Interpretable Machine Learning

Increasing Trust in Explanations





Learning goals

- Understand the aspects that undermine users trust in an explanation
- Learn diagnostic tools that could increase trust in an explanation
 - Learn diagnostic tools that could increase trust

 Local explanations should not only make a model interpretable but also reveal if eal if the model is the model is trustworthy



- Local explanations should not only make a model interpretable but also reveal if eal if the model is the model is trustworthy
- Interpretable != Why! did the model come up with this idecision? "on?"



- Local explanations should not only make a model interpretable but also reveal if eal if the model is the model is trustworthy
- Interpretable! "Why! did the model come up with this decision?" on?"
- Trustworthy: "How certain is this explanation?" n?"
 - accurate in sights into the inner workings of our modeledel
 - Failure cases generation is based on inputs in areas where the model odel was trained was trained with little or not training datal (extrapolation)



- Local explanations should not only make a model interpretable but also reveal if eal if the model is the model is trustworthy
- Interpretable! "Why! did the model come up with this decision?"
- Trustworthy: "How certain is this explanation?" n?"
 - accurate insights into the inner workings of our modeledel
 - Failure cases generation is based on inputs in areas where the model odel was trained was trained with little on not training data (extrapolation)
 - @robusts(i.e.dow.variance):e)
 - Expectation: similar explanations for similar data points with similar nilar predictions
 predictions multiple sources of uncertainty exist
 - However; multiple: sources of uncertainty exist mall changes in the input data or
 - --- measure how robust an IML method is to small changes in the input
 - ~ data or parameters out-of-distribution?
 - → Is an observation out-of-distribution?



- Local explanations should not only make a model interpretable but also reveal if eal if the model is the model is trustworthy
- Interpretable eWhyl did the model come up with this decision? on?
- Trustworthyn,"How certain is this explanation?"
 - accurate in sights into the inner workings of our modeledel
 - Failure cases generation is based on inputs in areas where the model odel was trained was trained with little on not training data (extrapolation)
 - @ robusts(i.e.dow variance):e)
 - Expectation: similar lex plantations for similar data points with similar nilar predictions
 predictions multiple sources of uncertainty exist
 - However, multiple sources of uncertainty exist mall changes in the input data or
 - we measure how robust an IML method is to small changes in the input
 - → data:or parameters out-of-distribution?
 - Failing in Is an observation out of distribution? trust in the explanations
- Failing inconerof these windermining users' trust in the explanations
 - wundermining trust in the model

- Models are unreliable in areas with little data support
 - Models are unreliable areas with little data supported and explanation methods are unreliable are explanations from local explanation methods are unreliable.



- Models are unreliable in areas with little data support
- → explanations from local explanation methods are unreliable
- explanations from local explanation methods are unreliable
 For local explanation methods, the following components could be
 out-of-distribution (OOD):

 Out-of-distribution (OOD):
 - The data for LIME's surrogate model
 - Counterfactuals themselves
 - "Shapley value is permuted observations to calculate the marginal contributions contributions grid data points
 - ICE curves grid data points



- Models are unreliable in areas with little data support
- explanations from local explanation methods are unreliable

 For local explanation methods, the following components could be

 out-of-distribution (OOD):

 explanations from local explanation methods are unreliable

 out-of-distribution (OOD):
 - The data for LIME's surrogate model
 - Counterfactuals themselves
 - "Shapley value's permuted observations to calculate the marginal contributions
 - contributions grid data points
 - Twoice curves grid data points approaches
- Two very simple and intuitive approaches
 - Classifier for out-of-distribution
 - Morciusteringated also possible, e.g., variational autoencoders [Daxberger et al. 2020]
- More complicated also possible, e.g., variational autoencoders [Daxberger et al. 2020]



OUT-OF-DISTRIBUTION DETECTION: : OOD-CLASSIFIER OOD-CLASSIFIER

- Problem: we have only in-distribution data
- Problem: ave have only in-distribution data by randomly sample data points
- Idea: Hallucinate new (out-of-distribution) data by randomly sample data points
- Learn a binary classifier to distinguish between the origins of the data

OUT-OF-DISTRIBUTION DETECTION: OOD-CLASSIFIER OOD-CLASSIFIER

- Problem: we have only in-distribution data
- Problem: we have only in-distribution idata n) data by randomly sample data points
- Idea: Hallucinate new (out-of-distribution) data by randomly sample data points
- Learn a binary classifier to distinguish between the origins of the data

 Study whether an explanation approach can be fooled
- Study whether an explanation capproach can be fooled and unitstated model for all
 - Hide bias in the true (deployed) model, but use an unbiased model for all
- Impout-of-distribution samples, explanation approach
- Important way to diagnose an explanation approach

OUT-OF-DISTRIBUTION DETECTION: CLUSTERING VIA DBSCAN VIA DBSCAN data clustering algorithm • Martin Ester et al. 1995

DBSCAN/is a data clustering algorithm (Applications with Noise)
 (Density-Based Spatial Clustering of Applications with Noise)



VIA DBSCAN data clustering algorithm • Martin Ester et al. 1996

- DBSCAN/is a data clustering algorithm (Applications with Noise)
- (Density-Based Spatial Clustering of Applications with Noise)
- For this method, we'define an ℓ -neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as Given a dataset $X = \{\mathbf{x}^{(l)}\}_{l=1}^n$, an ℓ -neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as $\mathcal{N}_{\ell}(\mathbf{x}) = \{\mathbf{x}^{(l)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(l)}) \leq \epsilon\}$.

$$\mathcal{N}_{\epsilon}(\mathbf{x}) = \{\mathbf{x}^{(l)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(l)}) \le \epsilon\}.$$
 $d(\cdot)$ is a distance measure (e.g., Euclidean or Gower distance)

 $d(\cdot)$ is a distance measure (e.g., Euclidean or Gower distance)



VIA DBSCAN a data clustering algorithm • Martin Ester et al. 1996

- DBSCAN/is a data clustering algorithm (Applications with Noise)
- (Density-Based Spatial Clustering of Applications with Noise)
- Forthis method, we'define an ϵ -neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as Given a dataset $X = \{\mathbf{x}^{(i)}\}_{i=1}^n$, an ϵ -neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as $\mathcal{N}_{\epsilon}(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(i)}) \leq \epsilon\}$.

$$\mathcal{N}_{\epsilon}(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(i)}) \le \epsilon\}.$$
 $d(\cdot)$ is a distance measure (e.g., Euclidean or Gower distance)

- od(C) is a distance measure (e.g., Euclidean or Gower distance)
- Core observations \mathbf{x} m data points within $\mathcal{N}_{\epsilon}(\mathbf{x})$
 - "Have at least who data points within syling horhood points
 - · Forms an own cluster with all its neighborhood points



VIA DESCAN data clustering algorithm Marsin Ester et al. 1998

- DBSCAN/is a data clustering algorithm (Applications with Noise)
- (Density-Based Spatial Clustering of Applications with Noise)
- Forthis method, we'define an ϵ -neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as Given a dataset $X = \{\mathbf{x}^{(i)}\}_{i=1}^n$, an ϵ -neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as $\mathcal{N}_{\epsilon}(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(i)}) \leq \epsilon\}$.

$$\mathcal{N}_{\epsilon}(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(i)}) \le \epsilon\}.$$
 $d(\cdot)$ is a distance measure (e.g., Euclidean or Gower distance)

- od() is a distance measure (e.g., Euclidean or Gower distance)
- Core observations \mathbf{x} m data points within $\mathcal{N}_{\epsilon}(\mathbf{x})$
 - "Have at least who data points within N. (1) horhood points
 - Ber Forms alsown cluster with all its neighborhood points
- Border boints N_ϵ(x)
 - Within Wa(x) uster defined by a core point
 - · Part of a cluster defined by a core point



VIA DESCAN a data clustering algorithm • Martin Ester et al. 1996

- DBSCAN/is a data glustering algorithm (Applications with Noise)
 - (Density-Based Spatial Clustering of Applications with Noise)
- For this method, we'define an ℓ -neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as Given a dataset $X = \{\mathbf{x}^{(i)}\}_{i=1}^n$, an ℓ -neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as $\mathcal{N}_{\epsilon}(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(i)}) \leq \epsilon\}$.

$$\mathcal{N}_{\epsilon}(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(i)}) \le \epsilon\}.$$
 $d(\cdot)$ is a distance measure (e.g., Euclidean or Gower distance)

- d() is a distance measure (e.g., Euclidean or Gower distance)
- ullet Core observations ${f x}$ m data points within ${\cal N}_{\epsilon}({f x})$
 - "Have at least who data points within N. Tigh borhood points
 - Ber Forms are own cluster with all its neighborhood points
- Border boints Λ_ε(x)
 - Within No (x) uster defined by a core point
- NoiPart of a cluster defined by a core point
- Noise points of within $\mathcal{N}_{\epsilon}(\mathbf{x})$
 - Are not within N. (x) ister
 - Not part of any cluster





Example for DBS CAN, circles display ϵ -neighborhoods, m=4

Green points A and B are core points and and form one cluster since they lie in hey lie in each others neighborhood, all yellow points are border points of this cluster.

Green points A and B are core points and and and are core points and are core points and are core points and and and are core points and and are core points and are core points and are core points and and are core points and are core points and are core points and and are core points a



Example for DBS CAN, circles display ϵ -neighborhoods, m=4

- Green points A and B are core points and and form one cluster since they lie in hey lie in each others neighborhood, all yellow points are border points of this cluster.

 Cluster
- Since D is not part of the neighborhood of Since D is not part of the Core points, it is a noise point neighborhood of core points, it is a noise point



- Green points A and B are core points and and form one cluster since they lie in hey lie in each others neighborhood, all yellow points are border points of this cluster.

 Cluster
- Since D is not part of the neighborhood of
 Since D is not part of the core points, it is a noise point neighborhood of core points, it is a noise point In-distribution: new point lies within a cluster
- In-distribution: new point lies within a cluster



Green points A and B are core points and and form one cluster since they lie in hey lie in each others neighborhood, all yellow points are border points of this cluster.

Cluster

- Since D is not part of the neighborhood of
 Since D is not part of the core points, it is a noise point neighborhood of core points, it is a noise point in-distribution: new point lies within a cluster
- In-distribution: new point lies within a point lies outside the cluster
- Out-of-distribution: new point lies outside the clusters



Example for DBS CAN, circles display ϵ -neighborhoods, m=4

- Disadvantages:
- Disadvantages nality"
 - Depending on the distance metric d(-). DBSCAN could suffer from the "curse of dimensionality"
 - The choice of ε and m is not clear a-priori

- Green points A and B are core points and and form one cluster since they lie in each other each others neighborhood, all yellow points are points are border points of this cluster
- Since D is not part of the neighborhood of
 Since D is not part of the. neighborhood of core points, it is a noise point distribution: new point lies within a cluster noise point
- In-distribution: new point lies within a point lies outside the clusters cluster
- Out-of-distribution: new point lies • Depending on the distance metric $a(\cdot)$, DBSCAN could suffer from the "curse of

ROBUSTNESS

- Differentiate between different kinds of uncertainty inty:
 - Explanation uncertainty: Change of explanation if we repeat the the process, e.g., the process, e.g., the explanation could differ depending on which subset of for the explanation data we use for the explanation method and which hyperparameters

ROBUSTNESS

- Differentiate between different kinds of uncertainty: inty:
 - Explanation uncertainty: Change of explanation if we repeat the the process, e.g., the process, e.g., the explanation could differ depending on which subset of for the explanation data we use for the explanation method and which hyperparameters
 - Process uncertainty: Change of explanation if the underlying model is let is changed changed ML models non-robust, e.g., because they are trained on noisy data?
 → are ML models non-robust, e.g., because they are trained on noisy data?

ROBUSTNESS

- Differentiate between different kinds of uncertainty: inty:
 - Explanation uncertainty: Change of explanation if we repeat the the process, e.g., the process, e.g., the explanation could differ depending on which subset of for the explanation data we use for the explanation method and which hyperparameters
 - Process uncertainty: Change of explanation if the underlying model is let is changed changed ML models non-robust, e.g., because they are trained on noisy data?
 - We pous on explanation uncertainty are trained on noisy data?
- Even with the same model and same (or similar) data points, we can receive different
 We focus on explanation uncertainty
 - Even with the same model and same (or similar) data points, we can receive different explanations

Objective: Similar explanations for similar inputs (in a neighborhood) od)



- Objective: Similar explanations for similar inputs (that neighborhood)
- For LIME and SHAP, notion of stability based on locally Lipschitz continuity wity
 Avanza Malis and Jaakkota 2018

An explanation method of :9x 4 Rm is locally Lipschitz fritz if

- for every x₀ ∈ X the relexist by 0 and we R∈ R
- \bullet such that $\|\mathbf{x}\| \times \mathbf{x}_0 \| \le \delta$ implies $\|\hat{g}(\mathbf{x})\| \times g(\mathbf{x}_0)\| < \omega \| \mathbf{x} \mathbf{x}_0 \| = \mathbf{x}_0 \|$

Note that, for LIME, g returns the m coefficients of the surrogate model



- Objective: Similar explanations for similar inputs (the ineighborhood)
- For LIME and SHAP hotion of stability based on locally Lipschitz continuity wity

 [Alvarez-Melis and Jackwara 2018]



- for every $\mathbf{x}_0 \in \mathcal{X}$ there exist $\mathbb{R} \searrow 0$ and $\mathbb{R} \subseteq \mathbb{R} \subseteq \mathbb{R}$
- such that $\|\mathbf{x} \times \mathbf{x}_0\| < \delta$ implies $\|\hat{g}(\mathbf{x}) \times g(\mathbf{x}_0)\| < \delta$

Note that, for LIME, g returns the m coefficients of the surrogate model

- According to this, we can quantify the robustness of explanation models he is in terms of ω : terms of ω : terms of ω : to 0, the more robust our explanation method is
 - \rightarrow The closer ω is to 0, the more robust our explanation method is



- Objective Similar explanations for similar inputs (that neighborhood) od)
- For LIME and SHAP, notion of stability based on locally Lipschitz continuity wity Alvarez-Melis and Jaakkola 2018



- for every x < X there exist > 0 and D < R ∈ R
- ◆ such that | x × x | fo < δ implies | fg(x) (× g(x o) (× c) | x = x | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o | | x o

Note that, for LIME, g returns the m coefficients of the surrogate model

- ullet According to this, we can quantify the robustness of expranation models riels in terms of ω : terms of ω closer ω is to 0, the more robust our explanation method is
- wisThe:dosenwis:to:0;the more robust:our explanation method is
- ω is rarely known a-priori but it could be estimated as follows: $\hat{\omega}_X(\mathbf{x}) \in \underset{\mathbf{x}^{(i)} \in \mathcal{N}_{\epsilon}(\mathbf{x})}{\text{arg max}} \frac{\|g(\mathbf{x})_{\mathcal{N}_{\epsilon}} \|g(\mathbf{x}^{(i)})\|_{2}}{d(\mathbf{x}, \mathbf{x}^{(i)})},$ where $\mathcal{N}_{\epsilon}(\mathbf{x})$ is the ϵ -neighborhood of \mathbf{x}

where $\mathcal{N}_{\epsilon}(\mathbf{x})$ is the ϵ -neighborhood of \mathbf{x}

