

# 可解释机器学习：反事实解释

汇报人：王亦琛

2022年4月8日





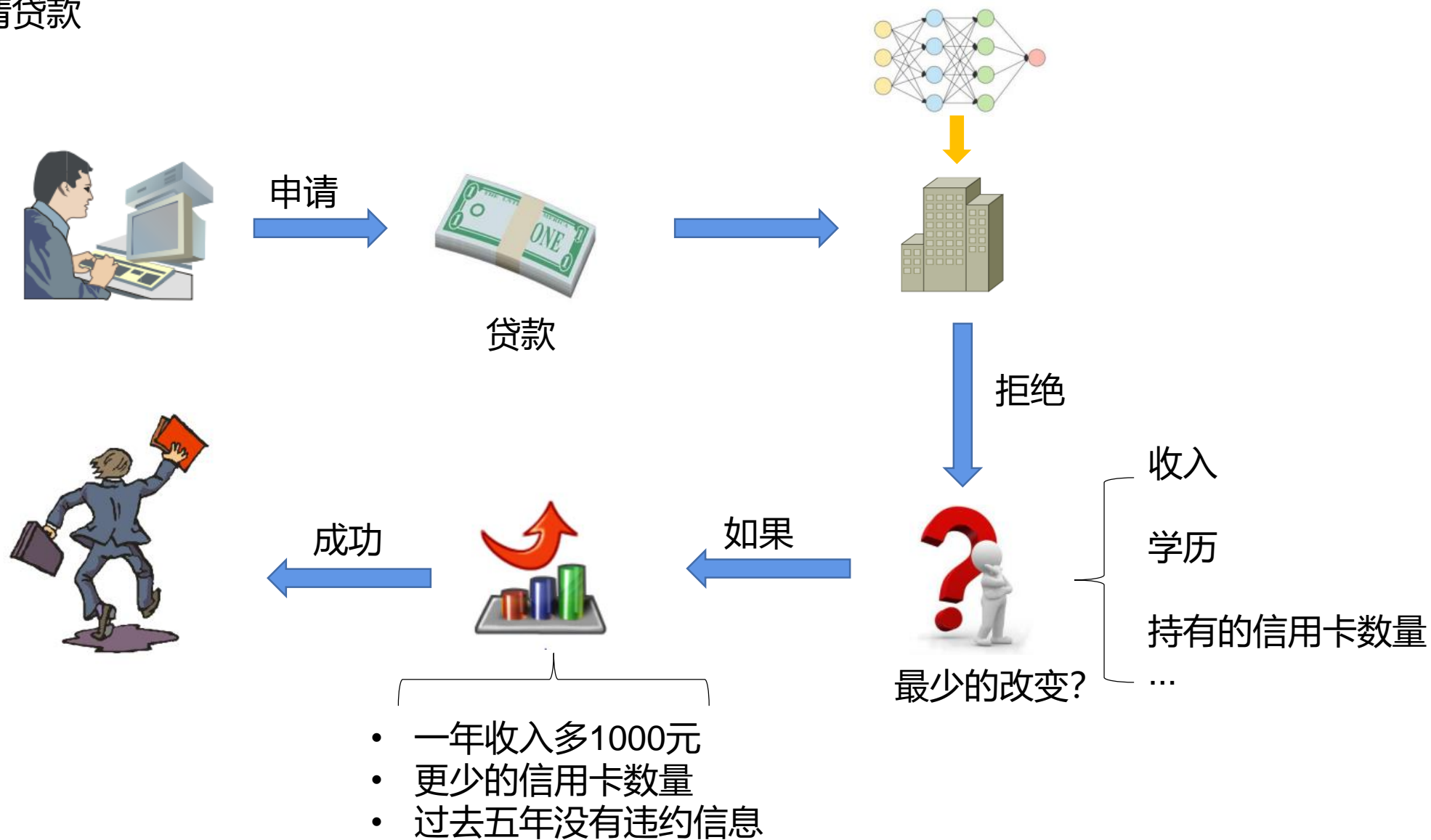
- ◆ **什么是反事实解释**
- ◆ **反事实解释的评价指标**
- ◆ **反事实解释方法分类**
- ◆ **图神经网络中的反事实解释**



- ◆ **什么是反事实解释**
- ◆ 反事实解释的评价指标
- ◆ 反事实解释方法分类
- ◆ 图神经网络中的反事实解释

# 反事实解释

例子：申请贷款



# 反事实解释

在上述的例子中，Bob想要进行贷款的行为被机器学习模型拒绝

Bob想知道：

- 为什么他的申请被拒绝了？
- 如何改进自己的行为能够使得将来贷款被批准？

假设Bob的特征向量表示为  $Bob = \{Income, CreditScore, Education, Age, \dots, etc\}$

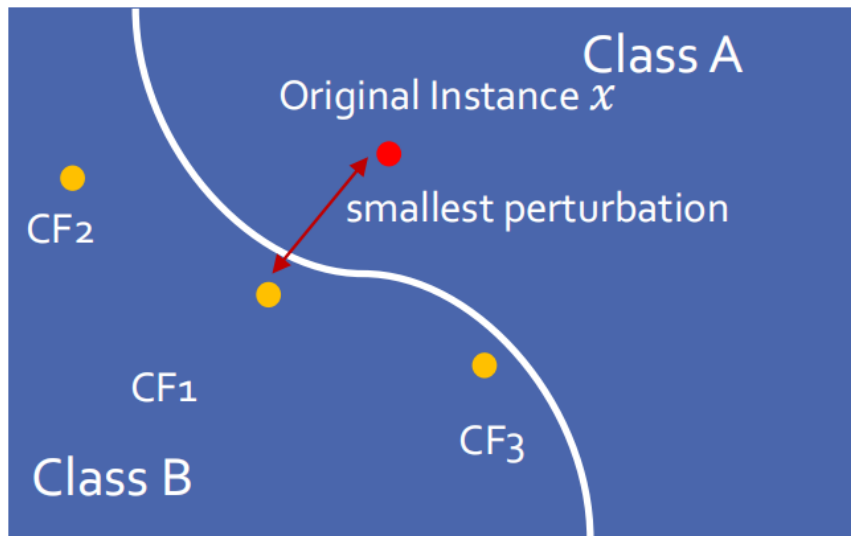
普通的机器学习解释方法  
反事实解释方法

信用分数过低，  
申请不通过  
提高1000元收入  
获得一个硕士学位

# 反事实解释

反事实解释：对输入特征的最小扰动，从而使预测结果变成另一个预定的输出。

Original Instance: 我们想要  
解释的实例  
CFx: 反事实解释



通过创建反事实实例，我们可以了解模型如何做出决策，以及如何解释单个实例

事实解释: Why

反事实解释: What if and Why



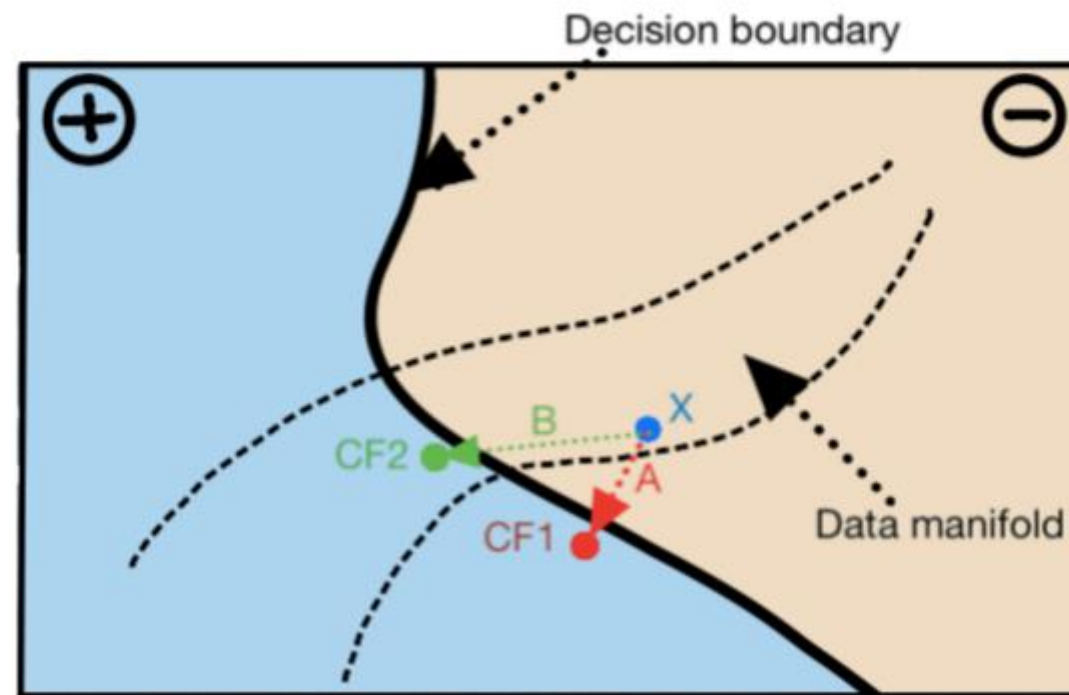
- ◆ 什么是反事实解释
- ◆ **反事实解释的评价指标**
- ◆ 反事实解释方法分类
- ◆ 图神经网络中的反事实解释

# 反事实解释的评价指标

名词：

Decision Boundary：决策边界

Data manifold：数据流形（数据分布的空间）



$$\arg \min_{x' \in \mathcal{A}} \max_{\lambda} \underbrace{\lambda(f(x') - y')^2}_{\text{有效性}} + \underbrace{d(x, x')}_{\text{接近度}} + \underbrace{g(x' - x)}_{\text{稀疏性}} + \underbrace{l(x'; \mathcal{X})}_{\text{数据流形相似性}}$$

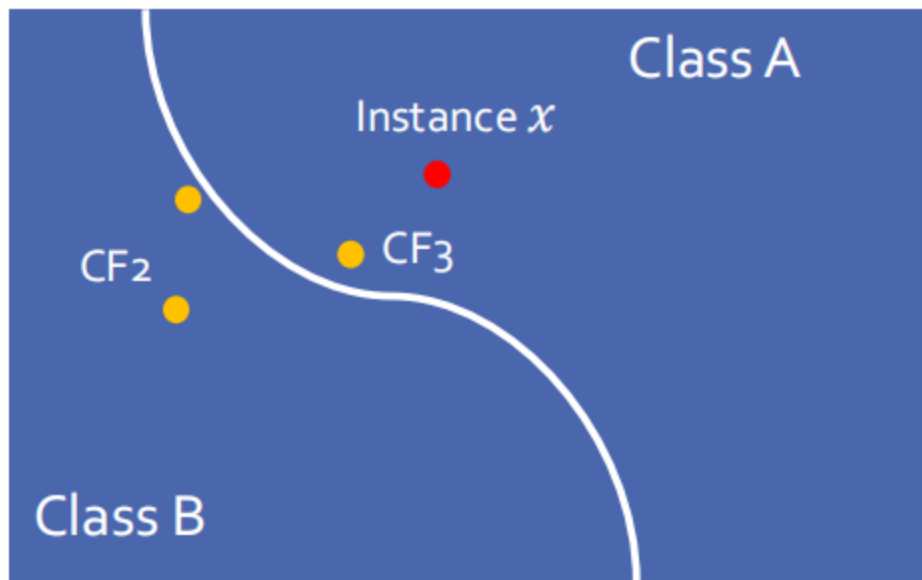


# ① Validity: 有效性

有效性：有效性衡量实际上具有所需类别标签的反事实

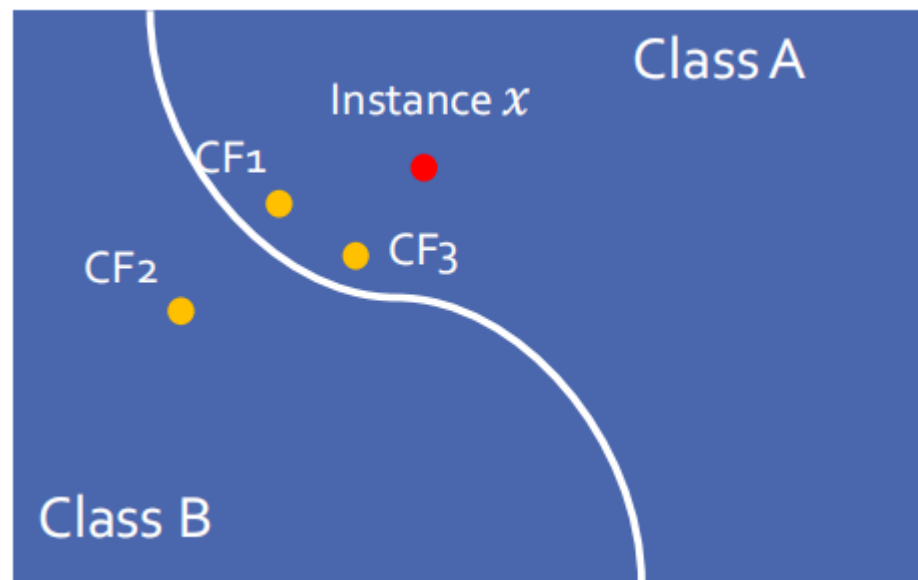
$$\text{Validity: } \frac{\sum_i I(f(CF'_i) == y'_i)}{K}$$

方法1:



Validity=2/3

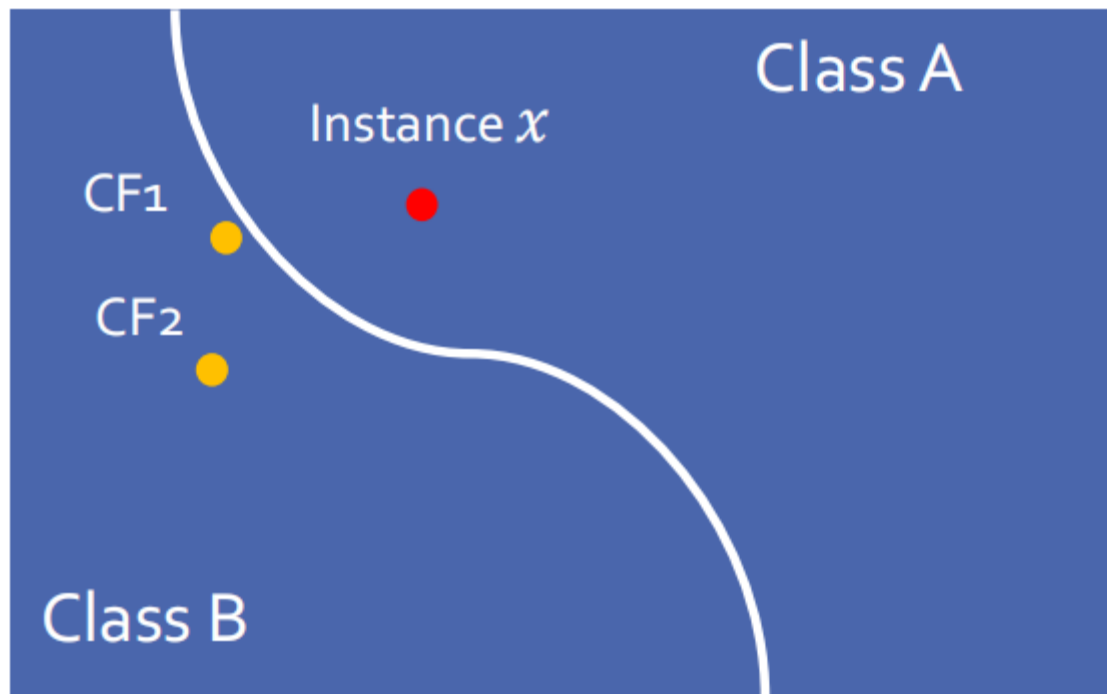
方法2:



Validity=1/3

## ②Proximity: 接近度

接近度：测量反事实与输入数据点的距离



CF1 比 CF2 更接近输入样例点

常见的连续距离度量：

- L1 范数
- L2 范数
- 马氏距离
- ...

由于一些特征是可变的（收入、年龄），而另一些特征是不可变的（种族、原国籍）。因此需要对反事实解释作出限制：

$$x' \in \mathcal{A}$$

### ③Sparsity: 稀疏性

稀疏性：原始输入与反事实解释之间的特征差异数

$$\text{Sparsity: } L_0(x - CF)$$

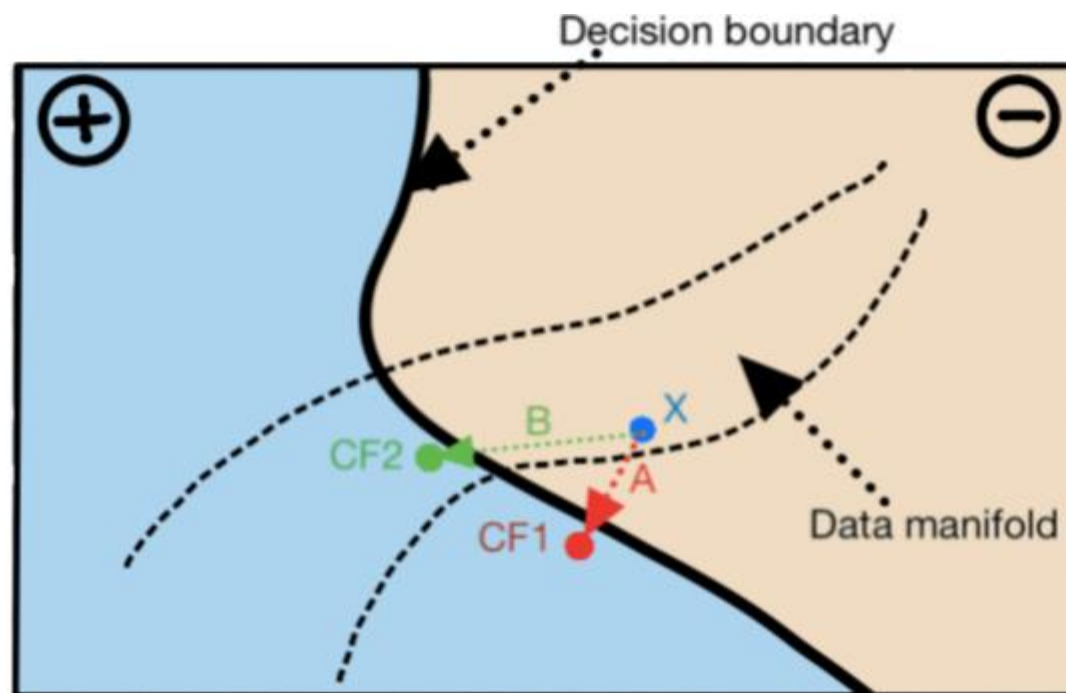
Features	Test Case	“Good” Counterfactual	“Bad” Counterfactual
<i>Weight</i>	80 kg	80 kg	80 kg
<i>Duration</i>	1 hr	<b>1.5 hrs</b>	<b>3 hrs</b>
<i>Gender</i>	Male	Male	<b>Female</b>
<i>Meal</i>	Empty	Empty	<b>Full</b>
<i>Units</i>	6	6	<b>6.5</b>
<i>Bac Level</i>	Over	Under	Under

Sparsity= 1

Sparsity= 4

## ④Data Manifold closeness: 数据流形相似性

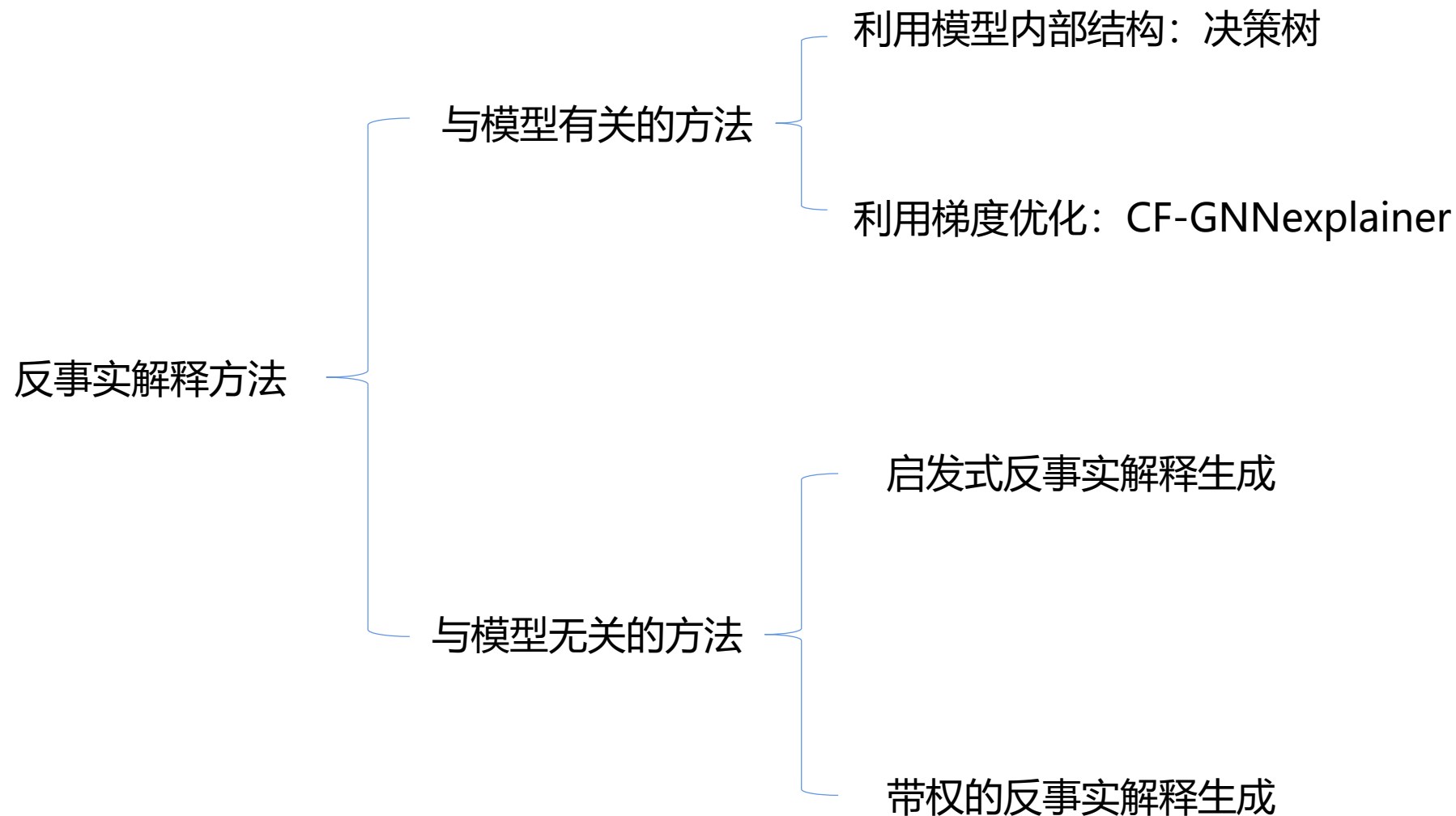
反事实解释器的结果也必须“符合事实”，使其能够被用户接受。所谓流形相近是指，生成的反事实应该和模型见过的观测数据（训练数据）比较接近，而不是生成一个完全偏离训练数据的样本。如下图所示，虽然X与CF1的距离要短与CF2的距离，但由于CF1在数据流形之外，因此选择更合理的CF2





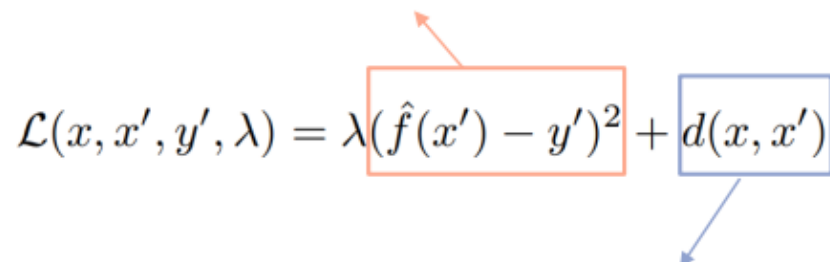
- ◆ 什么是反事实解释
- ◆ 反事实解释的评价指标
- ◆ **反事实解释方法分类**
- ◆ 图神经网络中的反事实解释

# 反事实解释方法分类



# 启发式反事实解释生成

the distance in predictions

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


the distance in instances

$$d(x, x') = \sum_{j=1}^p \frac{|x_j - x'_j|}{MAD_j}$$

---

## Algorithm 1: Counterfactual generation heuristic

---

- 1 sample a random instance as the initial  $x'$
  - 2 optimise  $L(x, x', y', \lambda)$  with initial  $x'$
  - 3 **while**  $|\hat{f}(x') - y'| > \varepsilon$  **do**
  - 4     increase  $\lambda$  by step-size  $\alpha$
  - 5     optimise  $L(x, x', y', \lambda)$  1 with new  $x'$
  - 6 return  $x'$
- 

- 启发式反事实解释生成方法，没有将目标转换为优化问题，而是利用启发式算法来解决。如Rafael Poyiadzi等人提出FACE，利用Dijkstra算法来找到反事实解释。
- 特点：不会生成新的数据点，不涉及数据流形紧密性问题。但可能会陷入局部最优解。同时，所有特征处于一样重要地位。

# 启发式反事实解释生成

启发式反事实生成 - LSAT 示例的问题

Score	GPA	LSAT	Race	GPA x'	LSAT x'	Race x'
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
-0.57	2.7	18.3	0	2.7	34.9	0
Instances				Counterfactual		

Predictions

Desired prediction : Score = 0

LSAT:法学院入学成绩

对于第 3 个人的反事实解释：  
如果您的 LSAT 为 33.5，并且您是“白人”，那么您的平均预测分数为 SCORE(o)  
但事实上，种族不应该被认为和其他特征一样重要



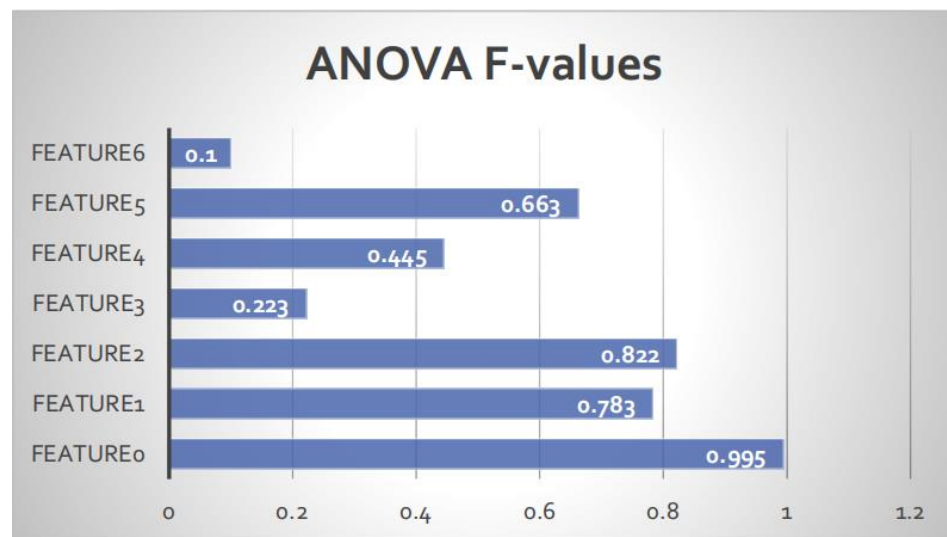
# 带权的反事实解释生成方法

$$\arg \min_{x' \in \mathcal{A}} \max_{\lambda} \lambda (f(x') - y')^2 + d(x, x') + g(x' - x) + l(x'; \mathcal{X})$$

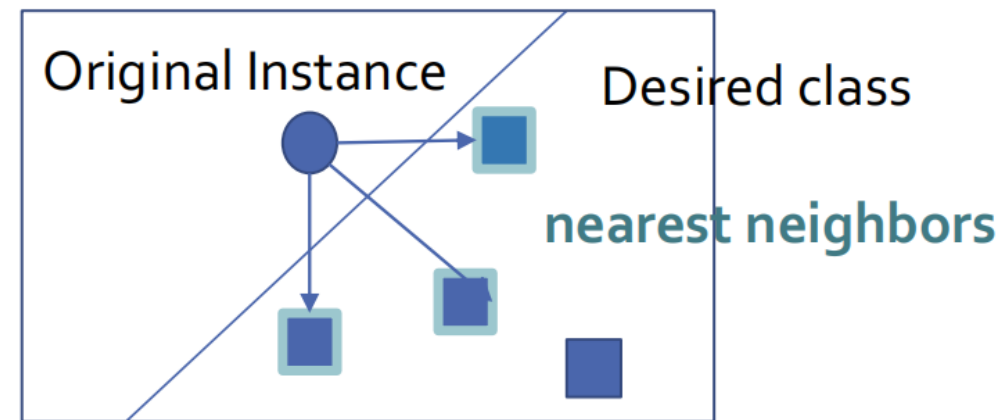
$$d_2(x, x') = \sum_{j=1}^p \frac{|x_j - x'_j|}{MAD_j} \theta_j$$

两种加权策略：

全局特征重要性



K近邻



Decision Boundary



- ◆ 什么是反事实解释
- ◆ 反事实解释的评价指标
- ◆ 反事实解释方法分类
- ◆ **图神经网络中的反事实解释**

# Counterfactual in GNN

---

## ❑ 反事实用于解释GNN

- 改变图结构和节点特征 (GNN-explainer)
- 只改变图结构 (CF-GNNexplainer)

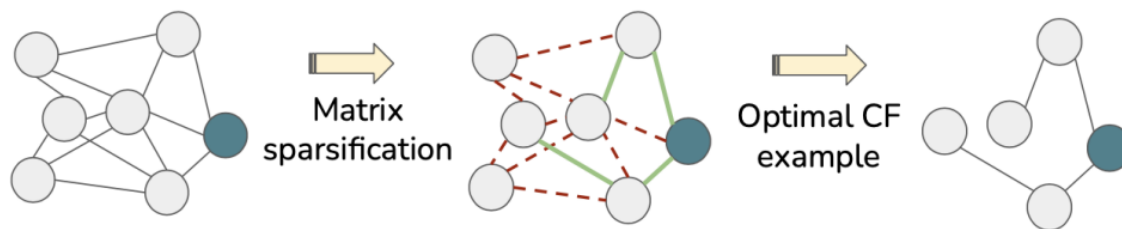
## ❑ 一些具体应用

- Counterfactual Graphs for Explainable Classification of Brain Networks (KDD2021)
- MEG: Generating Molecular Counterfactual Explanations for Deep Graph Networks (IJCNN2021)

# Counterfactual in GNN

## 只改变图的结构

CF-GNNEXPLAINER: 基于矩阵稀疏化技术迭代地从原始邻接矩阵中去除边。



$$\text{GNN model: } f(A_v, X_v; W) = \text{softmax} \left[ (D_v + I)^{-1/2} (A_v + I) (D_v + I)^{-1/2} X_v W \right] \quad \uparrow_{\text{update}}$$

$$\text{CF generation function: } g(A_v, X_v, W; P) = \text{softmax} \left[ \bar{D}_v^{-1/2} (P \odot A_v + I) \bar{D}_v^{-1/2} X_v W \right] \quad \begin{matrix} \uparrow_{\text{update}} & \uparrow_{\text{fix}} \end{matrix}$$

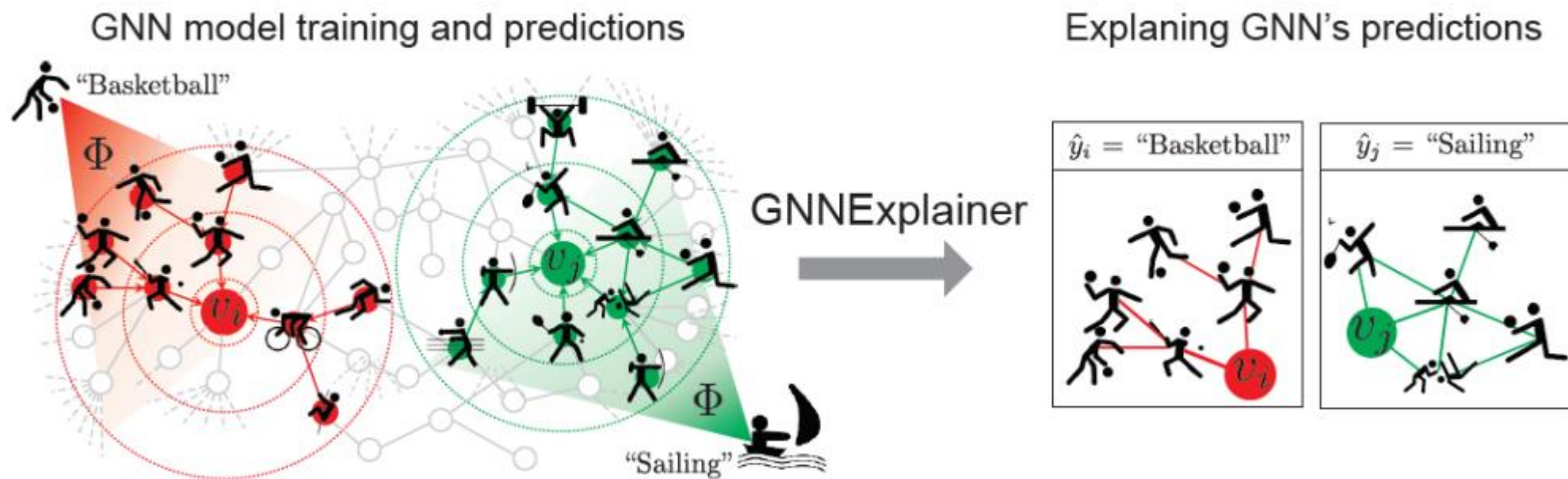
$$\mathcal{L} = \mathcal{L}_{pred}(v, \bar{v} \mid f, g) + \beta \mathcal{L}_{dist}(v, \bar{v} \mid d),$$

$$CF = P \odot A_v$$

# Counterfactual in GNN

## 更改图的结构和节点特征

GNN-explainer:



给定一个实例，GNN-explainer 识别出一个紧凑的子图结构和一小部分节点特征，它们在 GNN 的预测中起着重要的作用

## 更改图的结构和节点特征

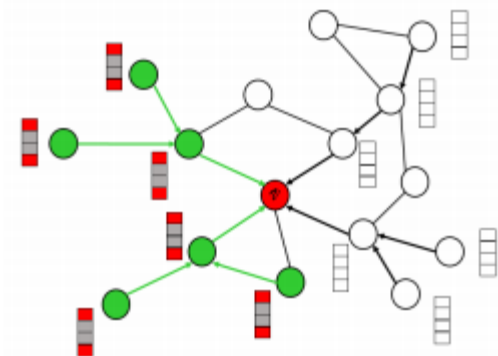
GNNExplainer 中的反事实解释：指定要预测的类别，将生成的目标子图和特征子集作为反事实解释

目标：

$$G_s \in G_c$$

$$X_s \in F_s$$

$$y_i \neq y_i$$



举例：  $v_i \in G_c(v_i)$ ,  $v_j \neq v_i$ , 如果去掉  $v_j$ , 导致预测分类  $y_i'$  的概率大大降低, 那么  $v_j$  可以看成是一个很好的反事实解释, 对特征同样适用。

# Counterfactual in GNN

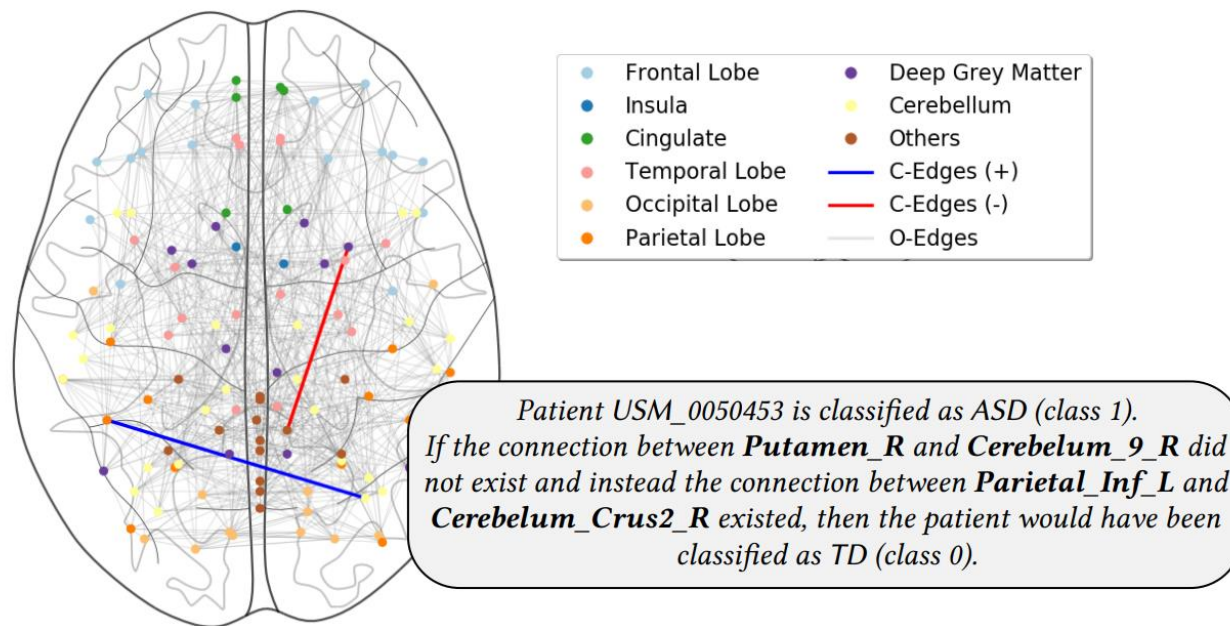
## 应用:Counterfactual Graphs for Explainable Classification of Brain Networks (KDD2021)

作者在文中提出了基于启发式搜索的反事实解释方法:

Step1:Oblivious Forward Search (OFS)

Step2:Oblivious Backward Search (OBS)

来实现基于反事实解释的病理检测。



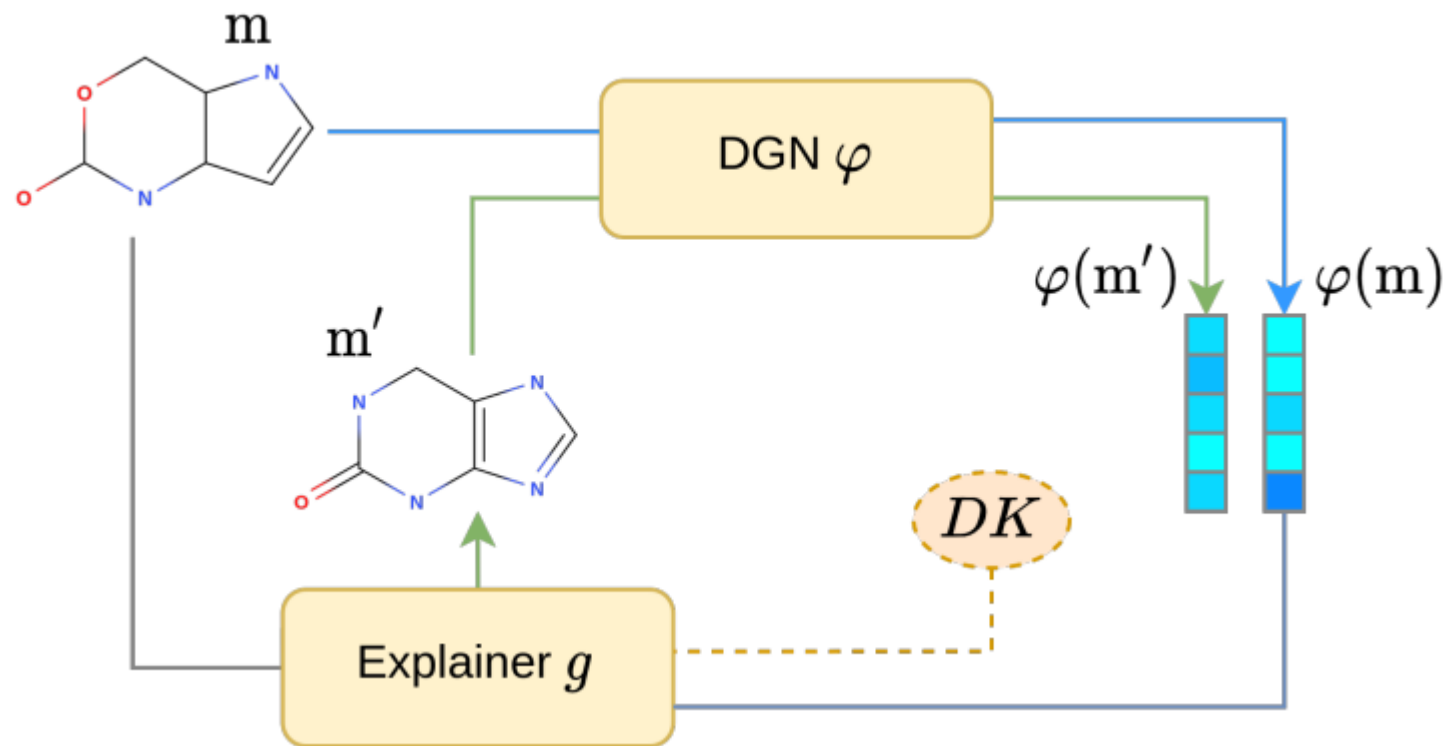
如上图所示: 该病人被机器诊断患有自闭症, 但仅仅改变两条边 (删除**蓝边**, 增加**红边**), 病人会被定义为正常发育。通过这种反事实解释, 可以帮助治疗脑部疾病。



# Counterfactual in GNN

## 应用:MEG: Generating Molecular Counterfactual Explanations for Deep Graph Networks

给定一个训练好的 DGN，作者训练一个基于强化学习的生成器来输出反事实解释。在每一步中，MEG 将当前候选的反事实输入 DGN，将得到的预测结果来奖励强化学习代理来指导流程。



DGN  $\varphi$  是经过训练的分子属性预测器（有毒与无毒），解释器  $g$  是产生反事实的生成代理，它受先验领域知识  $DK$  的约束。



# 参考文献

- [1]Verma S, Dickerson J, Hines K. Counterfactual explanations for machine learning: A review[J]. arXiv preprint arXiv:2010.10596, 2020.
- [2]Wang C, Li X H, Han H, et al. Counterfactual Explanations in Explainable AI: A Tutorial[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 4080-4081.
- [3]Ying Z, Bourgeois D, You J, et al. Gnnexplainer: Generating explanations for graph neural networks[J]. Advances in neural information processing systems, 2019, 32.
- [4]Grath R M, Costabello L, Van C L, et al. Interpretable credit application predictions with counterfactual explanations[J]. arXiv preprint arXiv:1811.05245, 2018.
- [5]Abrate C, Bonchi F. Counterfactual Graphs for Explainable Classification of Brain Networks[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 2495-2504.
- [6]Numeroso D, Bacciu D. Meg: Generating molecular counterfactual explanations for deep graph networks[C]//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.
- [7]Lucic A, ter Hoeve M, Tolomei G, et al. Cf-gnnexplainer: Counterfactual explanations for graph neural networks[J]. arXiv preprint arXiv:2102.03322, 2021.
- [8]Mothilal R K, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations[C]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020: 607-617.