NORSK REGNESENTRAL
NORWEGIAN COMPUTING CENTER

MCCE: Monte Carlo sampling of realistic counterfactual explanations

Methods for Explainable Machine Learning in Health Care
Amsterdam, February 4th, 2026

Kjersti Aas
Norsk Regnesentral

Joint work with Annabelle Redelmeier, Martin Jullum and Anders Løland

Counterfactual explanations

Source: finbucket.com

- Which features of the ML-model should be altered to obtain a different decision?
- Example:
  - Peter applies for a loan and gets rejected by the ML-method the bank uses for credit scoring.
  - He wonders why his application is rejected and how he might improve his chances to get a loan.
  - This question may be formulated as a counterfactual:

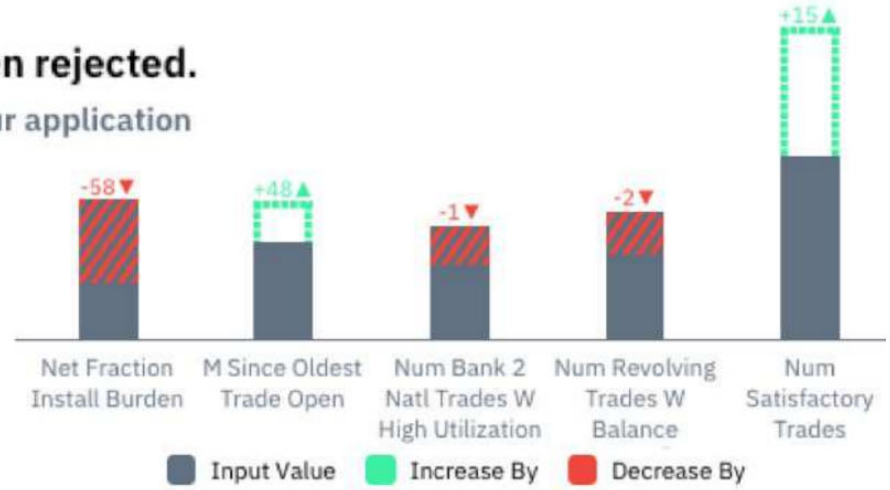
"What is the smallest change to Peter's features (e.g. income, age, number of credit cards) that would change the prediction from rejected to approved?"

Example

Sorry, your loan application has been rejected.

If instead you had the following values, your application would have been approved:

- MSinceOldestTradeOpen: **161**
- NumSatisfactoryTrades: **36**
- NetFractionInstallBurden: **38**
- NumRevolvingTradesWBalance: **4**
- NumBank2NatlTradesWHighUtilization: **2**



What is an appropriate explanation?

- **The explanation should be valid**
  - Desired decision
- **The explanation should be of low cost**
  - Few and small feature changes
- **The explanation should have actionable feature values**
  - Handle fixed features
- **The explanation should be on-manifold.**
  - Realistic range of, and correlation between, features

"I'd like to be a woman, please."

Source: cartoonstock.com

Methods

- The currently most used methods for computing counterfactual explanations are **optimization-based**.

$$\arg\min_{x'} \max_{\lambda} \lambda(f_w(x') - y')^2 + d(x_i, x')$$

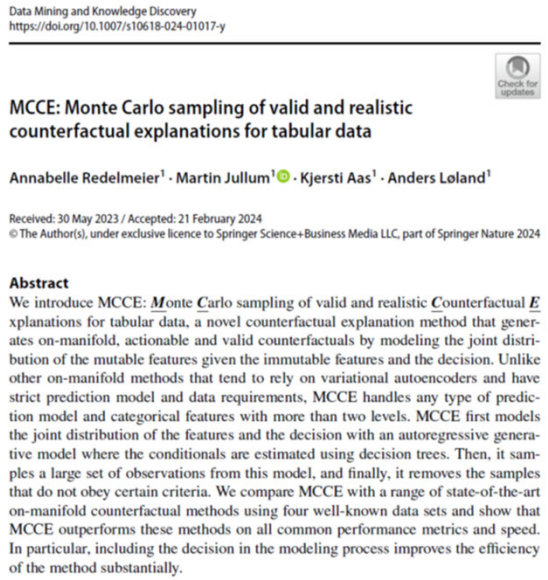
- Using these methods, the black-box model is usually assumed to be differentiable, meaning that they do not work for e.g. tree-based classifiers.
- Moreover, they tend to produce **unrealistic** counterfactuals, since they do not properly take the dependence between the variables into account.

Age: 17
Marital status: Widow
Profession: Professor

- NR has developed a method **MCCE**, which does not have these disadvantages.

MCCE

- **MCCE: M**onte **C**arlo sampling of valid and realistic **C**ounterfactual **E**xplanations for tabular data
- Three steps:
  1. Fits the joint distribution of the features and the decision with an **autoregressive generative model** where the conditionals are estimated using regression trees.
  2. Samples a large set of observations from this model
  3. Removes the samples that do not obey certain criteria.



Data Mining and Knowledge Discovery
https://doi.org/10.1007/s10618-024-01017-y

MCCE: Monte Carlo sampling of valid and realistic counterfactual explanations for tabular data

Annabelle Redelmeier[1] · Martin Jullum[1] · Kjersti Aas[1] · Anders Løland[1]

Received: 30 May 2023 / Accepted: 21 February 2024

Abstract
We introduce MCCE: *M*onte *C*arlo sampling of valid and realistic *C*ounterfactual *E*xplanations for tabular data, a novel counterfactual explanation method that generates on-manifold, actionable and valid counterfactuals by modeling the joint distribution of the mutable features given the immutable features and the decision. Unlike other on-manifold methods that tend to rely on variational autoencoders and have strict prediction model and data requirements, MCCE handles any type of prediction model and categorical features with more than two levels. MCCE first models the joint distribution of the features and the decision with an autoregressive generative model where the conditionals are estimated using decision trees. Then, it samples a large set of observations from this model, and finally, it removes the samples that do not obey certain criteria. We compare MCCE with a range of state-of-the-art on-manifold counterfactual methods using four well-known data sets and show that MCCE outperforms these methods on all common performance metrics and speed. In particular, including the decision in the modeling process improves the efficiency of the method substantially.

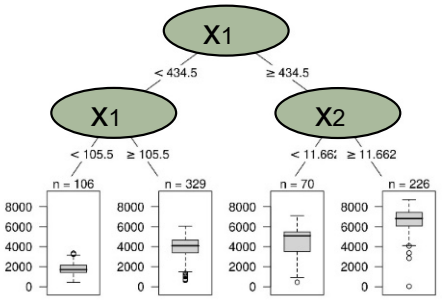
Step 1: Autoregressive generative model

- Decompose the distribution of the data ***X*** into products of conditional probability distributions* :

$$p(\boldsymbol{X}) = p(X_1) \times \prod_{i=2}^{q} p(X_i|X_1, \ldots, X_{i-1}).$$

- Fit a regression tree (CART) to each conditional distribution.

$x_3 \sim x_1 + x_2$.



From https://christophm.github.io/interpretable-ml-book/tree.html

Each end node contains the probability distribution for x_3 given certain values of x_1 and x_2.

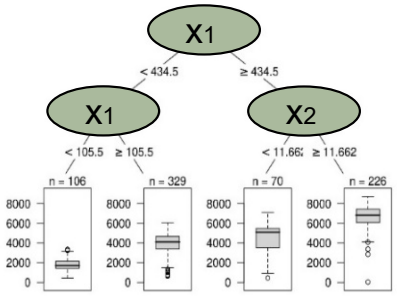
*Visit sequence corresponds to the order of the variables in the data set.

Step 2: Generation

- Step 2 consists of generating a K x q dimensional data set **D**, by sequentially sampling from the conditional distributions.

1. Generate K simulations $D_{1,1}, \ldots D_{K,1}$ by sampling with replacement from the values of variable X_1 in the training data set.
2. For $j = 2, \ldots q$ Variables
   - For $k = 1, \ldots K$ Samples
     - Find the end node in the tree T_j to which the sample $D_{k,1}, .., D_{k,j-1}$ belongs
     - Select $D_{k,j}$ by randomly sampling one observation from this node.

$X_3 \sim X_1 + X_2$.



From https://christophm.github.io/interpretable-ml-book/tree.html

Fixed features

- Features like age, sex and race are usually assumed to be fixed
- This can easily be taken into account by replacing step 1 in the generation procedure by
  - For $j = 1, \ldots p$
    - For $k = 1, \ldots K$
      - $D_{j,k} = x_j^*$

p is the number of fixed variables and x_j^* is the fixed value of variable j.

- and letting step 2 start at j=p+1 instead of j=2.

Step 3: Postprocessing

- Criteria 3 (actionable) and 4 (on-manifold) are already satisfied by construction.
- Further, most samples satisfy criterion 1 (valid), because we condition on the decision in addition to the fixed variables in the generation process.
- In the postprocessing step, criteria 2 (low cost) is satisfied as follows:
    1. Pick the rows in **D** for which the smallest number of features are changed.
    2. Of these rows, select the one with the smallest **Gower distance** to the factual.

Gower distance

- The Gower distance between the factual **x** and one row in **D** is computed as follows:

$$\text{Gower distance} = \frac{1}{p}\sum_{j=1}^{p}\delta_G(d_j, x_j) \in [0,1],$$

where

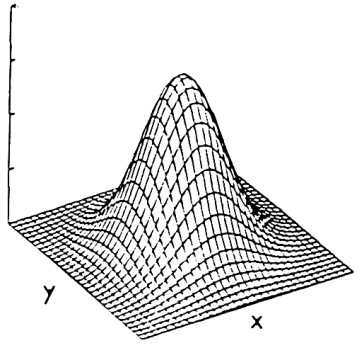
$$\delta_G(d_j, x_j) = \begin{cases} \frac{1}{R_j} \mid d_j - x_j \mid & \text{if } x_j \text{ is numerical,} \\ \mathbb{1}_{d_j \neq x_j} & \text{if } x_j \text{ is categorical,} \end{cases}$$

- R_j is the range of variable j.

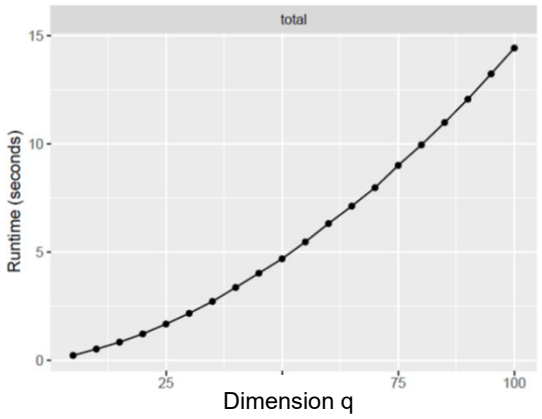
Computational complexity

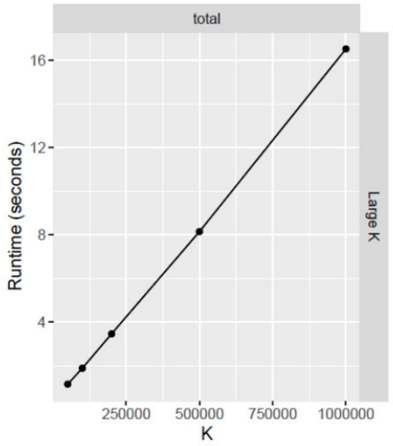
- Have simulated data from a q-dimensional Gaussian distribution.
- Assume a linear model
- Study computational time as a function of
  - Dimension q
  - Number of samples in generative model K
  - Number of training observations n_train
  - Number of test observations n_test

  keeping the other variables fixed



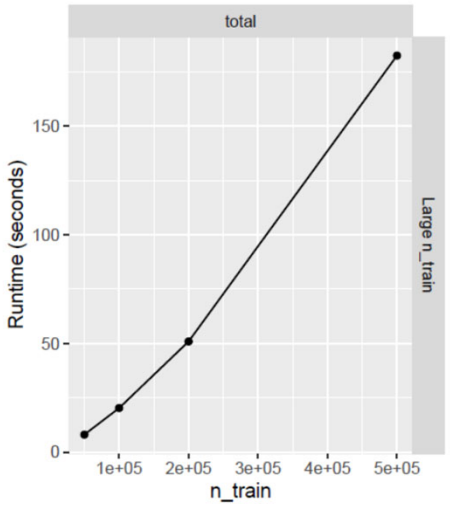
Computational complexity

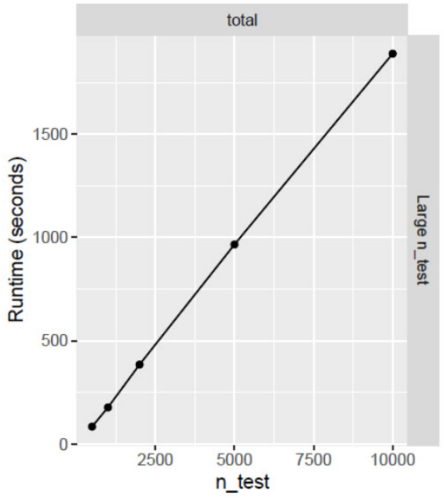




Quadratic increase in q

~linear increase in the other variables





Experiments

- Have compared MCCE to 6 other on-manifold counterfactual methods using 4 well-known data sets (“Adult”, ”GMC”, “German Credit” and “FICO”).
- All datasets have a binary response.
- Use a 3-layer ANN as prediction model*.
- Binarize categorical variables by partitioning them into the most frequent class and its counterpart**.
- Scaling continuous variables

*The competing methods do not handle tree-models.
** The competing methods do not handle categorical variables with more than two levels.

FICO

- Binary classification of customer being 90 days late with payment or not.
- 23 features, 21 continuous and 2 categorical
- *ExternalRiskEstimate* is set as immutable
- 10,459 observations

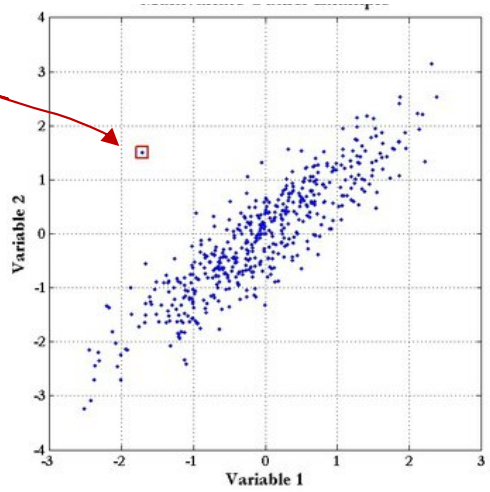
Data set: FICO, $n_{\text{test}} = 1000$, $K = 1000$						
Method	$L_0 \downarrow$	$L_1 \downarrow$	violation↓	success↑	$N_{\text{CE}} \uparrow$	t(s) all↓
C-CHVAE	21.99 (0.09)	2.08 (0.84)	**0.00 (0.00)**	**1.00**	**1000**	16.39
CEM-VAE	19.52 (2.46)	3.11 (0.87)	0.95 (0.21)	0.31	478	734.83
CLUE	23.00 (0.00)	2.64 (0.73)	1.00 (0.00)	**1.00**	4	3527.53
CRUDS	22.00 (0.00)	9.32 (2.20)	**0.00 (0.00)**	**1.00**	592	7639.12
FACE	19.11 (2.32)	3.01 (0.90)	0.98 (0.16)	**1.00**	**1000**	428.44
REViSE	21.71 (0.78)	**1.64 (0.54)**	**0.00 (0.00)**	**1.00**	926	5589.01
MCCE	**12.67 (1.91)**	1.95 (0.78)	**0.00 (0.00)**	**1.00**	**1000**	**15.47**

L_0 & L_1	(low cost)
Violation	(actionability)
Success	(validity)

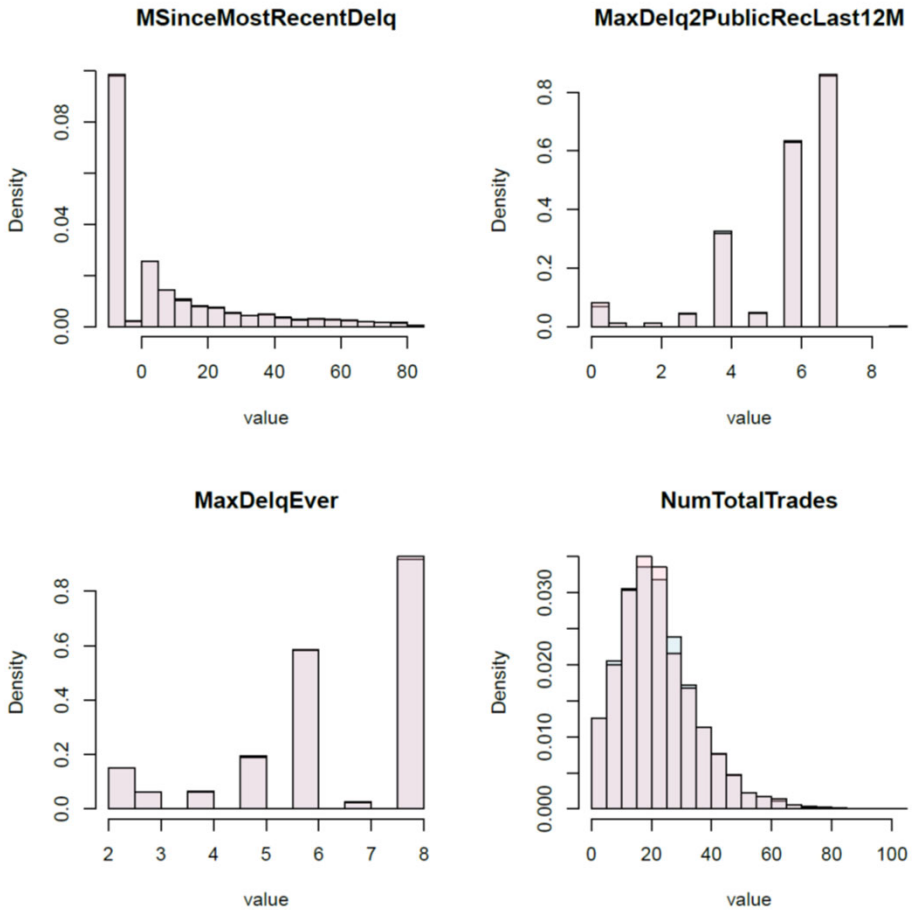
Data manifold closeness

- According to Guidotti (2022), a plausible counterfactual is “realistic” if it is “similar” to the known dataset and adheres to observed correlations among the features.
- We study the characteristics of the data generated in step 2 of the MCCE method.

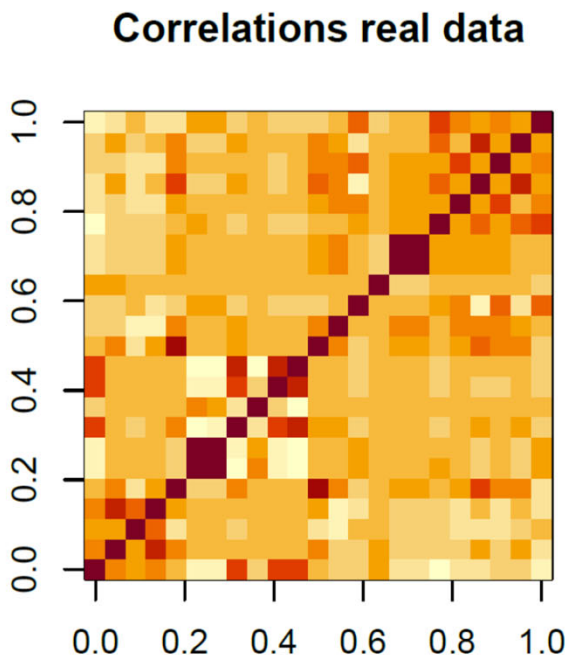
Not on data manifold

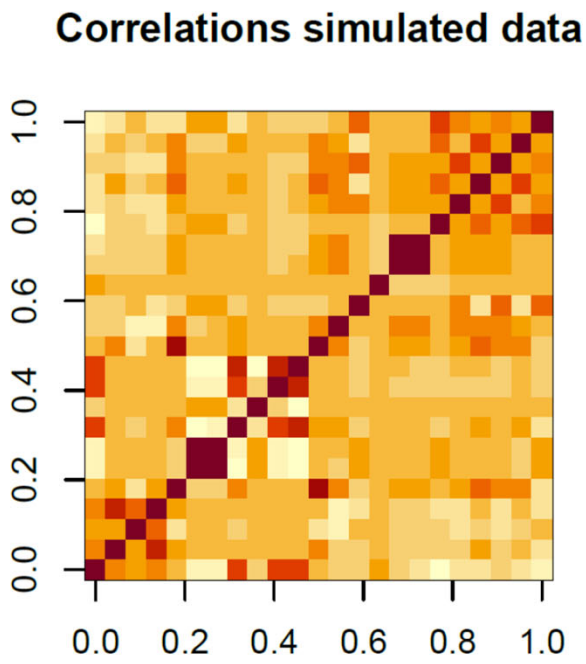


FICO: Marginal distributions



FICO: Correlations



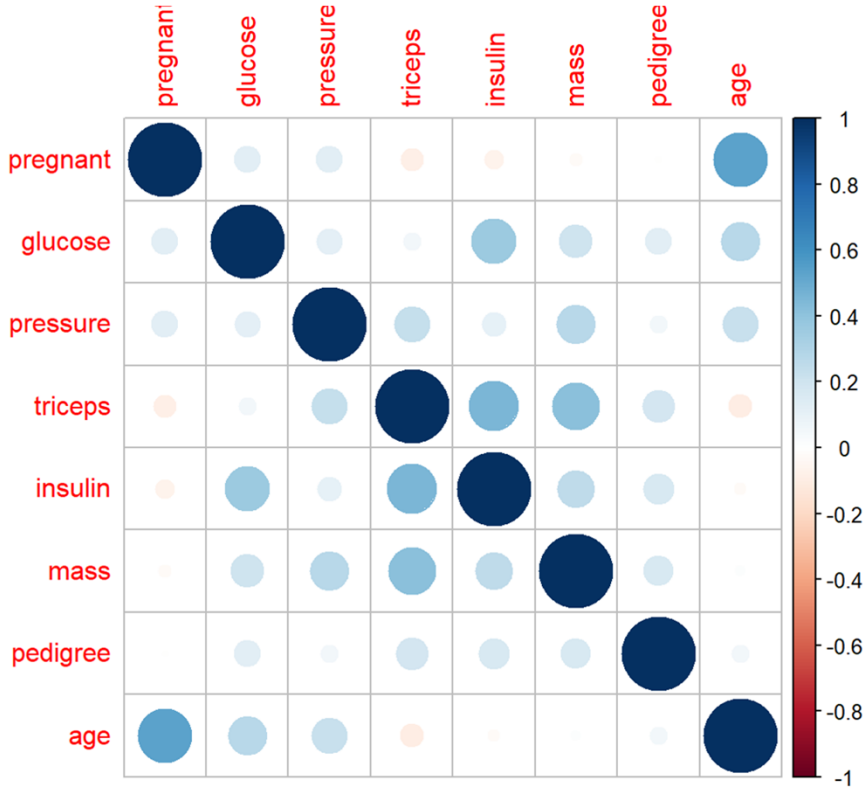


PIMA data set

- Predicting the onset of diabetes within 5 years in Pima Indians
- 768 observations and 8 explanatory variables
  - pregnant - Number of times pregnant
  - glucose - Plasma glucose concentration (glucose tolerance test)
  - pressure - Diastolic blood pressure (mm Hg)
  - triceps - Triceps skin fold thickness (mm)
  - insulin - 2-Hour serum insulin (mu U/ml)
  - mass - Body mass index (weight in kg/(height in m)^2)
  - pedigree - Genetic risk score used to estimate the likelihood of diabetes
  - age - Age (years)

Have used the version of the data set found here: https://www.openml.org/search?type=data&id=43483&sort=runs&status=active where missing values have been imputed. In the original data set triceps and insulin have 227 and 374 missing values, respectively.

Correlation matrix



ML-model

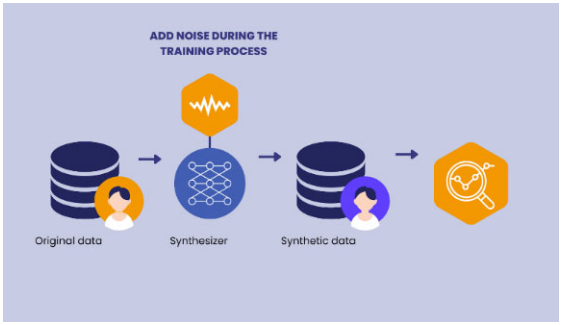
- Randomly divided the data set into a training (70%) and a test (30%) set.
- Used a Random forest classifier*
- AUC on test set is 0.88-0.90
- Study the observations with largest probability of onset of diabetes
  - Which variables should be changed, and how much, for the probability to be smaller than 0.2?
  - Keep age, pedigree and number of times pregnant fixed.

*ranger from the caret R-package trained with 10-fold cross-validation

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	prob
1:	12	114	82.0	18	102.5	30.0	0.528	58	0.166
	12	140	82.0	43	325.0	39.2	0.528	58	0.948
2:	7	92	84.0	31	102.5	39.9	0.331	41	0.130
	7	178	84.0	32	169.5	39.9	0.331	41	0.942
3:	8	91	68.0	19	48.0	30.1	0.615	60	0.070
	8	181	68.0	36	495.5	30.1	0.615	60	0.896
4:	3	96	78.0	17	45.0	27.8	0.970	31	0.148
	3	173	78.0	39	185.0	33.8	0.970	31	0.872
5:	8	90	106.0	17	77.0	37.6	0.165	43	0.142
	8	167	106.0	46	231.0	37.6	0.165	43	0.866

Differential privacy

- MCCE combines feature values in new ways.
- Hence, counterfactuals are almost certainly not observed in the training data.
- However, if a counterfactual is too close to another training observation than the one we want to explain, it may lead to a breach in privacy.
- This might be avoided by binning.
- In addition, differential privacy techniques might be used in the generation process.



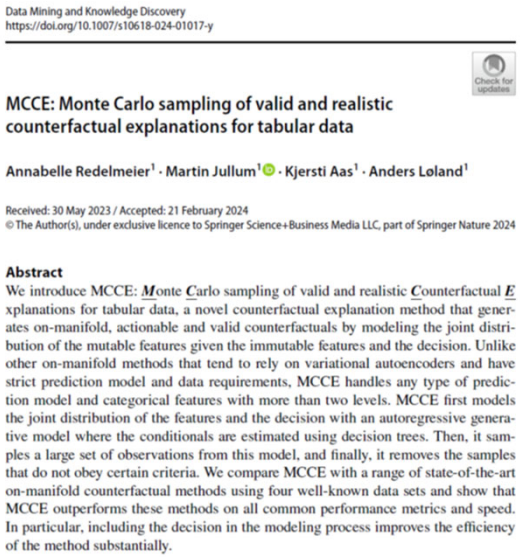
Source: Statice

MCCE: Summary

- Is quite fast
- Does not restrict the black-box model to be differentiable.
- Does properly handle fixed features.
- Does produce realistic counterfactuals
- Does handle categorical variables with more than two levels.

For R-code, see
https://github.com/NorskRegnesentral/mcceR

For Python code, see
https://github.com/NorskRegnesentral/mccepy



Data Mining and Knowledge Discovery
https://doi.org/10.1007/s10618-024-01017-y

MCCE: Monte Carlo sampling of valid and realistic counterfactual explanations for tabular data

Annabelle Redelmeier[1] · Martin Jullum[1] · Kjersti Aas[1] · Anders Løland[1]

Received: 30 May 2023 / Accepted: 21 February 2024

Abstract
We introduce MCCE: *M*onte *C*arlo sampling of valid and realistic *C*ounterfactual *E*xplanations for tabular data, a novel counterfactual explanation method that generates on-manifold, actionable and valid counterfactuals by modeling the joint distribution of the mutable features given the immutable features and the decision. Unlike other on-manifold methods that tend to rely on variational autoencoders and have strict prediction model and data requirements, MCCE handles any type of prediction model and categorical features with more than two levels. MCCE first models the joint distribution of the features and the decision with an autoregressive generative model where the conditionals are estimated using decision trees. Then, it samples a large set of observations from this model, and finally, it removes the samples that do not obey certain criteria. We compare MCCE with a range of state-of-the-art on-manifold counterfactual methods using four well-known data sets and show that MCCE outperforms these methods on all common performance metrics and speed. In particular, including the decision in the modeling process improves the efficiency of the method substantially.

Thank you for your attention!

kjersti@nr.no

www.nr.no