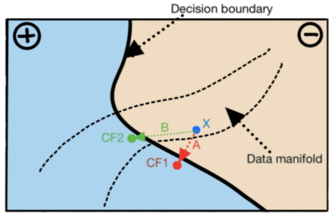
Interpretable Machine Learning

CE: Optimization Problem and Objectives



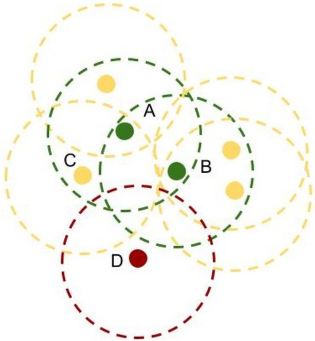
Learning goals

- Formulate CEs as optimization problem
- Identify key objectives (proximity, sparsity)
- Understand trade-offs in CE generation

magentaDaxberger et al. 2020

Interpretable Machine Learning

Local Explanations: Increasing Trust in Explanations



Learning goals

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

MATHEMATICAL PERSPECTIVE

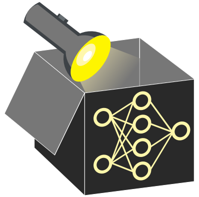
Terminology:

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

MOTIVATION & IMPORTANT PROPERTIES

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

MATHEMATICAL PERSPECTIVE



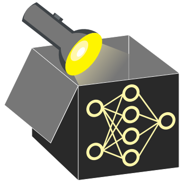
Terminology:

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

A **valid** counterfactual $\mathbf{x}'$ satisfies two criteria:

1. **Prediction validity:** CE’s prediction $\hat{f}(\mathbf{x}')$ is equal to the desired prediction y'
2. **Proximity:** CE $\mathbf{x}'$ is as close as possible to the original input $\mathbf{x}$

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?”

MATHEMATICAL PERSPECTIVE

Terminology:

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

A **valid** counterfactual $\mathbf{x}'$ satisfies two criteria:

1. **Prediction validity:** CE's prediction $\hat{f}(\mathbf{x}')$ is equal to the desired prediction y'
2. **Proximity:** CE $\mathbf{x}'$ is as close as possible to the original input $\mathbf{x}$

Reformulate these two objectives as optimization problem:

$$\arg\min_{\mathbf{x}'} \lambda_1 o_{target}(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_{proximity}(\mathbf{x}', \mathbf{x})$$

- λ_1 and λ_2 balance the two objectives
- o_{target}: distance in target space
- $o_{proximity}$: distance in feature space

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)

OBJECTIVE FUNCTIONS ▸ Dandl et al. (2020)

Distance in target space o_{target}:

- **Regression:** L_1 distance $o_{target}(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- **Classification:**
  - For predicted probabilities: $o_{target} = |\hat{f}(\mathbf{x}') - y'|$
  - For predicted hard labels: $o_{target} = \mathbb{I}\{\hat{f}(\mathbf{x}') \neq y'\}$

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?

OBJECTIVE FUNCTIONS ▸ Dandl et al. (2020)

Distance in target space o_{target}:

- **Regression:** L_1 distance $o_{target}(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- **Classification:**
  - For predicted probabilities: $o_{target} = |\hat{f}(\mathbf{x}') - y'|$
  - For predicted hard labels: $o_{target} = \mathbb{I}\{\hat{f}(\mathbf{x}') \neq y'\}$

Distance in input space $o_{proximity}$: Gower distance (mixed feature types)

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

- $\delta_G(x_j', x_j) = \mathbb{I}\left\{x_j' \neq x_j\right\}$ if x_j is categorical
- $\delta_G(x_j', x_j) = \frac{1}{\widehat{R}_j} |x_j' - x_j|$ if x_j is numerical

$\rightsquigarrow$ $\widehat{R}_j$ is the range of feature j in the training set to ensure $\delta_G(x_j', x_j) \in [0, 1]$

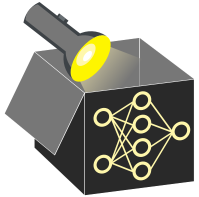
MOTIVATION & IMPORTANT PROPERTIES

- 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?
- Failing in one of these $\rightsquigarrow$ undermining users' trust in the explanations $\rightsquigarrow$ undermining trust in the model

FURTHER OBJECTIVES: SPARSITY

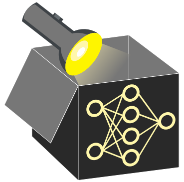


Additional constraints can improve the explanation quality of the corresponding CEs
⇝ popular constraints include **sparsity** and **plausibility**

Sparsity Favor explanations that change few features

- End-users often prefer short over long explanations

OUT-OF-DISTRIBUTION (OOD) DETECTION



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

FURTHER OBJECTIVES: SPARSITY

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

Sparsity Favor explanations that change few features

- End-users often prefer short over long explanations
- Sparsity could be integrated into $o_{proximity}$
  e.g., using L_0-norm (number of changed features) or L_1-norm (LASSO)

OUT-OF-DISTRIBUTION (OOD) 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 obs. to calculate the marginal contribs
  - ICE curves grid data points

FURTHER OBJECTIVES: SPARSITY

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

Sparsity Favor explanations that change few features

- End-users often prefer short over long explanations
- Sparsity could be integrated into $o_{proximity}$
  e.g., using L_0-norm (number of changed features) or L_1-norm (LASSO)
- Alternative: Include separate objective measuring sparsity, e.g., via L_0-norm

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

OUT-OF-DISTRIBUTION (OOD) 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 obs. to calculate the marginal contribs
  - 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 2020

FURTHER OBJECTIVES: PLAUSIBILITY

Plausibility:

- CEs should suggest realistic (i.e., plausible) alternatives
  ⇝ Implausible: increase income *and* become unemployed

OOD DETECTION: OOD-CLASSIFIER

- Problem: we have only in-distribution data
- Idea: Hallucinate new (ood) data by randomly sampling data points

⇝ Learn a binary classifier to distinguish between the origins of the data

FURTHER OBJECTIVES: PLAUSIBILITY

Plausibility:

- CEs should suggest realistic (i.e., plausible) alternatives
  $\rightsquigarrow$ Implausible: increase income *and* become unemployed
- CEs should adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ Avoid unrealistic combinations of feature values

OOD DETECTION: OOD-CLASSIFIER

- Problem: we have only in-distribution data
- Idea: Hallucinate new (ood) data by randomly sampling data points

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

- Study whether an explanation approach can be fooled ▸ Slack 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

FURTHER OBJECTIVES: PLAUSIBILITY

Plausibility:

- CEs should suggest realistic (i.e., plausible) alternatives
  $\leadsto$ Implausible: increase income *and* become unemployed
- CEs should adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\leadsto$ Avoid unrealistic combinations of feature values
- Estimating joint distribution is hard, especially for mixed feature spaces
  $\leadsto$ Common proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$

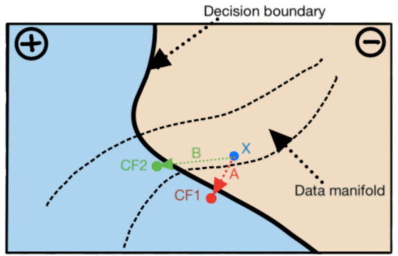
OOD DETECTION: CLUSTERING VIA DBSCAN

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

FURTHER OBJECTIVES: PLAUSIBILITY

Plausibility:

- CEs should suggest realistic (i.e., plausible) alternatives
  $\rightsquigarrow$ Implausible: increase income *and* become unemployed
- CEs should adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ Avoid unrealistic combinations of feature values
- Estimating joint distribution is hard, especially for mixed feature spaces
  $\rightsquigarrow$ Common proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$



Example from Verma et al. (2020)

- Input $\mathbf{x}$ originally classified as $\ominus$
- Two valid CEs in class $\oplus$: CF1 and CF2
- Path **A** (CF1) is shorter (but unrealistic)
- Path **B** (CF2) is longer but in data manifold

OOD DETECTION: CLUSTERING VIA DBSCAN

- DBSCAN is a data clustering algorithm Ester 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)

FURTHER OBJECTIVES

Plausibility term: Encourage counterfactuals close to observed data.

- Define $\mathbf{x}^{[1]}$ as the nearest neighbor of $\mathbf{x}'$ in the training set $\mathbf{X}$
- Use Gower distance between $\mathbf{x}'$ and $\mathbf{x}^{[1]}$ to define plausibility objective:

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

OOD DETECTION: CLUSTERING VIA DBSCAN

- DBSCAN is a data clustering algorithm ▸ Ester 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

FURTHER OBJECTIVES

Plausibility term: Encourage counterfactuals close to observed data.

- Define $\mathbf{x}^{[1]}$ as the nearest neighbor of $\mathbf{x}'$ in the training set $\mathbf{X}$
- Use Gower distance between $\mathbf{x}'$ and $\mathbf{x}^{[1]}$ to define plausibility objective:

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

Extended optimization: Add sparsity and plausibility terms to the objective

$$\arg\min_{\mathbf{x}'} \lambda_1 o_{\text{target}}(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_{\text{proximity}}(\mathbf{x}', \mathbf{x}) + \lambda_3 o_{\text{sparse}}(\mathbf{x}', \mathbf{x}) + \lambda_4 o_{\text{plausible}}(\mathbf{x}', \mathbf{X})$$

OOD DETECTION: CLUSTERING VIA DBSCAN

- DBSCAN is a data clustering algorithm ▸ Ester 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

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

OOD DETECTION: CLUSTERING VIA DBSCAN

- DBSCAN is a data clustering algorithm ▸ Ester 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

REMARKS: THE RASHOMON EFFECT

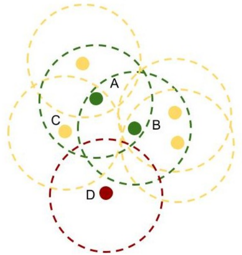
Issue (Rashomon effect):

- 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 $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one or few CEs (but: by which criterion should guide this choice?)

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

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

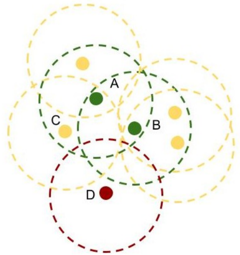
Possible solutions:

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

Note:

- Nonlinear models can produce diverse and inconsistent CEs
  $\rightsquigarrow$ suggest both increasing and decreasing credit duration (confusing for users)
- Handling this **Rashomon effect** remains an open problem in interpretable ML

OUT-OF-DISTRIBUTION DETECTION



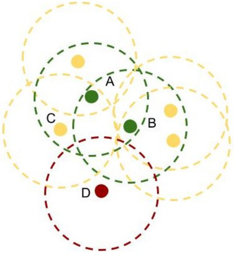
Example for DBSCAN, circles display ϵ-neighborhoods, $m = 4$

- 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

REMARKS: MODEL OR REAL-WORLD

- CEs explain model predictions, but may appear to explain the real-world users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- **Example:** CE suggests increasing age by 5 to receive a loan
  $\rightsquigarrow$ The applicant waits 5 years and reapplies

OUT-OF-DISTRIBUTION DETECTION



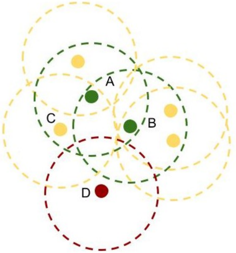
Example for DBSCAN, circles display ϵ-neighborhoods, $m = 4$

- 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

REMARKS: MODEL OR REAL-WORLD

- CEs explain model predictions, but may appear to explain the real-world users
  ⇝ Transfer of model explanations to explain real-world is generally not permitted
- **Example:** CE suggests increasing age by 5 to receive a loan
  ⇝ The applicant waits 5 years and reapplies
- **Problem:** Other features may change in the meantime (e.g., job status, income)
  ⇝ ▸ Karimi et al. (2020) propose CEs that respect causal structure
- **Model drift:** Bank's algorithm itself may change over time
  ⇝ Past CEs may become invalid

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