Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

An introduction to counterfactuals in explainable AI

Annabelle Redelmeier

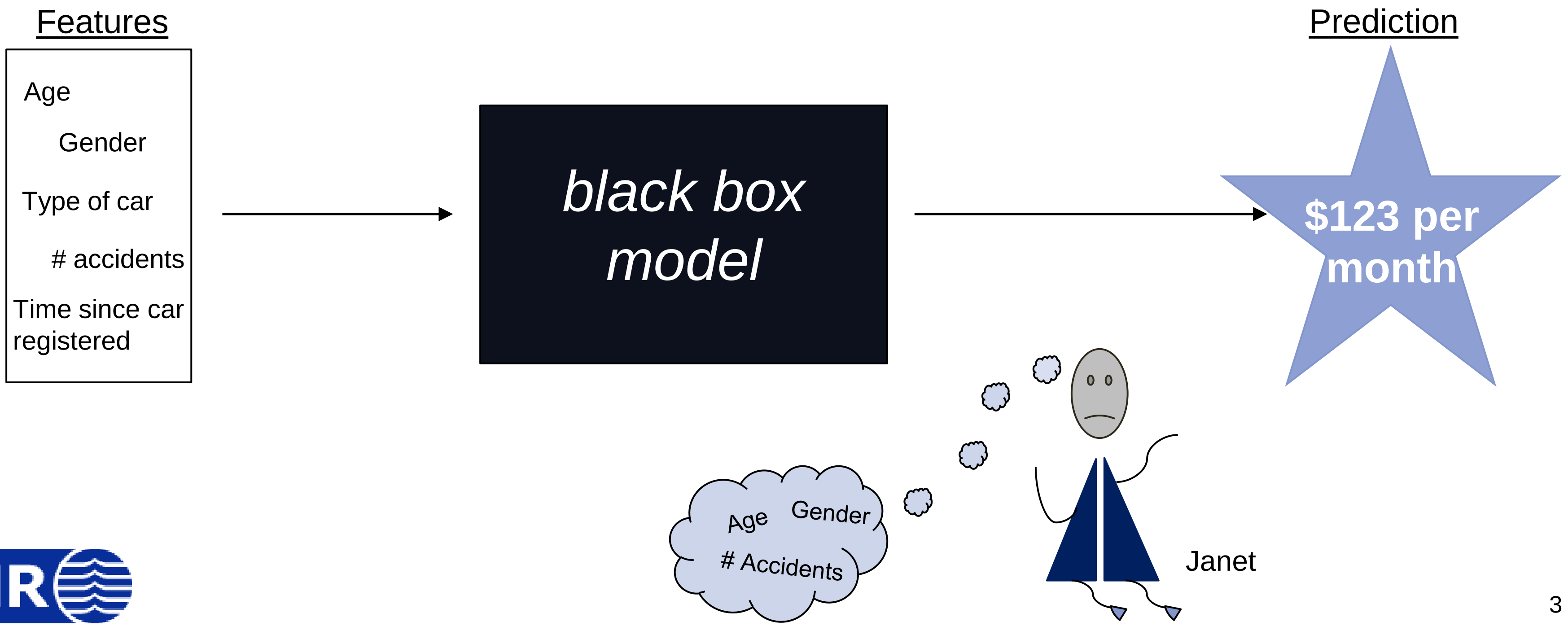
April 15th, 2021

Content

1. Explainable AI (what is it?).
2. Intuition behind counterfactuals (what/why/how).
3. Desirable counterfactual properties.
4. 2 different methods to calculate counterfactuals.
5. Summary.

Explanation problem

- Suppose we have a black-box model predicts the price of car insurance based on some features.
- How can we explain the prediction of a black-box model to a customer?



1. Explainable AI (XAI)

- For the last 10 years people have wanted to explain complex machine learning/statistical models.
- Also called “opening the black box”.

Explain a specific prediction

Used for any ML model

Specific to a model like xgboost or regression

	Model agnostic	**Model specific**
Local explanation	LIME, Shapley values, Explanation Vectors, Counterfactuals explanations, Saliency map	DeepExplain (understanding convolutional networks), tf-explain, RISE,
Global explanation	Partial dependence plots, Activation maximization, Model distillation,	Decision trees, Rule lists,

Understanding the whole logic of the model

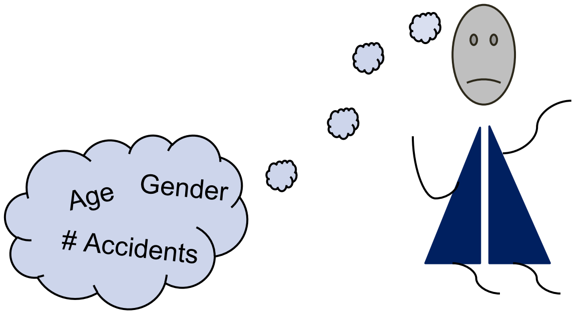
Google Trends Popularity Index (max value 100) of “Explainable AI” over last 5 years.

2. Counterfactuals in XAI (*what*)

- A counterfactual explanation takes the form:

“If Janet had *less car accidents in a year*, she would have *cheaper car insurance*”.

- If *A*, then *desired outcome.*
- Counterfactuals try to answer the question: *How could Janet’s features change to get a different prediction?*



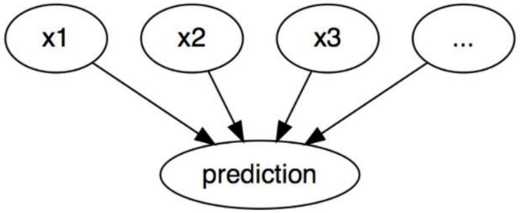
2. Counterfactuals in XAI (*why*)

- According to Wachter et al., 2017, explanations are useful to:
  1. Help the individual understand why a decision was reached;
  2. Provide grounds to contest the decision if the outcome is undesired;
  3. Understand what needs to change to receive a desired result in the future.
- They “enhance the autonomy of people subjected to automated decision”[1].
- They “help people recognize when they should contest decisions” [1].
- They are “human-friendly explanations” and “selective, meaning they usually focus on a small number of feature changes” [2].
- This is a type of explanation in *consequential decision making*[3].

[1]Barocas, Solon and Selbst, Andrew D and Raghavan, Manish (2020)
[2] Ch 6.1 Interpretable ML book by Dandl and Molnar
[3] Karimi, Amir-Hossein, et al. "Model-agnostic counterfactual explanations for consequential decisions." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

2. Counterfactuals in XAI (*how*)

- What would it take for Janet to have car insurance that costs \$100?



Ch 6.1 Interpretable ML book by Dandl and Molnar

Naïve method:

- Given a predictive model and individual:
  1. Pick a *different* predicted value: $\hat{Y} = 100$.
  2. Try every combination of features in the training data and keep the ones that give the chosen predicted value: $f(\hat{X}) = \hat{Y}$.

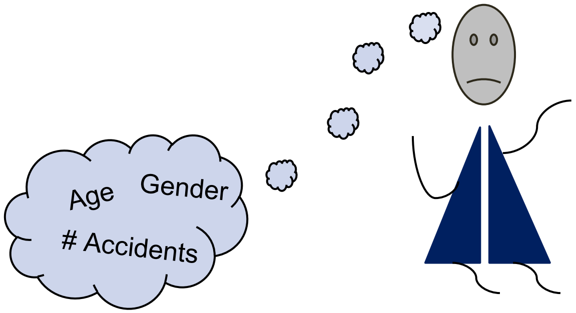
2. Counterfactuals in XAI (*example*)

- How do we define a “good” counterfactual explanation?

Current features	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Age = 55	Age = 70	Age = 65	Age = 55	Age = 55
Gender = F	Gender = F	Gender = F	Gender = F	Gender = F
Car = Volvo	Car = Volvo	Car = Subaru	Car = B&W	Car = Volvo
# accidents = 3	# accidents = 0	# accidents = 0	# accidents = 1	# accidents = 1
Time since car registered = 3	Time since car registered = 3	Time since car registered = 2	Time since car registered = 3	Time since car registered = 1

Closeness: a counterfactual that is *closer* to the starting feature vector is better.

How can we represent “closeness”?



Distance function

- “Closeness” is defined using a **distance function** between the original feature vector $\boldsymbol{x}'$ and the new counterfactual vector $\boldsymbol{x}$.
- One way to define the distance is:

$$d(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k \in F} (x_k - x'_k)^2$$

where F is the set of features.

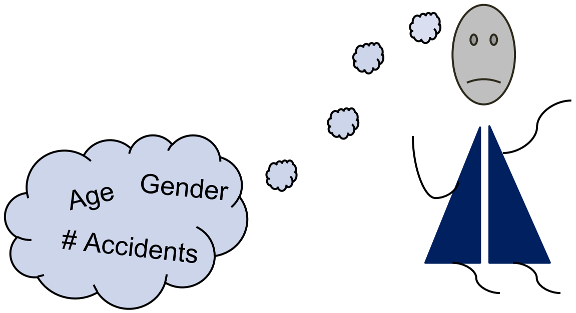
2. Counterfactuals in XAI (*example 2)*

- How do we define a “good” counterfactual explanation?

Current features	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Age = 55	Age = 55	Age = 55	Age = 55	Age = 58
Gender = F	Gender = F	Gender = F	Gender = F	Gender = F
Car = Volvo	Car = Volvo	Car = Volvo	Car = Volvo	Car = Volvo
# accidents = 3	# accidents = 1	# accidents = 1	# accidents = 0	# accidents = 1
Time since car registered = 3	Time since car registered = 1	Time since car registered = 2	Time since car registered = 3	Time since car registered = 3
Distance	**1.5**	**1.4**	**1.7**	**1.5**

Diversity: A series of counterfactuals that are different from each other are better.

Can we find a way to remove scenarios that are ***almost the same***?



2. Counterfactuals in XAI (*example 3)*

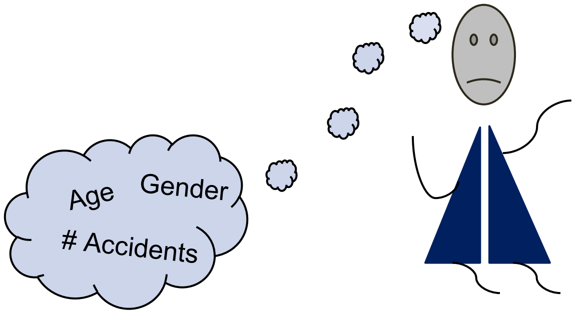
- How do we define a "good" counterfactual explanation?

Current features	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Age = 55	Age = 36	Age = 55	Age = 55	Age = 55
Gender = F	Gender = F	Gender = M	Gender = F	Gender = F
Car = Volvo	Car = Volvo	Car = Volvo	Car = B&W	Car = Volvo
# accidents = 3	# accidents = 3	# accidents = 3	# accidents = -1	# accidents = 1
Time since car registered = 3	Time since car registered = 3	Time since car registered = 3	Time since car registered = 3	Time since car registered = 0
Distance	**1**	**2.2**	**2.8**	**3**

Actionability: Counterfactuals that are impossible (decreasing age, changing gender) are useless.

Can we find a way to avoid scenarios that are "**unactionable**?"

How do we adjust our naïve approach to produce scenarios that are:
- **Easy to get to** ("close")?
- **Different** from each other?
- **Not "impossible"**?



3. Desirable properties of counterfactuals

Sparsity

Feasibility

Diversity

Actionability

Proximity

Causality

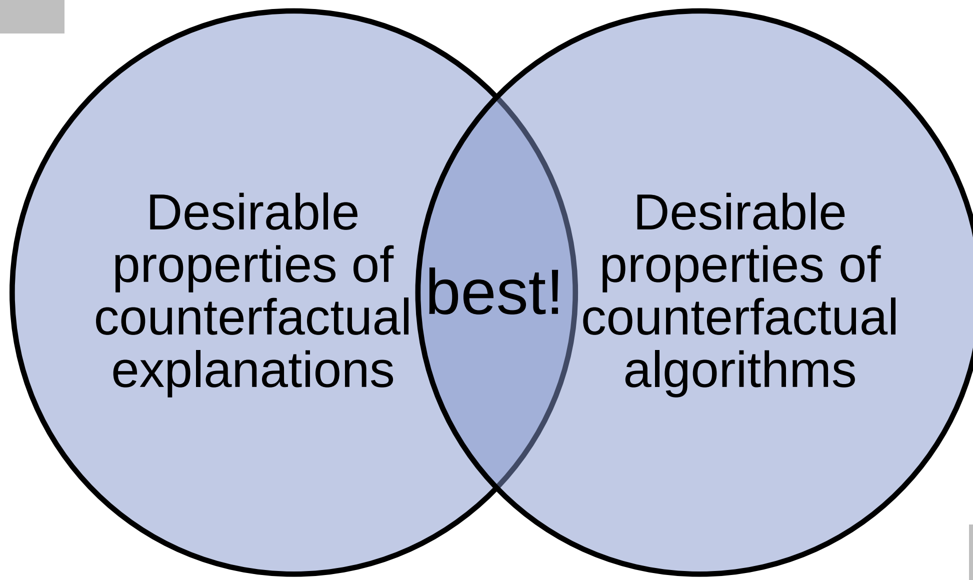
Flexibility in predictive model $f()$

Flexibility in distance function $d()$

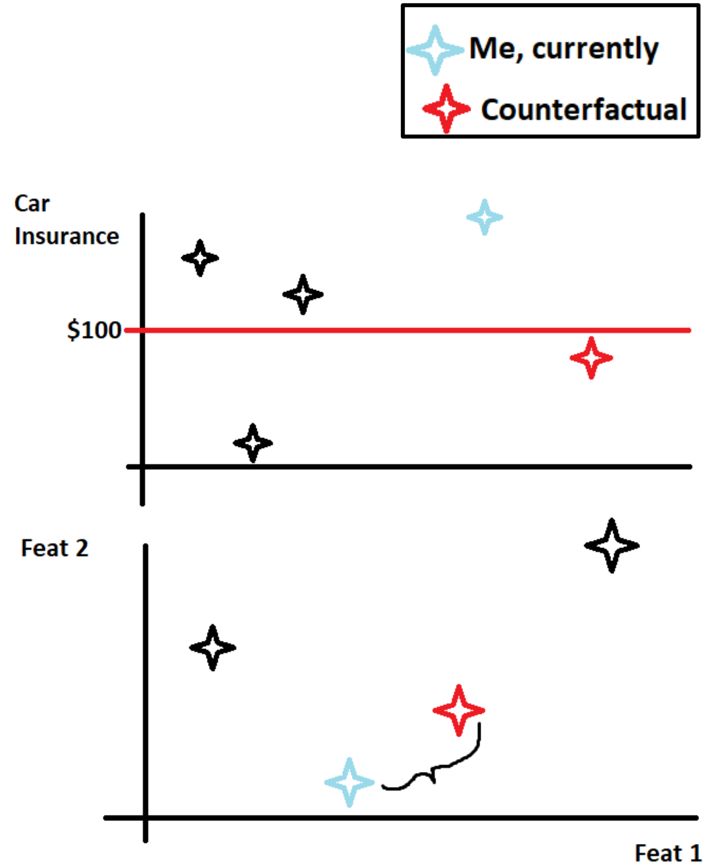
Flexibility in feature types

Speed

Convergence



Desirable properties of counterfactual explanations



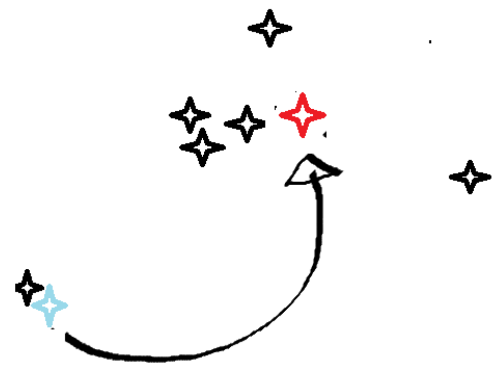
1. **Response-Proximity**: *Changing my features to these will give me a response that is close to my desired response.*

2. **Feature-Proximity**: *The counterfactual is close to my current feature vector.*

3. **Sparsity**: *The counterfactual changes only a few of my features.*

[1, 0, 1, 0.5] --> [1, 0, 2, 0.5]

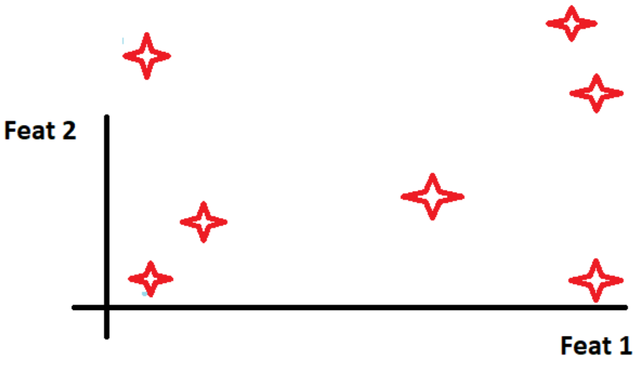
Desirable properties of counterfactual explanations



4. **Feasibility**. *The counterfactual lies in a high-density region in the feature space*.

5. **Causality.** *The counterfactual obeys causal constraints.*

6. **Diversity**. *The counterfactuals span a wide range of possibilities. This gives me many different choices/ways to change my prediction value.*



7. **Actionability/plausibility**. *If I had to, I could change all these features*.

History of counterfactuals in XAI

Mar 2014

Nov 2017

Explaining data-driven document classifications
Martens and Provost

Counterfactual explanations without opening the black box: Automated decisions and the GDPR
Wachter, Mittelstadt, and Russell

Pivotal paper 1

Paper: Counterfactual explanations without opening the black box: Automated decisions and the GDPR (Wachter et al., 2017)

- Suppose we have:
  - Training data and a model $f()$,
  - An individual x_i with response y,
  - A desired response y'.
- We wish to find a counterfactual x' as close to the original point x_i as possible such that f(x') = y'.
- How? We can set up a **loss function** that
  1. Minimizes f(x') - y' AND
  2. Minimizes the **distance** between x' and x.

$$L(x, x', \lambda) = \lambda(\hat{f}(x') - y')^2 + d(x_i, x')$$

A larger λ → we prefer counterfactuals that are very close to y'.
A smaller λ → we prefer counterfactuals that are very close to the original feature vector.

- Then we can **solve for the vector** x' that minimizes this loss using any optimization algorithm.

Loss = $\underbrace{[\text{distance to } y']}_{\text{Response-Proximity}}$ + $\underbrace{[\text{distance to } x]}_{\text{Feature-Proximity}}$

Pivotal paper 1

Paper: Counterfactual explanations without opening the black box: Automated decisions and the GDPR (Wachter et al., 2017)

- Loss function: $L(x, x', \lambda) = \lambda(\hat{f}(x') - y')^2 + d(x_i, x')$

- We still have to define a **distance function**. Options:

1. (Un-normalized) L$_1$: $d(x_i, x_k) = \sum_{k \in F} |x_{i,k} - x'_k|$.
2. (Un-normalized) L$_2$: $d(x_i, x_k) = \sum_{k \in F} \left(x_{i,k} - x'_k\right)^2$.

Other distances include the Gower distance, Mahalanobis distance…

- We can also **normalize** these differences by:

1. $std_{j \in P}(x_{j,k})$, for feature *k.*
2. $MAD_k = median_{i \in \{1,\ldots,n\}} (|X_{i,k} - median_{l \in \{1,\ldots,n\}}(X_{l,k})|)$, for feature *k.*

- How to choose? We'll see!

MAD is equivalent to the variance of a feature but takes the median rather than the mean.

Example: LSAT data set

- Predict a student's first year average grade based on:
  - *Race (0 = white, 1 = black),*
  - *GPA (from undergrad)*
  - *LSAT score.*

gpa	lsat	isblack	fya
3.1	39.0	0	-0.98
3.6	36.0	0	-0.10

- The average grade is *normalized* so that if it is > 0 → better than average, < 0 → worse than average.
- Counterfactual: What features should an individual change to get an average **test score of 0** (i.e average)?

$$d(x_i, x_k) = \sum_{k \in F} (x_{i,k} - x'_k)^2$$

Example: LSAT data set

Unnormalized L_2

Original Data				Counterfactuals			Counterfactual Hybrid		
score	GPA	LSAT	Race	GPA	LSAT	Race	GPA	LSAT	Race
0.17	3.1	39.0	0	3.0	39.0	0.3	1.5	38.4	0
0.54	3.7	48.0	0	3.5	47.9	0.9	-1.6	45.9	0
-0.77	3.3	28.0	1	3.5	39.8	0.4	3.4	33.4	0
-0.83	2.4	28.5	1	2.7	37.4	0.2	2.6	35.7	0

Two things to mention:

1. The counterfactuals for *Race* are nonsense decimal values.
   - To fix this, they set *Race* = 1 and solve the optimizer. Then they set *Race* = 0 and solve the optimizer again. They take the closest counterfactual as the result.
2. The counterfactuals always changes GPA more than LSAT.
   - This is due to the **chosen distance function** which prefers small changes spread uniformly across all variables. And because GPA varies less, this is changed more.

LSAT data set Try #2

$$d(x_i, x_k) = \sum_{k \in F} \frac{\left(x_{i,k} - x'_k\right)^2}{std_{j \epsilon P}(x_{j,k})}$$

Normalized L_2

Original Data				Counterfactual Hybrid		
score	GPA	LSAT	Race	GPA	LSAT	Race
0.17	3.1	39.0	0	3.0	34.0	0
0.54	3.7	48.0	0	3.5	33.1	0
-0.77	3.3	28.0	1	3.4	33.4	0
-0.83	2.4	28.5	1	2.6	35.7	0

How can we ensure that GPA changes **less** than LSAT? **Normalize the distance function!**

First try: use the standard deviation of the feature.

New problem: How can we make sure that the counterfactual explanation **doesn't change** every feature?

LSAT data set Try #3

$$d(x_i, x_k) = \sum_{k \in F} \frac{|x_{i,k} - x'_k|}{MAD_k}$$

Normalized L_1

Original Data				Counterfactual Hybrid		
score	GPA	LSAT	Race	GPA	LSAT	Race
0.17	3.1	39.0	0	3.1	34.0	0
0.54	3.7	48.0	0	3.7	32.4	0
-0.77	3.3	28.0	1	3.3	33.5	0
-0.83	2.4	28.5	1	2.4	35.8	0

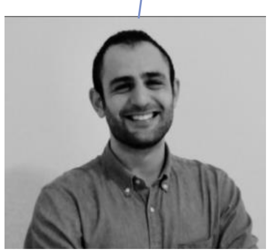
It turns out that using the L_1 norm (rather than the L_2 norm) normalized by the MAD **makes sparser counterfactuals**!

Notes:

- Fixing the discrete problem is time consuming (imagine if race had 100 levels!)
- They do not ensure that x' is an *actionable* data point (changing race?!)
- This algorithm solves for exactly one counterfactual.

History of counterfactuals in XAI

- **Mar 2014** — Explaining data-driven document classifications. **Martens and Provost**
- **Nov 2017** — Counterfactual explanations without opening the black box: Automated decisions and the GDPR. **Wachter, Mittelstadt, and Russell**
- **Jan 2019** — Actionable Recourse in Linear Classification. **Ustun, Spangher, and Liu**
- **Jan 2019** — Efficient Search for Diverse Coherent Explanations. **Russell**
- **May 2019** — Model-Agnostic Counterfactual Explanations for Consequential Decisions. **Karimi, Barthe, Balle, Valera**
- **Dec 2019** — Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations (DiCE). **Mothilal, Sharma, Tan**



Paper #2

Paper: Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations (Mothilal et al., 2019)

- Extends the Wachter et al. paper to handle ***feasibility*** and ***diversity*** among the counterfactuals presented.

> **Feasibility**: Feature-Proximity + Actionability + Sparsity + Causality
>
> **Diversity**: Counterfactuals are all different from each other.

- Feature-Proximity: through proximity constraint.
- Actionability + sparsity: through postprocessing.
- Diversity: through point process.

Note: They have a different definition of "feasibility" than the one defined on slide 14.

Paper #2

Paper: Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations (Mothilal et al., 2019)

- We begin with the same loss function as before:

$$\boxed{L(x, x', \lambda) = \lambda(\hat{f}(x') - y')^2 + d(x_i, x')}$$

- But now we want to **generate k counterfactuals** $\{c_1, \dots, c_k\}$. We can add a sum term to the loss:

$$\boxed{L(c_1, \dots, c_k, x', \lambda) = \frac{1}{k}\sum_{i=1}^{k} \text{yloss}(\hat{f}(c_i), y) + \frac{\lambda}{k}\sum_{i=1}^{k} d(c_i, x')}$$

- But remember out example:

Current features	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Age = 55	Age = 55	Age = 55	Age = 55	Age = 55
Gender = F	Gender = F	Gender = F	Gender = F	Gender = F
Car = Volvo	Car = Volvo	Car = Volvo	Car = Volvo	Car = Volvo
# accidents = 3	# accidents = 1	# accidents = 1	# accidents = 0	# accidents = 1
Time since car registered = 3	Time since car registered = 1	Time since car registered = 2	Time since car registered = 3	Time since car registered = 3
Distance	**1.5**	**1.5**	**1.7**	**1.5**

Counterfactuals are **not useful** if they are all the same!

Paper #2

Paper: Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations (Mothilal et al., 2019)

- To make sure we have **diversity**, we add an additional term to our loss function that increases our loss if our counterfactuals are **close together**.
- How do we measure **closeness? ** Our favourite **distance** function!
- If c_i and c_j are two counterfactuals that are **close** (we want to penalize our loss),
  - $dist(c_i, c_j)$ will be **small**,
  - So, $1/dist(c_i, c_j)$ will be **large.**
- Because we have *k* counterfactuals, we use the matrix **K** where

$$\boxed{\mathbf{K}_{i,j} = \frac{1}{1 + dist(\mathbf{c}_i, \mathbf{c}_j)}}$$

- And it turns out that the **determinant** of a symmetric matrix with **large** values in [0,1] will be **small** (close to 0).
- To make our loss function **bigger** if the determinant is **small**, we **subtract** det(**K**):

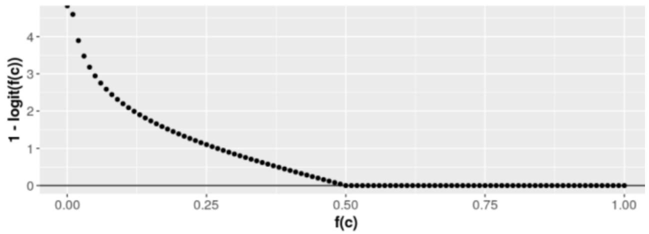
$$\boxed{L(c_1, \ldots, c_k, x', \lambda_1, \lambda_2) = \frac{1}{k}\sum_{i=1}^{k} \text{yloss}(\hat{f}(\boldsymbol{c}_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} d(c_i, x') - \lambda_2 \det \boldsymbol{K}}$$

$$\boxed{L(c_1,\ldots,c_k,x',\lambda_1,\lambda_2)=\frac{1}{k}\sum_{i=1}^{k}\text{yloss}(\hat{f}(\boldsymbol{c}_i),y)+\frac{\lambda_1}{k}\sum_{i=1}^{k}d(c_i,x')-\lambda_2\det\boldsymbol{K}}$$

Paper: Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations (Mothilal et al., 2019)

- Some additional notes:
  - They define yloss:

$$\boxed{\max(0,-1*logit(f(c)))}$$



Here the response is {0,1} and pr(x) > 0.5 → response of 1.

  - The **distance** function is the same as Wachter for continuous features and for categorical:

$$\boxed{\frac{1}{dcat}\sum_{p=1}^{dcat} I(c^P \neq x^P)}$$

- We summarize:

$$\text{Loss} = \underbrace{[\text{distance to } y']}_{\text{Response-Proximity}} + \underbrace{[\text{distance to } x]}_{\text{Feature-Proximity}} + \underbrace{[\text{diversity between chosen counterfactuals}]}_{\text{Diversity}}$$

Conclusion & Summary

- Counterfactual explanation is a straightforward method to provide explanations in terms of “what-if scenarios”.
- There are lots of ways to calculate the scenarios/counterfactuals.
- Some counterfactuals are “better” than others:
  - Response-proximity
  - Feature-proximity
  - Sparse
  - Feasible
  - Obey causal constraints
  - Actionable.

Next presentation

- We will go into depth of three advanced counterfactual methods: (probably)
  1. Dandl, Susanne, et al. "Multi-objective counterfactual explanations." *International Conference on Parallel Problem Solving from Nature*. Springer, Cham, 2020.
  2. Ustun, Berk and Spangher, Alexander and Liu, Yang (2019)Actionable recourse in linear classificationProceedings of the Conference on Fairness, Accountability, and Transparency
  3. Poyiadzi, Rafael, et al. "FACE: feasible and actionable counterfactual explanations." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020.
  4. Joshi, Shalmali and Koyejo, Oluwasanmi and Vijitbenjaronk, Warut and Kim, Been and Ghosh, Joydeep(2019) Towards realistic individual recourse and actionable explanations in black-box decision making systems arXiv preprint arXiv:1907.09615

Suggestions?

List of papers metioned

- Wachter, Sandra and Mittelstadt, Brent and Russell, Chris (2017)Counterfactual explanations without opening the black box: Automated decisions and the GDPRHarv. JL & Tech.31, 841
- Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020.
- Dandl, Susanne and Molnar, Christoph and Binder, Martin and Bischl, Bernd (2020) Multi-objective counterfactual explanations International Conference on Parallel Problem Solving from Nature
- Barocas, Solon and Selbst, Andrew D and Raghavan, Manish (2020)
- Ch 6.1 Interpretable ML book by Dandl and Molnar
- Karimi, Amir-Hossein, et al. "Model-agnostic counterfactual explanations for consequential decisions." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.