relation is local if it is valid in a neighborhood of a point.

NB!

Observe:

- LIME: is based on the following notions without explicitly defining them: (this technique would be considered in this lecture).
  - *locally faithful*
  - *fidelity*
  - *interpretable* in the sense of being understandable, whereas qualitative explanation is targeted, at least in LIME.
- SHAP is based on so called SHAP values describing importance of each feature in the decision making.

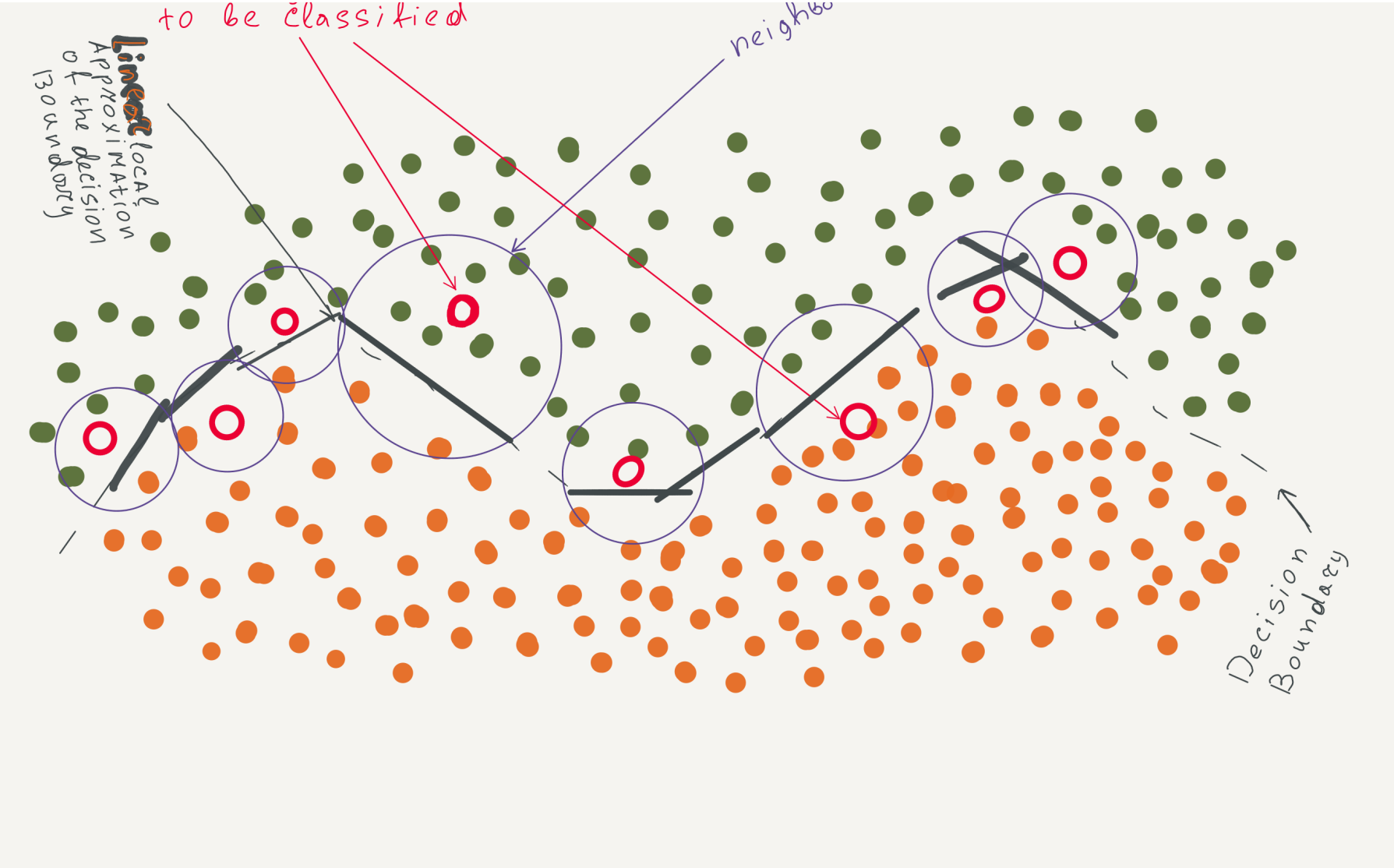
Assumptions

Let $D \in \mathbb{R}^n$ be the dataset of the interest explaining the decision made by an arbitrary classifier C relies on the following assumptions:

- There is a set of models which results are human interpretable.
- There exist sufficiently small, positive ε such that classification process is linear in the ε - neighborhood of each point in D.

Roughly speaking LIME and SHAP are trying to construct linear approximation of the C in some ε neighborhood of each point to be classified.

Approximation of an arbitrary classifier by means of the linear one



Additive feature attribution

NB! Here we use different notations just to distinguish between the ideas stated above and those used by the other authors.

- Let f be the classifier of interest and g the explanation model (another classifier we can understand).
- Denote $h_x : \mathbb{R}^n \to \mathbb{R}^n$ (analogue of projection we discussed above but works in the opposite direction!!!) $x = h_x(x')$.
- Explanation techniques are trying to ensure: $g(z') = h(f(z'))$ when $z' \approx x'$
- According to Lunberg: Additive feature attribution methods have an explanation model that is a linear function of binary variables:

$$g(z') = \phi_0 \sum_{i=1}^{M} \phi_i z_i'$$

where z' is a vector of binary elements.

Lime

- Let f be the classifier of interest and g the explanation model (another classifier we can understand).
- Denote $h_x : \mathbb{R}^n \to \mathbb{R}^n$ (analogue of projection we discussed above) $x = h_x(x')$. Denote loss function as L.
- LIME minimizes:

$$\xi = \text{argmin}_{g \in G} L(f, g, \pi_{x'}) + \Omega(g).$$

Here $\Omega(g)$ is the function to penalize complexity of g and π is the local weight kernel.

SHAP

- Shapley regression values may be treated as the feature importance values.
- Denote $S \subset F$ where F is the set of all features.
- Shapeley values for the model f are defined as follows:

$$\phi_i = \sum_{s \subset F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)].$$

Properties of the explaining model

- Local accuracy.
- Missingness.
- Consistency.

Implementation issues

- How to choose the neighborhood?
- How to define loss function?
- How to define penalizing function?
- ... .