Data Manifold Closeness

- CF should be realistic and not anomalous to the training data points.
- CF should be near the training data
- CF should adhere to the observed correlations among the features
- Loss function can include another penalty for adhering to the data manifold.

$$\arg \min_{x' \in \mathcal{A}} \max_{\lambda} \underbrace{\lambda(f(x') - y')^2}_{\text{validity}} + d(x, x') + \underbrace{g(x - x')}_{\text{sparsity}} + \underbrace{l(x, x')}_{\text{data manifold closeness}}$$

Causality

- Changing one feature in the real world affects other features.
- Educational Degree/ age increase
- CF should maintain any non causal relations between the features.

Amortised Inference

- Property of the algorithm, but not part of the objective function.
- Learning to predict a CFE allows to predict the CF for any new x
- Examples
  - Generative technique for amortised inference of CFEs.
  - Use RL to predict a CF.

Black-box access

- Property of the algorithm, but not part of the objective function.
- Can work with proprietary ML models.
- Requires to use only the "predict" function
- Genetic algorithm-based
- RL-based

Model Agnosticity

- Property of the algorithm, but not part of the objective function.
- Such an approach can work with different kinds of ML models.
- An approach that requires black-box access to the ML model is a model-agnostic approach.

Algorithmic Recourse

- Algorithmic recourse takes into account the actionability of the prescribed changes.
- The difference between CFE generation and algorithmic recourse has now blurred.

Inverse Classification

- Aims to perturb an input in a meaningful way in order to classify it into its desired class.
- Prescribes the actions to be taken in order to get the desired classification.
- Has the same goals as CFEs.

Contrastive Explanation

- An input x is classified as y because features $f_1, f_2, \ldots, f_k$ are present and $f_n, \ldots, f_r$ are absent.
- Pertinent Positives: minimally sufficient for a classification
- Pertinent Negatives: absence is necessary for final classification
- CFEs are related to Contrastive explanations
- Need to solve an optimisation problem to find the minimum perturbations needed to maintain the same class label or change it.

Adversarial Learning (AL)

- Closely related, but not interchangeable with CFEs.
- AL aims to generates the least amount of change in a given input to classify to classify it differently (resulting in a highly confident misclassification).
- Applied to images, the goal of AL is an imperceptible change in the input image.
  - This is often at odds with the CFE's goal of sparsity and parsimony. (Single pixel attack is an exception)
  - AL does not consider data manifold, actionability, causality.
  - CFEs produce plausible or semantically meaningful data points.

Properties of Counterfactual Algorithms

Model Access

CF generation algorithm

- Access to complete Model internals — ML prediction
- Access to gradients (Used by majority)
- Access to only the Prediction function — (Black box)

- Access to complete models: required if the algorithm uses a solver-based method like mixed integer programming or they operate on decision trees.
- Access to Gradients: is needed to solve the optimisation objective, but this is restricted to differentiable models.
- Black-box approaches use gradient free optimisation algorithms such as Nelder-Mead, growing spheres, FISTA and ASP or genetic algorithms. $f(x)$

- Some approaches do not cast the goal into an optimisation problem and solve it using heuristics.
  - FACE that uses Dijkstra's algorithm
  - Graph traversal to find the closest CFEs
- Approaches that propose new training routines
  - Adversarial loss during the training of an ML model to have a higher probability of having a recourse for the training datapoints. After training, any CFE generation method can be used.
  - CounterNet can predict the class and generate the CFE of a datapoint.

Model Agnostic

- Algorithms based on solvers require linear or piece wise linear models.
- Algorithms that are model specific only work for those models, like tree ensembles.
- Black-box methods have no restriction on the underlying models.

Optimisation Amortisation

- For each input data point, solve the optimisation problem for each counterfactual that was generated
- Some can generate several counterfactuals for each data point. (With some metric of diversity)
  - A trained VAE can generate multiple counterfactuals
  - RL
  - CGAN with umbrella sampling
  - GAN
  - Prediction model.
  - Partially evaluate the classifier for the static features, speeding up CFE generation.

Counterfactual Attributes

- Sparsity
- Data manifold adherence
- Causality
- Sparsity:
  - Methods using solvers explicitly constraint sparsity.
  - Black Box methods constrain L0 norm.
  - Gradient based methods constrain L1 norm.
  - Change fixed number of features/ flip minimum split nodes/ post hoc induction of sparsity in the generated CFE using greedy method.

- Data Manifold adherence
  - Training VAE — constraining distance to k nearest training points — kernel density estimator to estimate the PDF of data manifold — cycle consistency loss of the GAN — using GMM to estimate the density — using feature correlations
- Causality
  - Using the causal graph, consider the causal relations between features when generating the counterfactuals.

Counterfactual (CF) optimisation problem attributes

- Considering feature actionability: classify features into immutable, mutable and actionable types. Credit score cannot be changed directly.
- Handling categorical features in gradient-based methods can be complicated. — Gumble softmax to relax categorical features into continuous ones. — generally a separate distance function is used for categorical features.

Bias in Machine Learning

Sources of Bias

- Bias leads to unfairness
- Bias can exist in many shapes and forms.
- Bias can exist in data origin, collection, processing

Bias can be in

- Data
- Data Attributes
- Data Collection
- User Interface
- User interaction
- Modelling assumptions

Types of Bias

- Historical Bias
  This is the already existing bias.
  Perfect sampling and feature selection cannot overcome this bias
- Representation Bias
  Happens in the way we define and sample from the population.

Imagenet

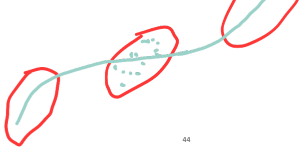
- Measurement Bias recedivism
  - arises from the way we choose, utilize or measure a particular feature
  - happens due to the way the minority groups are assessed and controlled.
- Evaluation Bias: Happens during model evaluation, the benchmark datasets are themselves biased

- Aggregation Bias
  Making general assumptions about different subgroups in a population can result in aggregation bias.
  (in other words, these are false assumptions about a population).
- Population Bias
  Arises due to the user population represented in the dataset or the platform
  and the target population.

- Simpson's Paradox – occurs when analyzing heterogeneous data composed of subgroups.
  (trend/association/characteristic).

BBG	Control Group (No Drug) 60		Treatment Group (Took Drug) 60	
	Heart attack	*No heart attack*	*Heart attack*	*No heart attack*
Female	5% 1	20 19	7.5% 3	40 37
Male	30% 12	40 28	40% 8	20 12
Total	13 21.6%	47	11 18.3%	49

- Longitudinal data fallacy
  Longitudinal data tracks the same sample at different points of time.
  The cross-sectional snapshot of a population can contain different cohorts.
- Sampling Bias/ Selection Bias
  arises due to non-random sampling of sub-groups.
  Due to the differences in sampling, the trends estimated from one subgroup may not generalize to data collected from a new population.



44

- Behavioural Bias
  Arises due to different user behaviour across [platforms, contexts, datasets]
- Content Production Bias
  difference in the use of language across different gender and age groups.
  [structural, lexical, semantic, syntactic differences]

45

- Linking Bias
  network attributes/network measures [obtained from user connections, activities, or interactions] differ and misrepresent the true behaviour of the user.
  -- low degree nodes
  -- social link patterns may not depict the actual user interaction.
- Instead of judging based on network links, judge based on content and behaviour

46

- Temporal bias:
  arises when the population itself changes with time, or the behaviour of the population changes with time.
- Popularity bias: Items that are more popular tend to be exposed more.
  But, popularity metrics are subject to manipulation.
- Algorithmic Bias:
  Bias is not in the input data
- User interaction bias:
  triggered by user interface
  triggered by user's self selected behaviour
- Social Bias:
  when our judgement gets biased by other people's actions or content.

47

- Emergent Bias
  - Generally a user interface is designed keeping in mind some prospective users, their cultural values, population, societal values.
  - But if the real users interacting with the interface are different, then they would tend to reflect on the emergent bias.

    As a result of use and interaction with real users.

48

- Self Selection Bias:
  subjects of the research select themselves.
  survey takers/ successful students
- Omitted Variable Bias
  one or more important variables are left out of the model.
- Cause-effect Bias
  Interpreting correlation as a cause-effect
- Observer Bias: researchers influence the participants.
- Funding Bias: Reporting of biased results.

49

Feedback Loop

- Model makes decisions that produce outcomes.
- These outcomes affect future data for subsequent training`