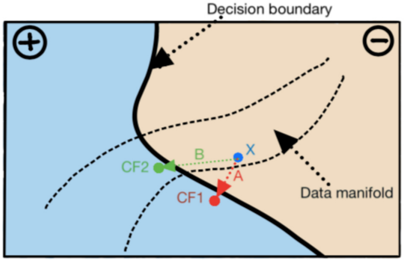
Interpretable Machine Learning

Counterfactual Explanations

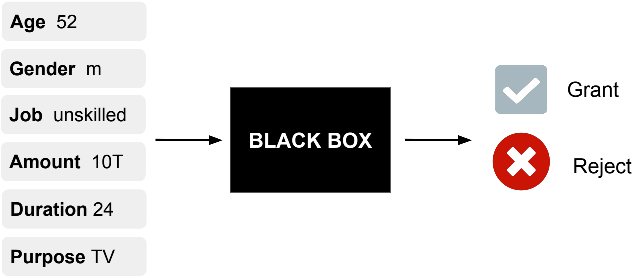


Learning goals

- Understand the motivation behind CEs
- See the mathematical foundation of CEs

EXAMPLE: CREDIT RISK APPLICATION

- **x**: customer and credit information
- y: grant or reject credit

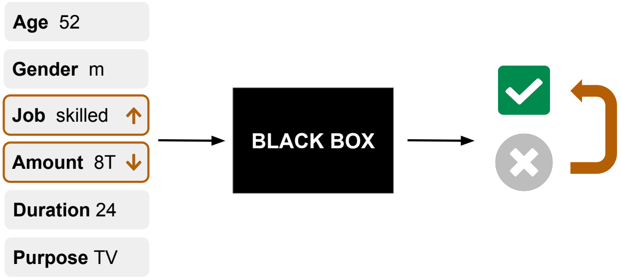


Questions:

- Why was the credit rejected?
- Is it a fair decision?
- **How should x be changed so that the credit is accepted?**

EXAMPLE: CREDIT RISK APPLICATION

Counterfactual Explanations provide answers in the form of "What-If"-scenarios.



“If the person was more skilled and the credit amount had been reduced to $8.000, the credit would have been granted."

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA

- Counterfactual explanations == counterfactuals == CEs

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA

- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA

- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome
- Represent **close neighbors** of a data point we are interested in, but belonging to the **desired outcome**

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA

- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome
- Represent **close neighbors** of a data point we are interested in,
  but belonging to the **desired outcome**
- Reveal which minimal changes to the input are sufficient to receive a different outcome
  $\rightsquigarrow$ Useful if there is a chance to change the input features (e.g., by changing behaviour)

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA

- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome
- Represent **close neighbors** of a data point we are interested in, but belonging to the **desired outcome**
- Reveal which minimal changes to the input are sufficient to receive a different outcome
  $\leadsto$ Useful if there is a chance to change the input features (e.g., by changing behaviour)
- The targeted audience of CEs are often end-users

AIMS & ROLES

CEs can serve various purposes, the user can decide what to learn from them. For example:

“If the person had been **one year older** and the **credit amount had been increased** to \$12.000, the credit would have been granted."

AIMS & ROLES

CEs can serve various purposes, the user can decide what to learn from them. For example:

“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

AIMS & ROLES

CEs can serve various purposes, the user can decide what to learn from them. For example:

“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

- **Provide reasons:**
  Interesting, I did not know that age plays a role in loan applications.

AIMS & ROLES

CEs can serve various purposes, the user can decide what to learn from them. For example:

“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

- **Provide reasons:**
  Interesting, I did not know that age plays a role in loan applications.

- **Provide grounds to contest the decision:**
  How dare you, I do not want to be discriminated for my age in an application.

AIMS & ROLES

CEs can serve various purposes, the user can decide what to learn from them. For example:

“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

- **Provide reasons:**
  Interesting, I did not know that age plays a role in loan applications.

- **Provide grounds to contest the decision:**
  How dare you, I do not want to be discriminated for my age in an application.

- **Detect model biases:**
  There is a bug, an increase in amount should not increase approval rates.

PHILOSOPHICAL BASIS

Counterfactuals have a long-standing tradition in analytic philosophy
$\rightsquigarrow$ Accoding to Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If S was the case, Q would have been the case."

PHILOSOPHICAL BASIS

Counterfactuals have a long-standing tradition in analytic philosophy
$\rightsquigarrow$ Accoding to ▸ Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If *S* was the case, *Q* would have been the case."

- *S* is an event that must relate to a past event that didn't occur
  $\rightsquigarrow$ counterfactuals run **contrary** to the **facts**

PHILOSOPHICAL BASIS

Counterfactuals have a long-standing tradition in analytic philosophy
$\rightsquigarrow$ Accoding to ▸ Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If S was the case, Q would have been the case."

- S is an event that must relate to a past event that didn't occur
  $\rightsquigarrow$ counterfactuals run **contrary** to the **facts**
- Above statement is true, if in all possible worlds most similar to the actual world where S had been the case, Q would have been the case

PHILOSOPHICAL BASIS

Counterfactuals have a long-standing tradition in analytic philosophy
$\rightsquigarrow$ Acoding to ▸ Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If S was the case, Q would have been the case."

- S is an event that must relate to a past event that didn't occur
  $\rightsquigarrow$ counterfactuals run **contrary** to the **facts**
- Above statement is true, if in all possible worlds most similar to the actual world where S had been the case, Q would have been the case
- A world is similar to another if laws are maximally preserved between the worlds and only a few facts are changed

PHILOSOPHICAL BASIS

- Counterfactuals have largely been studied to explain causal dependence

PHILOSOPHICAL BASIS

- Counterfactuals have largely been studied to explain causal dependence
- Causal dependence underlies the explanatory power
  $\rightsquigarrow$ good CEs point to critical causal factors that drove the algorithmic decision

PHILOSOPHICAL BASIS

- Counterfactuals have largely been studied to explain causal dependence
- Causal dependence underlies the explanatory power
  $\rightsquigarrow$ good CEs point to critical causal factors that drove the algorithmic decision
- If maximal closeness is relaxed, causally irrelevant factors can become part of the explanation
  $\rightsquigarrow$ e.g., decreasing loan amount by $20.000 and being one year older is recommended by the explainer although only loan amount might be causally relevant

PHILOSOPHICAL BASIS

- Counterfactuals have largely been studied to explain causal dependence
- Causal dependence underlies the explanatory power
  $\rightsquigarrow$ good CEs point to critical causal factors that drove the algorithmic decision
- If maximal closeness is relaxed, causally irrelevant factors can become part of the explanation
  $\rightsquigarrow$ e.g., decreasing loan amount by \$20.000 and being one year older is recommended by the explainer although only loan amount might be causally relevant
- CEs are often contrastive, i.e., they explain a decision by referring to an alternative outcome
  $\rightsquigarrow$ e.g., if the loan applicant was 30 instead of 60 years old, the approved loan would have been over \$100.000 instead of \$40.000

MATHEMATICAL PERSPECTIVE

Terminology:

- $\mathbf{x}$: original/factual datapoint whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired prediction ($y' = 1000$ or $y' =$ “grant credit") or interval ($y' = [1000, \infty[$)

MATHEMATICAL PERSPECTIVE

Terminology:

- $\mathbf{x}$: original/factual datapoint whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired prediction ($y' = 1000$ or $y' =$ “grant credit") or interval ($y' = [1000, \infty[$)

A **valid** counterfactual $\mathbf{x}'$ is a datapoint:

1. whose prediction $\hat{f}(\mathbf{x}')$ is equal to the desired prediction y'
2. that is maximally close to the original datapoint $\mathbf{x}$

MATHEMATICAL PERSPECTIVE

Terminology:

- $\mathbf{x}$: original/factual datapoint whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired prediction ($y' = 1000$ or $y' =$ “grant credit") or interval ($y' = [1000, \infty[$)

A **valid** counterfactual $\mathbf{x}'$ is a datapoint:

1. whose prediction $\hat{f}(\mathbf{x}')$ is equal to the desired prediction y'
2. that is maximally close to the original datapoint $\mathbf{x}$

Reformulate these two objectives (denoted by o_1 and o_2) as optimization problem:

$$\arg\min_{\mathbf{x}'} \lambda_1 o_p(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_f(\mathbf{x}', \mathbf{x})$$

- λ_1 and λ_2 balance the two objectives
- Choice of o_p (distance on prediction space) and of o_f (distance on feature space) is crucial

MATHEMATICAL PERSPECTIVE ‣ Dandl et al. (2020)

- Regression: o_p could be the L_1-distance $o_p(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- Classification: L_1-distance for scores and 0-1 Loss for labels, e.g., $o_p(\hat{f}(\mathbf{x}'), y') = \mathcal{I}_{\{\hat{f}(\mathbf{x}') \neq y'\}}$

MATHEMATICAL PERSPECTIVE ▸ Dandl et al. (2020)

- Regression: o_p could be the L_1-distance $o_p(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- Classification: L_1-distance for scores and 0-1 Loss for labels, e.g., $o_p(\hat{f}(\mathbf{x}'), y') = \mathcal{I}_{\{\hat{f}(\mathbf{x}') \neq y'\}}$
- o_f could be the Gower distance (suitable for mixed feature space):

$$o_f(\mathbf{x}', \mathbf{x}) = d_G(\mathbf{x}', \mathbf{x}) = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x_j', x_j) \in [0, 1]$$

The value of δ_G depends on the feature type (numerical or categorical):

$$\delta_G(x_j', x_j) = \begin{cases} \frac{1}{\widehat{R}_j} |x_j' - x_j| & \text{if } x_j \text{ is numerical} \\ \mathcal{I}_{\{x_j' \neq x_j\}} & \text{if } x_j \text{ is categorical} \end{cases}$$

with $\widehat{R}_j$ as the value range of feature j in the training dataset (to ensure that $\delta_G(x_j', x_j) \in [0, 1]$)

FURTHER OBJECTIVES

Additional constraints can improve the explanation quality of the corresponding CEs
$\rightsquigarrow$ popular constraints include sparsity and plausibility

Sparsity:

- End-users often prefer short over long explanations
  $\rightsquigarrow$ counterfactuals should be **sparse**

FURTHER OBJECTIVES

Additional constraints can improve the explanation quality of the corresponding CEs
$\rightsquigarrow$ popular constraints include sparsity and plausibility

Sparsity:

- End-users often prefer short over long explanations
  $\rightsquigarrow$ counterfactuals should be **sparse**
- Objective o_f can take the number of changed features into account (but does not have to)
  $\rightsquigarrow$ e.g., the L_0- and the L_1-norm (similar to LASSO) can do this

FURTHER OBJECTIVES

Additional constraints can improve the explanation quality of the corresponding CEs
$\rightsquigarrow$ popular constraints include sparsity and plausibility

Sparsity:

- End-users often prefer short over long explanations
  $\rightsquigarrow$ counterfactuals should be **sparse**
- Objective o_f can take the number of changed features into account (but does not have to)
  $\rightsquigarrow$ e.g., the L_0- and the L_1-norm (similar to LASSO) can do this
- Independently from o_f, sparsity in the changes can be additionally considered by another objective that counts the number of changed features via the L0-norm:

$$o_s(\mathbf{x}', \mathbf{x}) = \sum_{j=1}^{p} \mathcal{I}_{\{x'_j \neq x_j\}}$$

FURTHER OBJECTIVES

Plausibility:

- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time

FURTHER OBJECTIVES

Plausibility:

- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time
- CEs should be realistic and adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ avoid unrealistic combinations of feature values

FURTHER OBJECTIVES

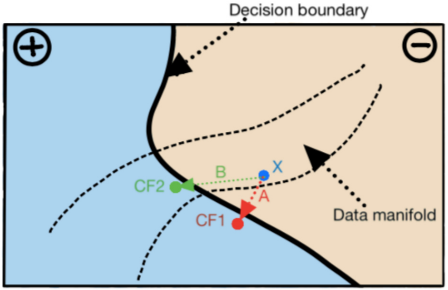
Plausibility:

- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time
- CEs should be realistic and adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ avoid unrealistic combinations of feature values
- Estimating joint distribution of training data is complex, especially for mixed feature spaces
  $\rightsquigarrow$ Proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$

FURTHER OBJECTIVES

Plausibility:

- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time
- CEs should be realistic and adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ avoid unrealistic combinations of feature values
- Estimating joint distribution of training data is complex, especially for mixed feature spaces
  $\rightsquigarrow$ Proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$



Example from ▸ Verma et al. (2020)

- Two possible paths for $\mathbf{x}$, originally classified to $\ominus$
- Two valid CEs in class $\oplus$: CF1 and CF2
- Path **A** for CF1 is shorter
- Path **B** for CF2 is longer but adheres to data manifold

FURTHER OBJECTIVES

To ensure plausibility, o_4 could, e.g., be the Gower distance of $\mathbf{x}'$ to its nearest data point of the training dataset which we denote $\mathbf{x}^{[1]}$:

$$o_4(\mathbf{x}', \mathbf{X}) = d_G(\mathbf{x}', \mathbf{x}^{[1]}) = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x'_j, x_j^{[1]})$$

We can extend the previous optimization problem by adding o_s (for sparsity) and o_4 (for plausibility):

$$\arg\min_{\mathbf{x}'} \lambda_1 o_p(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_f(\mathbf{x}', \mathbf{x}) + \lambda_3 o_s(\mathbf{x}', \mathbf{x}) + \lambda_4 o_4(\mathbf{x}', \mathbf{X})$$

REMARKS: THE RASHOMON EFFECT

Issue (Rashomon effect):

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

REMARKS: THE RASHOMON EFFECT

Issue (Rashomon effect):

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

Possible solutions:

- Present all CEs for a given $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one or few CEs (but: by which criterion should they be selected?)

REMARKS: THE RASHOMON EFFECT

Issue (Rashomon effect):

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

Possible solutions:

- Present all CEs for a given $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one or few CEs (but: by which criterion should they be selected?)

Note:

- As the model is generally non-linear, inconsistent and diverse CEs can arise
  e.g. suggesting either an increase or decrease in credit duration (confuses the explainee)
- How to deal with the Rashomon effect is considered an open problem in IML

REMARKS: MODEL OR REAL-WORLD

- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted

REMARKS: MODEL OR REAL-WORLD

- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- Consider a CE that proposes to increase the feature age by 5 to obtain the loan
  $\rightsquigarrow$ a loan applicant takes this information and applies 5 years later for the loan

REMARKS: MODEL OR REAL-WORLD

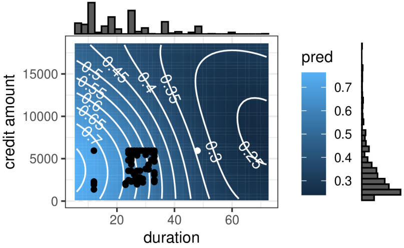
- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- Consider a CE that proposes to increase the feature age by 5 to obtain the loan
  $\rightsquigarrow$ a loan applicant takes this information and applies 5 years later for the loan
- However, by then, many other feature values might have changed
  $\rightsquigarrow$ not only age, also other causally dependent features e.g. job status might have changed
  $\rightsquigarrow$ ▸ Karimi et al. (2020) avoid this by considering causal dependencies between features

REMARKS: MODEL OR REAL-WORLD

- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- Consider a CE that proposes to increase the feature age by 5 to obtain the loan
  $\rightsquigarrow$ a loan applicant takes this information and applies 5 years later for the loan
- However, by then, many other feature values might have changed
  $\rightsquigarrow$ not only age, also other causally dependent features e.g. job status might have changed
  $\rightsquigarrow$ ▸ Karimi et al. (2020) avoid this by considering causal dependencies between features
- Also, the bank's algorithm might change and previous CEs are not applicable anymore

Interpretable Machine Learning

Methods & Discussion of CEs



Learning goals

- See two strategies to generate CEs
- Know problems and limitations of CEs

OVERVIEW OF METHODS

Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\rightsquigarrow$ so far, all methods remain in the supervised learning paradigm

OVERVIEW OF METHODS

Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\rightsquigarrow$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data

OVERVIEW OF METHODS

Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\rightsquigarrow$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data
- **Feature space:** Some methods can only handle numerical features, few can process mixed (numerical and discrete) feature spaces

OVERVIEW OF METHODS

Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\rightsquigarrow$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data
- **Feature space:** Some methods can only handle numerical features, few can process mixed (numerical and discrete) feature spaces
- **Objectives:** Many methods focus on action guidance, plausibility and sparsity, few on other objectives like fairness or individual preferences

OVERVIEW OF METHODS

Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\rightsquigarrow$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data
- **Feature space:** Some methods can only handle numerical features, few can process mixed (numerical and discrete) feature spaces
- **Objectives:** Many methods focus on action guidance, plausibility and sparsity, few on other objectives like fairness or individual preferences
- **Model access:** Methods either require access to complete model internals, access to gradients, or only to prediction functions $\Rightarrow$ Model-agnostic and model-specific methods exist

OVERVIEW OF METHODS

Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\rightsquigarrow$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data
- **Feature space:** Some methods can only handle numerical features, few can process mixed (numerical and discrete) feature spaces
- **Objectives:** Many methods focus on action guidance, plausibility and sparsity, few on other objectives like fairness or individual preferences
- **Model access:** Methods either require access to complete model internals, access to gradients, or only to prediction functions $\Rightarrow$ Model-agnostic and model-specific methods exist
- **Optimization tool:** Gradient-based algorithms (only for differentiable models), mixed-integer programming (only linear), or gradient-free algorithms e.g. Nelder-Mead, genetic algorithm

OVERVIEW OF METHODS

Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\rightsquigarrow$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data
- **Feature space:** Some methods can only handle numerical features, few can process mixed (numerical and discrete) feature spaces
- **Objectives:** Many methods focus on action guidance, plausibility and sparsity, few on other objectives like fairness or individual preferences
- **Model access:** Methods either require access to complete model internals, access to gradients, or only to prediction functions $\Rightarrow$ Model-agnostic and model-specific methods exist
- **Optimization tool:** Gradient-based algorithms (only for differentiable models), mixed-integer programming (only linear), or gradient-free algorithms e.g. Nelder-Mead, genetic algorithm
- **Rashomon Effect:** Many methods return a single counterfactual per run, some multiple counterfactuals, others prioritize CEs or let the user choose

FIRST OPTIMIZATION METHOD ▸ Wachter et. al (2018)

Introduced counterfactual explanations in the context of ML predictions by solving

$$\arg\min_{\mathbf{x}'} \max_{\lambda} \lambda \underbrace{(\hat{f}(\mathbf{x}') - y')^2}_{o_p(\hat{f}(\mathbf{x}'),y')} + \underbrace{\sum\nolimits_{j=1}^{p} |x'_j - x_j| / MAD_j}_{o_f(\mathbf{x}',\mathbf{x})} \tag{1}$$

MAD_j is the median absolute deviation of feature j. In each iteration, optimizers like Nelder-Mead solve the equation for $\mathbf{x}'$ and then λ is increased until a sufficiently close solution is found

This optimization problem has several shortcomings:

- We do not know how to choose λ a priori
- Due to the maximization of λ, we focus primarily on the minimization of o_p
  $\leadsto$ only if $\hat{f}(\mathbf{x}') = y'$, we focus on minimizing o_f
- Definition of o_f only covers numerical features
- Other objectives such as sparsity and plausibility of counterfactuals are neglected

MULTI-OBJECTIVE COUNTERFACTUAL EXPLANATIONS ▸ Dandl et al. (2020)

- **Multi-Objective Counterfactual Explanations (MOC):** Instead of collapsing objectives into a single objective, we could optimize all four objectives simultaneously

$$\arg\min_{\mathbf{x}'} \left(o_p(\hat{f}(\mathbf{x}'), y'), o_f(\mathbf{x}', \mathbf{x}), o_s(\mathbf{x}', \mathbf{x}), o_4(\mathbf{x}', \mathbf{X}) \right).$$

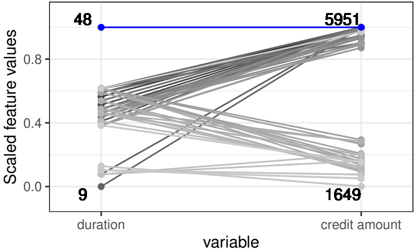
- Note that weighting parameters like λ are not necessary anymore
- Uses an adjusted multi-objective genetic algorithm (NSGA-II) to produce a set of diverse counterfactuals for mixed discrete and continuous feature spaces
- Instead of one, MOC returns multiple counterfactuals that represents different trade-offs between the objectives and are constructed to be diverse in feature space

EXAMPLE: CREDIT DATA

- Model: SVM with RBF kernel
- **x**: First data point of credit data with $\mathbb{P}(y = good) = 0.34$ of being a “good" customer
- Goal: Increase the probability to $[0.5, 1]$
- MOC (with default parameters) found 69 CEs after 200 iterations that met the target
- All counterfactuals proposed changes to credit duration and many of them to credit amount

EXAMPLE: CREDIT DATA ▸ Dandl et al. (2020)

- We can visualize feature changes with a parallel plot and 2-dim surface plot
- Parallel plot reveals that all counterfactuals had values equal to or smaller than the values of **x**

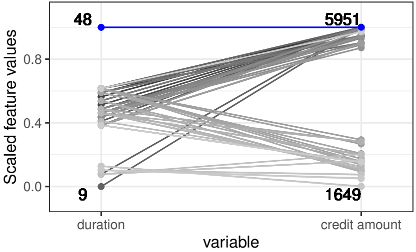


Parallel plot: Grey lines show feature values of CEs $\mathbf{x}'$, blue line are values of **x**. Features without proposed changes are omitted. Bold numbers refer to range of numeric features.

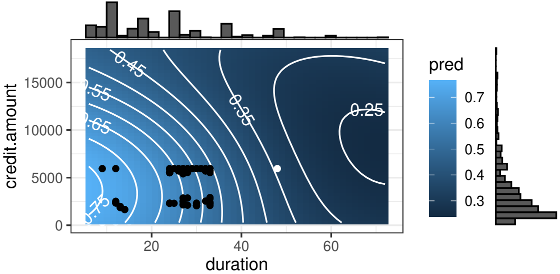
EXAMPLE: CREDIT DATA

▸ Dandl et al. (2020)

- We can visualize feature changes with a parallel plot and 2-dim surface plot
- Parallel plot reveals that all counterfactuals had values equal to or smaller than the values of $\mathbf{x}$
- Surface plot illustrates why these feature changes are recommended
- Counterfactuals in the lower left corner seem to be in a less favorable region far from $\mathbf{x}$, but they are in high density areas close to training samples (indicated by histograms)



Parallel plot: Grey lines show feature values of CEs $\mathbf{x}'$, blue line are values of $\mathbf{x}$. Features without proposed changes are omitted. Bold numbers refer to range of numeric features.



Surface plot: White dot is $\mathbf{x}$, black dots are CEs $\mathbf{x}'$. Histograms show marginal distribution of training data $\mathbf{X}$.

PROBLEMS, PITFALLS, & LIMITATIONS

- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives which reduces complexity, but is limited in explanatory power
  $\rightsquigarrow$ Psychologists have shown that although perceived model understanding of end-users increases, the objective model understanding remains unchanged

PROBLEMS, PITFALLS, & LIMITATIONS

- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives which reduces complexity, but is limited in explanatory power
  $\rightsquigarrow$ Psychologists have shown that although perceived model understanding of end-users increases, the objective model understanding remains unchanged
- **Right metric:** Similarity measures are crucial to find good CEs (depends on context/domain)
  $\rightsquigarrow$ e.g., L_1 can be reasonable for tabular data but not for image data
  $\rightsquigarrow$ sparsity can be desirable for end-users but not for data scientists searching for model bias

PROBLEMS, PITFALLS, & LIMITATIONS

- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives which reduces complexity, but is limited in explanatory power
  $\leadsto$ Psychologists have shown that although perceived model understanding of end-users increases, the objective model understanding remains unchanged
- **Right metric:** Similarity measures are crucial to find good CEs (depends on context/domain)
  $\leadsto$ e.g., L_1 can be reasonable for tabular data but not for image data
  $\leadsto$ sparsity can be desirable for end-users but not for data scientists searching for model bias
- **Confusing Model and Real-World:** Model explanations are not easily transferable to reality
  $\leadsto$ End-users need to be aware that CE provide insights into a model not the real world

PROBLEMS, PITFALLS, & LIMITATIONS

- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives which reduces complexity, but is limited in explanatory power
  $\rightsquigarrow$ Psychologists have shown that although perceived model understanding of end-users increases, the objective model understanding remains unchanged
- **Right metric:** Similarity measures are crucial to find good CEs (depends on context/domain)
  $\rightsquigarrow$ e.g., L_1 can be reasonable for tabular data but not for image data
  $\rightsquigarrow$ sparsity can be desirable for end-users but not for data scientists searching for model bias
- **Confusing Model and Real-World:** Model explanations are not easily transferable to reality
  $\rightsquigarrow$ End-users need to be aware that CE provide insights into a model not the real world
- **Disclosing too much information:**
  CEs can reveal too much information about the model and help potential attackers

PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  $\rightsquigarrow$ No perfect solution, depends on end-users computational resources and knowledge

PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  $\rightsquigarrow$ No perfect solution, depends on end-users computational resources and knowledge
- **Actionability vs. fairness:** Some authors suggest to focus only on the actionability of CEs
  $\rightsquigarrow$ Counteract contestability, e.g., if ethnicity is not changed in a CE since it is not actionable, this could hide racial biases in the model

PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  $\rightsquigarrow$ No perfect solution, depends on end-users computational resources and knowledge
- **Actionability vs. fairness:** Some authors suggest to focus only on the actionability of CEs
  $\rightsquigarrow$ Counteract contestability, e.g., if ethnicity is not changed in a CE since it is not actionable, this could hide racial biases in the model
- **Assumption of constant model:** To provide guidance for the future, CEs assume that their underlying model does not change in the future
  $\rightsquigarrow$ in reality this assumption is often violated and CEs are not reliable anymore

PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  $\rightsquigarrow$ No perfect solution, depends on end-users computational resources and knowledge
- **Actionability vs. fairness:** Some authors suggest to focus only on the actionability of CEs
  $\rightsquigarrow$ Counteract contestability, e.g., if ethnicity is not changed in a CE since it is not actionable, this could hide racial biases in the model
- **Assumption of constant model:** To provide guidance for the future, CEs assume that their underlying model does not change in the future
  $\rightsquigarrow$ in reality this assumption is often violated and CEs are not reliable anymore
- **Attacking CEs:** Researchers can create models with great performance, which generate arbitrary explanations specified by the ML developer
  $\rightsquigarrow$ how faithful are CEs to the models underlying mechanism?

Interpretable Machine Learning

Local Explanations: Adversarial Examples

Learning goals

- Understand the definition of ADEs
- Understand first methods that generate ADEs
- Discuss potential causes of ADEs and standard defenses against them

ADVERSARIAL MACHINE LEARNING

- What happens if a computer system gets an erroneous input?
- Even worse:
  What happens if someone feeds in a malicious input on purpose to attack a system?

$\rightsquigarrow$ **Robustness** is important to ensure a safe service!

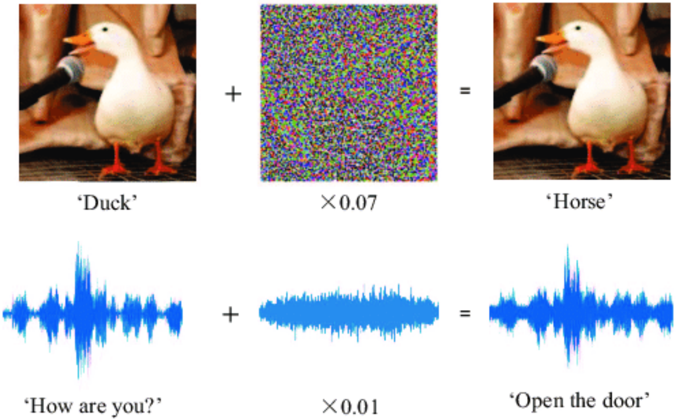
- **Adversarial ML** studies the robustness of machine learning (ML) algorithms to malicious input
- Two different kinds of attacks:
  - **Evasion attacks** mislead an employed ML model with manipulated inputs (our focus)
  - **Data Poisoning**: Malicious inputs to the training dataset

ADVERSARIAL EXAMPLES

- **Informal Definition**: An ADE is an input to a model that is deliberately designed to "fool" the model into misclassifying it
- Even possible with low generalization error
- Both deep learning models (e.g., CNNs) and classical ML can be vulnerable to such attacks
- ADEs created from a real data observation $\mathbf{x}$ can be indistinguishable from $\mathbf{x}$ by a human observer
- Since the model misclassifies this input, it does not seem to have a real understanding of the underlying concepts of the provided inputs

EXAMPLES: MODEL-ATTACKS

Gong & Poellabauer 2018



- Is this a duck or a horse?
- Small (hard-to-see) noise can change the prediction

EXAMPLES: IMAGE DATA ▸ Eykholt et al. (2018) ▸ Athalye et al. (2018)

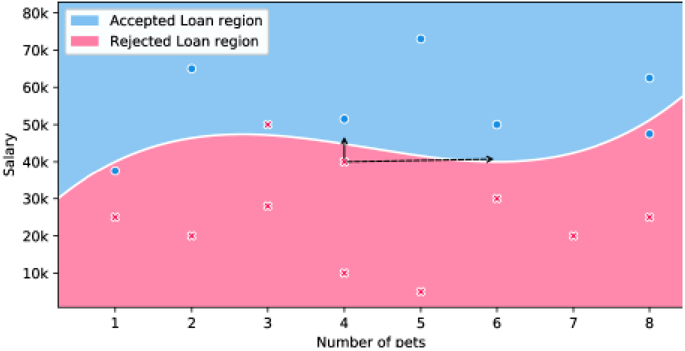
- Stop signs can be missclassified
  e.g., because of graffiti
- With some well-placed patches, the model identifies it as a “right of way” sign

- 3D-print of a turtle
- Misclassified as a rifle (from every angle)
- Video: ▸ MITCSAIL (2017)

EXAMPLE: TABULAR DATA ▸ Ballet (2019)

What is imperceptibility on tabular data?

- Idea: experts focus on the most important features in their judgment
- An ADE arises from manipulating features the model deems important but experts do not



Decision boundary of a classifier deciding loan applications. ADE via “number of pets”

ADE AND INTERPRETABILITY

1. ADEs show where models fail $\rightsquigarrow$ improved model understanding
2. Because of ADEs, we need more interpretability
3. Interpretation can lead to robustness against ADEs
4. Explanations can be used to construct ADEs (e.g., see numer of pets on previous slide)

FORMAL DEFINITION

Adversarial Input

Let $\epsilon > 0$, $f : \mathcal{X} \to \mathcal{Y}$ be an ML model and $\mathbf{x} \in \mathcal{X}$ be a real data point that is correctly classified: $f(\mathbf{x}) = y_{\mathbf{x},true}$.

We call $\mathbf{a_x}$ an **adversarial input** to $\mathbf{x}$ if:

$$\|\mathbf{a_x} - \mathbf{x}\| < \epsilon \text{ and } f(\mathbf{a_x}) \neq y_{\mathbf{a_x},true} = f(\mathbf{x})$$

- $\mathbf{a_x}$ an is a data point close to a real, correctly classified input that is misclassified
- $\mathbf{a_x}$ is called **targeted** if the class it is assigned to is determined $f(\mathbf{a_x}) = y'$ with y' being a desired prediction
- Can be generalized to regression problems

WHY DO ADE EXIST?

Non-exhaustive list of hypotheses:

1. Low-probability spaces hypotheses: ADEs live in low-probability yet dense spaces in the data manifold that are not well represented in the training samples ▸ Szegedy et al. (2013)

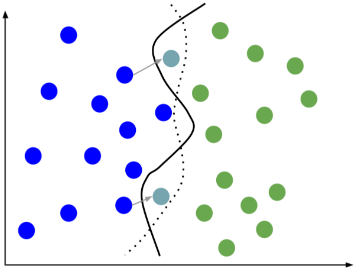


Figure: Binary classification example (dark blue vs. green dots). Dotted line represents the true decision boundary, bold line the trained one. Low probability space close to decision boundary allow for adversarial examples (turquoise dot).

WHY DO ADE EXIST?

Non-exhaustive list of hypotheses:

2. Linearity hypotheses (most popular):
Adversarial examples are omnipresent in the data manifold
$\rightsquigarrow$ occur, because commonly used models often show linear behavior
$\rightsquigarrow$ small changes of ϵ in every feature cause a change of $\epsilon||\boldsymbol{\theta}||_1$ in prediction ▸ Goodfellow et al. (2014)

Example: linear model
Original: $f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\theta}$
Small changes: $f(\mathbf{x} + \epsilon) = (\mathbf{x} + \epsilon)^T\boldsymbol{\theta}$
Difference: $f(\mathbf{x} + \epsilon) - f(\mathbf{x}) = \epsilon \cdot \boldsymbol{\theta}$

WHY DO ADE EXIST?

Non-exhaustive list of hypotheses:

3. The boundary tilting hypothesis: Linearity is neither necessary nor sufficient to explain ADEs
$\rightsquigarrow$ ADEs mostly result from overfitting the sampled manifold ▸ Tanay and Griffin (2016)

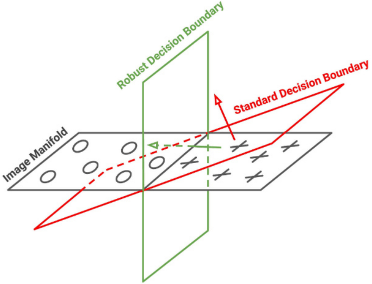


Figure: Linear binary classification example. Due to overfitting the decision boundary (red) is close to the manifold of the training data. Techniques like regularization could help to make the decision boundary more robust (green). ▸ Kim et al. (2019)

WHY DO ADE EXIST?

Non-exhaustive list of hypotheses:

4. Human-centric hypotheses: ML models make use of predictive but non-robust features – meaning they are highly correlated with the prediction target, but not used by humans

▸ Ilyas et al. (2019)

WAYS TO GENERATE ADE

Different ways for constructing ADEs: There exist various ways in the literature to generate ADEs for a given model in feasible time

- Formulate the search for ADEs as an **optimization problem**, e.g.

$$\underset{\mathbf{x}' \in \mathcal{X}}{\text{argmin}} \; \|\mathbf{x} - \mathbf{x}'\|_{\mathcal{X}} + \lambda \, \|f(\mathbf{x}') - y'\|_{\mathcal{Y}}$$

- Use **sensitivity analysis** to identify features that influence the target class
- Train a generative adversarial network (GAN) ▸ Goodfellow et al. (2014)

Moreover, depending on the attacker's model access, we can distinguish between

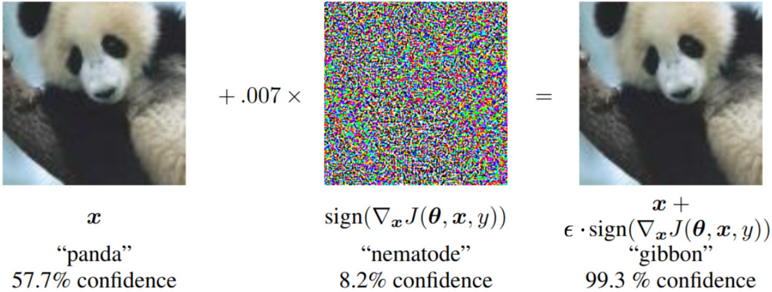
- **Full-access attacks**: the attacker has full access to the internals of the model
- **Black-box attacks**: the attacker can only query the model on some inputs and receives the model's outputs

FAST-GRADIENT-SIGN-METHOD (FGSM) ▸ Goodfellow et al. (2015)

- FGSM is based on the linearity hypothesis
- FGSM finds ADEs from:

$$a_{\mathbf{x}} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y_{\mathbf{x},true}))$$

where $\text{sign}(\nabla_{\mathbf{x}} J(\theta, x, y_{\mathbf{x},true}))$ describes the component-wise signum of the gradient of cost function J in $\mathbf{x}$ with true label $y_{\mathbf{x},true}$



NOTES ON FGSM ▸ Goodfellow et al. (2015) ▸ Lyu et al. 2015

- FGSM works particularly well for linear(-like) models in high-dimensional spaces, e.g., LSTMs, logistic regressions or CNNs with ReLU activations
- Not every $\mathbf{a_x}$ generated by FGSM is an ADE, especially if ϵ is too small
- FGSM attacks can be also generated without model access by approximating the gradient, e.g. with finite difference methods
- The notion of similarity in FGSM is based on $\|\cdot\|_\infty \rightsquigarrow$ there are generalizations of FGSM to other norms

BLACK-BOX ATTACKS WITH SURROGATES ▸ Papernot et al. (2016)

- So far, we assumed full access to the predictive model
- Black-box attacks only assume query-access
- Large risk of attacks since often one can query predictive models many times

1. Query the model you aim to attack as often as allowed on data similar to the training data
2. Use the labeled data you received to train a surrogate model
3. Generate ADEs for the surrogate model
4. Use these ADEs to attack the original model

$\rightsquigarrow$ Known as the **transferability** of ADEs.

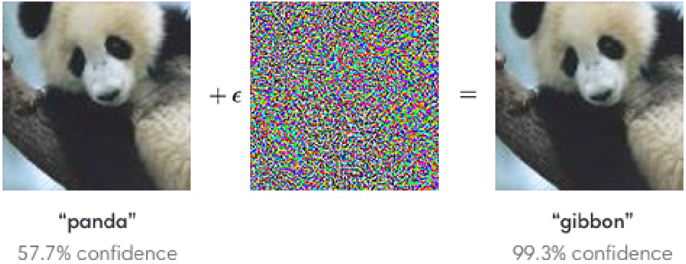
DEFENSES AGAINST ADE

There are several ways to protect your network against such attacks – we distinguish between two broad types of defenses, differing in the position in which they act

- **Guards** act on the inputs a model receives
  - **Detect anomalies:** e.g., statistical testing, or discriminator networks from GANs
  - **Conduct transformations** on inputs (e.g. PCA)
- **Defense by design** act on the model itself
  - **Adversarial training**: train model on adversarials
  - **Architectural defenses**: e.g., removing low predictive features from the model

SUMMARY

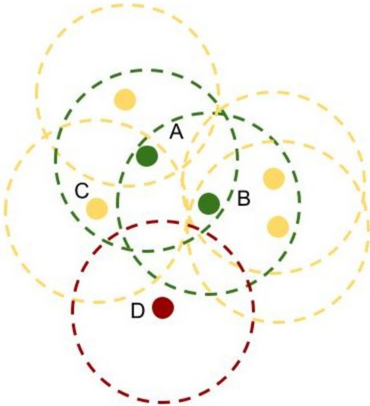
- ADEs are not explanations themselves but are conceptually connected to them
- ADEs can be generated in diverse settings $\rightsquigarrow$ crucial modeling decisions are the distance measure, the local environment, and the target level (model or process)
- There are various hypotheses on the existence of ADEs which also motivate different defense strategies



▸ Goodfellow et al. (2017)

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

MOTIVATION & IMPORTANT PROPERTIES

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

MOTIVATION & IMPORTANT PROPERTIES

- Local explanations should not only make a model interpretable but also reveal if the model is trustworthy
- **Interpretable**: “Why did the model come up with this decision?”

MOTIVATION & IMPORTANT PROPERTIES

- Local explanations should not only make a model interpretable but also reveal if the model is trustworthy
- **Interpretable**: “Why did the model come up with this decision?”
- **Trustworthy**: “How certain is this explanation?”
  1. accurate insights into the inner workings of our model
     - Failure case: generation is based on inputs in areas where the model was trained with little or no training data (extrapolation)

MOTIVATION & IMPORTANT PROPERTIES

- Local explanations should not only make a model interpretable but also reveal if the model is trustworthy
- **Interpretable**: “Why did the model come up with this decision?”
- **Trustworthy**: “How certain is this explanation?”
  1. accurate insights into the inner workings of our model
     - Failure case: generation is based on inputs in areas where the model was trained with little or no training data (extrapolation)
  2. robust (i.e. low variance)
     - Expectation: similar explanations for similar data points with similar predictions
     - However, multiple sources of uncertainty exist

     $\rightsquigarrow$ measure how robust an IML method is to small changes in the input data or parameters

     $\rightsquigarrow$ Is an observation out-of-distribution?

MOTIVATION & IMPORTANT PROPERTIES

- Local explanations should not only make a model interpretable but also reveal if the model is trustworthy
- **Interpretable**: “Why did the model come up with this decision?”
- **Trustworthy**: “How certain is this explanation?”
  1. accurate insights into the inner workings of our model
     - Failure case: generation is based on inputs in areas where the model was trained with little or no training data (extrapolation)
  2. robust (i.e. low variance)
     - Expectation: similar explanations for similar data points with similar predictions
     - However, multiple sources of uncertainty exist

     $\leadsto$ measure how robust an IML method is to small changes in the input data or parameters

     $\leadsto$ Is an observation out-of-distribution?
- Failing in one of these $\leadsto$ undermining users’ trust in the explanations
  $\leadsto$ undermining trust in the model

OUT-OF-DISTRIBUTION DETECTION

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

OUT-OF-DISTRIBUTION DETECTION

- Models are unreliable in areas with little data support
  $\leadsto$ explanations from local explanation methods are unreliable
- For local explanation methods, the following components could be out-of-distribution (OOD):
  - The data for LIME's surrogate model
  - Counterfactuals themselves
  - Shapley value's permuted observations to calculate the marginal contributions
  - ICE curves grid data points

OUT-OF-DISTRIBUTION DETECTION

- Models are unreliable in areas with little data support
  $\rightsquigarrow$ explanations from local explanation methods are unreliable
- For local explanation methods, the following components could be out-of-distribution (OOD):
  - The data for LIME's surrogate model
  - Counterfactuals themselves
  - Shapley value's permuted observations to calculate the marginal contributions
  - ICE curves grid data points
- Two very simple and intuitive approaches
  - Classifier for out-of-distribution
  - Clustering
- More complicated also possible, e.g., variational autoencoders [Daxberger et al. 2020]

OUT-OF-DISTRIBUTION DETECTION: OOD-CLASSIFIER

- Problem: we have only in-distribution data
- Idea: Hallucinate new (out-of-distribution) data by randomly sample data points

$\rightsquigarrow$ Learn a binary classifier to distinguish between the origins of the data

OUT-OF-DISTRIBUTION DETECTION: OOD-CLASSIFIER

- Problem: we have only in-distribution data
- Idea: Hallucinate new (out-of-distribution) data by randomly sample data points

$\rightsquigarrow$ Learn a binary classifier to distinguish between the origins of the data

- Study whether an explanation approach can be fooled ▸ Dylan Slack et al. 2020
  - Hide bias in the true (deployed) model, but use an unbiased model for all out-of-distribution samples

$\rightsquigarrow$ Important way to diagnose an explanation approach

OUT-OF-DISTRIBUTION DETECTION: CLUSTERING VIA DBSCAN

- DBSCAN is a data clustering algorithm ▸ Martin Ester et al. 1996
  (Density-Based Spatial Clustering of Applications with Noise)

OUT-OF-DISTRIBUTION DETECTION: CLUSTERING VIA DBSCAN

- DBSCAN is a data clustering algorithm ▸ Martin Ester et al. 1996
  (Density-Based Spatial Clustering of Applications with Noise)
- For this method, we define an ϵ-neighborhood:
  Given a dataset $X = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$, an ϵ-neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as

  $$\mathcal{N}_{\epsilon}(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(i)}) \leq \epsilon\}.$$

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

OUT-OF-DISTRIBUTION DETECTION: CLUSTERING VIA DBSCAN

- DBSCAN is a data clustering algorithm ▸ Martin Ester et al. 1996
  (Density-Based Spatial Clustering of Applications with Noise)
- For this method, we define an ϵ-neighborhood:
  Given a dataset $X = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$, an ϵ-neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as

  $$\mathcal{N}_\epsilon(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(i)}) \leq \epsilon\}.$$

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

OUT-OF-DISTRIBUTION DETECTION: CLUSTERING VIA DBSCAN

- DBSCAN is a data clustering algorithm ▸ Martin Ester et al. 1996
  (Density-Based Spatial Clustering of Applications with Noise)
- For this method, we define an ϵ-neighborhood:
  Given a dataset $X = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$, an ϵ-neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as

  $$\mathcal{N}_\epsilon(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(i)}) \leq \epsilon\}.$$

  $d(\cdot)$ is a distance measure (e.g., Euclidean or Gower distance)
- Core observations $\mathbf{x}$
  - Have at least m data points within $\mathcal{N}_\epsilon(\mathbf{x})$
  - Forms an own cluster with all its neighborhood points
- Border points
  - Within $\mathcal{N}_\epsilon(\mathbf{x})$
  - Part of a cluster defined by a core point

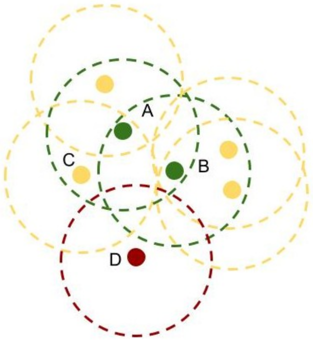
OUT-OF-DISTRIBUTION DETECTION: CLUSTERING VIA DBSCAN

- DBSCAN is a data clustering algorithm ▸ Martin Ester et al. 1996
  (Density-Based Spatial Clustering of Applications with Noise)
- For this method, we define an ϵ-neighborhood:
  Given a dataset $X = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$, an ϵ-neighborhood for $\mathbf{x} \in \mathcal{X}$ is defined as

$$\mathcal{N}_{\epsilon}(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} | d(\mathbf{x}, \mathbf{x}^{(i)}) \leq \epsilon\}.$$

  $d(\cdot)$ is a distance measure (e.g., Euclidean or Gower distance)
- Core observations $\mathbf{x}$
  - Have at least m data points within $\mathcal{N}_{\epsilon}(\mathbf{x})$
  - Forms an own cluster with all its neighborhood points
- Border points
  - Within $\mathcal{N}_{\epsilon}(\mathbf{x})$
  - Part of a cluster defined by a core point
- Noise points
  - Are not within $\mathcal{N}_{\epsilon}(\mathbf{x})$
  - Not part of any cluster

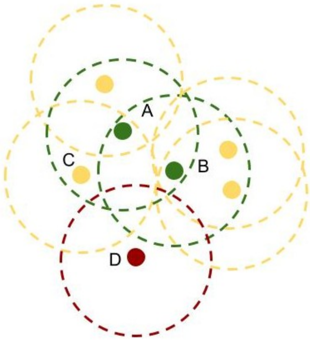
OUT-OF-DISTRIBUTION DETECTION



Example for DBSCAN, circles display ϵ-neighborhoods, $m = 4$

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

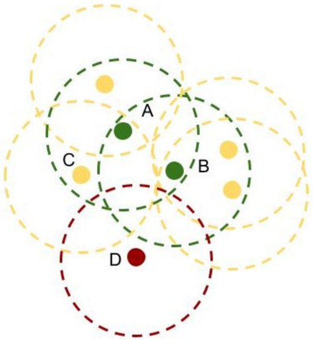
OUT-OF-DISTRIBUTION DETECTION



Example for DBSCAN, circles display ϵ-neighborhoods, $m = 4$

- Green points A and B are core points and form one cluster since they lie in each others neighborhood, all yellow points are border points of this cluster
- Since D is not part of the neighborhood of core points, it is a noise point

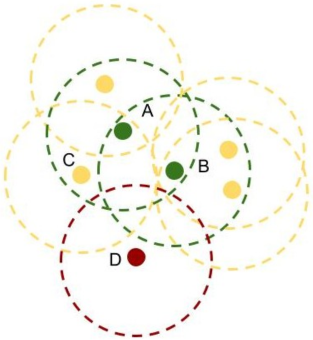
OUT-OF-DISTRIBUTION DETECTION



Example for DBSCAN, circles display ϵ-neighborhoods, $m = 4$

- Green points A and B are core points and form one cluster since they lie in each others neighborhood, all yellow points are border points of this cluster
- Since D is not part of the neighborhood of core points, it is a noise point
- In-distribution: new point lies within a cluster

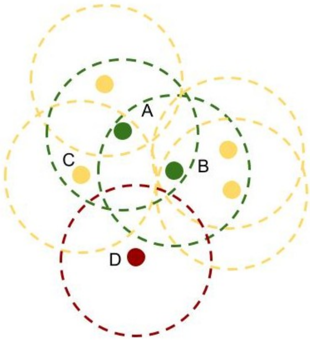
OUT-OF-DISTRIBUTION DETECTION



Example for DBSCAN, circles display ϵ-neighborhoods, $m = 4$

- Green points A and B are core points and form one cluster since they lie in each others neighborhood, all yellow points are border points of this cluster
- Since D is not part of the neighborhood of core points, it is a noise point
- In-distribution: new point lies within a cluster
- Out-of-distribution: new point lies outside the clusters

OUT-OF-DISTRIBUTION DETECTION



Example for DBSCAN, circles display ϵ-neighborhoods, $m = 4$

- Green points A and B are core points and form one cluster since they lie in each others neighborhood, all yellow points are border points of this cluster
- Since D is not part of the neighborhood of core points, it is a noise point
- In-distribution: new point lies within a cluster
- Out-of-distribution: new point lies outside the clusters

- Disadvantages:
  - Depending on the distance metric $d(\cdot)$, DBSCAN could suffer from the “curse of dimensionality”
  - The choice of ϵ and m is not clear a-priori

ROBUSTNESS

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

ROBUSTNESS

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

ROBUSTNESS

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

ROBUSTNESS MEASURE FOR LIME AND SHAP

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

ROBUSTNESS MEASURE FOR LIME AND SHAP

- Objective: Similar explanations for similar inputs (in a neighborhood)
- For LIME and SHAP, notion of stability based on **locally Lipschitz continuity** ▸ Alvarez-Melis and Jaakkola 2018:

  An explanation method $g : \mathcal{X} \to \mathbb{R}^m$ is locally Lipschitz if

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

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

ROBUSTNESS MEASURE FOR LIME AND SHAP

- Objective: Similar explanations for similar inputs (in a neighborhood)
- For LIME and SHAP, notion of stability based on **locally Lipschitz continuity** ▸ Alvarez-Melis and Jaakkola 2018:

  An explanation method $g : \mathcal{X} \rightarrow \mathbb{R}^m$ is locally Lipschitz if

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

  Note that, for LIME, g returns the m coefficients of the surrogate model
- According to this, we can quantify the robustness of explanation models in terms of ω:

  $\rightsquigarrow$ The closer ω is to 0, the more robust our explanation method is

ROBUSTNESS MEASURE FOR LIME AND SHAP

- Objective: Similar explanations for similar inputs (in a neighborhood)
- For LIME and SHAP, notion of stability based on **locally Lipschitz continuity** ▸ Alvarez-Melis and Jaakkola 2018:

  An explanation method $g : \mathcal{X} \to \mathbb{R}^m$ is locally Lipschitz if
  - for every $\mathbf{x}_0 \in \mathcal{X}$ there exist $\delta > 0$ and $\omega \in \mathbb{R}$
  - such that $||\mathbf{x} - \mathbf{x}_0|| < \delta$ implies $||g(\mathbf{x}) - g(\mathbf{x}_0)|| < \omega ||\mathbf{x} - \mathbf{x}_0||$

  Note that, for LIME, g returns the m coefficients of the surrogate model
- According to this, we can quantify the robustness of explanation models in terms of ω:

  ⇝ The closer ω is to 0, the more robust our explanation method is
- ω is rarely known a-priori but it could be estimated as follows:

$$\hat{\omega}_X(\mathbf{x}) \in \underset{\mathbf{x}^{(i)} \in \mathcal{N}_\epsilon(\mathbf{x})}{\arg\max} \frac{||g(\mathbf{x}) - g(\mathbf{x}^{(i)})||_2}{d(\mathbf{x}, \mathbf{x}^{(i)})},$$

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