阿里妈妈 / DataFunCon 2021

因果推断在营销科学的应用

张磊 阿里巴巴集团 阿里妈妈 数据技术专家

目录 CONTENT

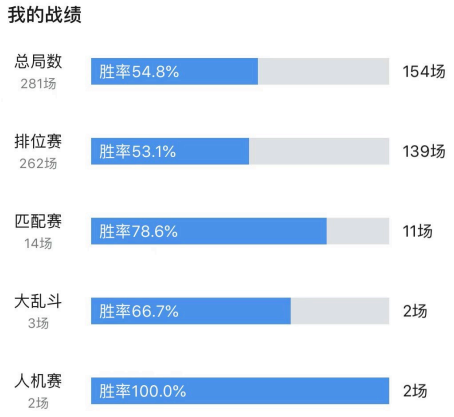
01

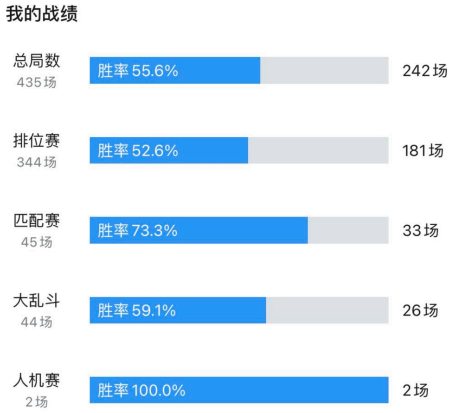
因果科学

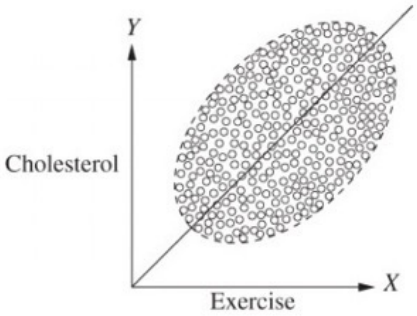
辛普森悖论

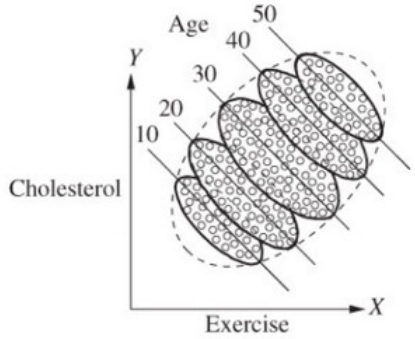
相关性不等于因果性

Population			
	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male			
	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female			
	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%









- 相关性不等于因果性
- 相关关系可以完全的被第三个变量扭曲（混淆变量、内生性），而我们往往一无所知

因果之梯

因果推断在统计与机器学习领域炙手可热

- 因果之梯：科学研究需要从观察，到干预，最终到反事实推理
  - 首先是底层，指的是对于事物现象的一般性观察，并根据观察的现象，发现其中的关联性。
  - 中间一层则是干预，即通过对变量的改变，研究这一变量对结果的影响，包括是否改变结果的性质，以及改变的强度。
  - 最后则是反事实，即通过模拟控制其他变量，仅翻转被研究的变量，探讨可能的发展。

- 11年图灵奖获得者Judea Pearl认为：当前统计机器学习主要关注对表征的拟合，寻找的是变量之间的相关性，而非潜在的因果性。这样的认识会使科学研究停留在较浅的关联层面，导致模型的鲁棒性和可解释性丧失，阻断了进一步探究干预变量，以及反事实推断的能力
- 19年图灵奖得主Yoshua Bengio认为：深度学习已经走到了瓶颈期，将因果关系整合到AI当中已经成为目前的头等大事

如何计算因果关系

PO与SCM两大派系

$P(Y = y|X = x)$

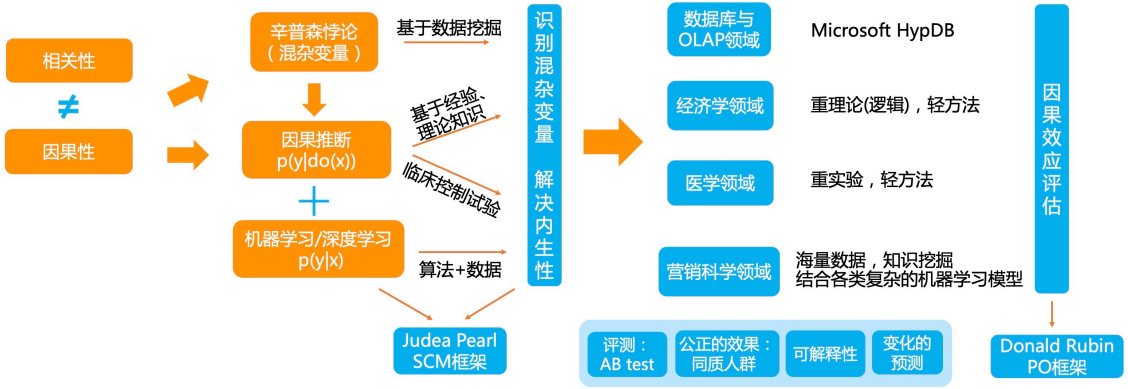
- Condition一个变量，不会改变其分布，目标是判断或预测（即观察自然发生的x并推断y的可能值），监督学习领域有出色表现

$P(Y = y|do(X = x))$

- Intervene一个变量，改变其分布，引起其他变量的改变，目标是根据估计的条件控制或选择x

$P(y_x|x', y')$

- Counterfactual, 假如希望Y变化，需要对x做出什么样的改变



实验

- 控制实验(controlled experiment)，难点是同一个体不可能同时进入实验组和对照组
- 随机实验(randomized controlled experiment)，A/B Test

两大派系

- 以Donald Rubin为代表的Potential Outcome（PO），在经济学和社会科学中有大量的应用实例
- 以Judea Pearl为代表的Structural Causal Model(SCM)，以DAG表示因果关系，深受计算机学者喜爱

PO框架

Potential Outcome理论

- 同质人群对比
  - 如何评估广告投放对转化的真实效果
  - 广告触达的转化率-未触达的转化率？广告触达的人群相比未触达的更活跃，而这波人本身更容易转化
- 将因果识别问题简化
  - 干预 Treatment T： $T_i \epsilon 0,1$ 表示广告触达与否的二值干预
  - 潜在结果Potential outcome： $Y_1(x), Y_0(x)$
  - 混淆变量Confounder
  - 对于单个用户，我们希望得到Individual Treatment Effect (ITE)，也就是ITE= $Y_1(x) - Y_0(x)$
  - 对于整体，通常为Average Treatment Effect (ATE)，ATE=E($Y_1(x) - Y_0(x)$)
- 成熟的估计方法
  - Regression
  - Matching
  - Weighting
  - Stratification
  - IV(Instrumental Variable)

Group	Potential Outcomes	
	Y^1	Y^0
Treatment group (D = 1)	Observable **$E[Y^1 \mid D = 1]$**	Counterfactual **$E[Y^0 \mid D = 1]$**
Control group (D = 0)	Counterfactual **$E[Y^1 \mid D = 0]$**	Observable **$E[Y^0 \mid D = 0]$**

ATE（Average Treatment Effect）为：

$$
\begin{aligned}
E[\delta] &= E[Y^1 - Y^0] \\
&= E[Y^1] - E[Y^0] \\
&= \{\pi E[Y^1 \mid D = 1] + (1 - \pi)\, E[Y^1 \mid D = 0]\} \\
&\quad - \{\pi E[Y^0 \mid D = 1] + (1 - \pi)\, E[Y^0 \mid D = 0]\} \\
&= \pi\{E[Y^1 \mid D = 1] - E[Y^0 \mid D = 1]\} + \\
&\quad (1 - \pi)\{E[Y^1 \mid D = 0] - E[Y^0 \mid D = 0]\}
\end{aligned}
$$

因果图SCM框架

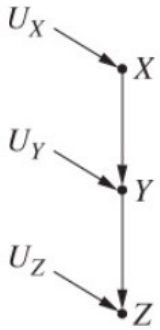
Structural Causal Model理论

- 因果图表示
  - 外生变量U: exogenous variables
  - 内生变量V: endogenous variables
  - 函数集合F: A variable X is a direct cause of a variable Y if X appears in the function that assign Y's value
- chains, forks, colliders
  - chains：condition on Y，X和Z是独立的
  - forks : condition on X，Y和Z是独立的
  - colliders： X和Y是相互独立的；如果condition on Z 那么X和Y就是非独立了
  - d-separation： Z 阻断了X 到 Y 的所有路径，那么称 Z d分离 X 和 Y，记为 $(X \perp Y|Z)_G$ 用于确定X与Y之间独立，需要控制哪些变量
- 因果关系识别
  - 后门准则（backdoor criterion）
  - 前门准则（frontdoor criterion）

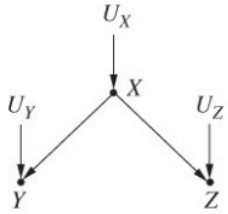
学校的经费(X), 平均成绩(Y), 年度录取率(Z)

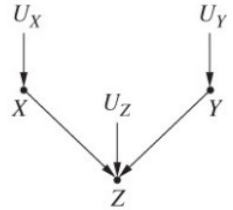
$$U = \{U_x, U_y, U_z\}, V = \{X, Y, Z\}, F = \{f_x, f_y, f_z\}$$

$$f_x: X = U_x \qquad f_y: Y = \frac{x}{3} + U_y \qquad f_z: Z = \frac{y}{16} + U_z$$



$$P(X = x, Y = y, Z = z) = P(X = x)P(Y = y|X = x)P(Z = z|Y = y)$$

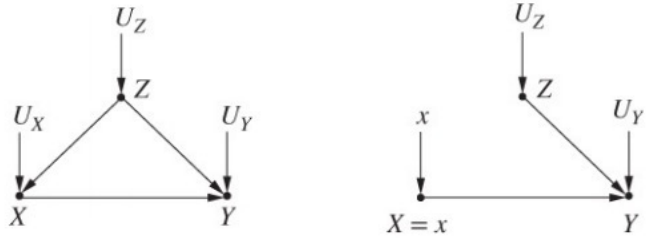




因果图SCM框架

Structural Causal Model理论

- SCM与PO
  - 通过后门准则找到干预的变量集z，adjust for z
  - Regression、Matching和Weighting等方法，也适用于SCM
  - SCM基于图表征因果关系，对于计算机非常友好，而PO框架提供了丰富的因果效应求解方法
  - 微软因果推断工具DoWhy：使用SCM表征因果关系，PO求解因果效应
- SCM的构造
  - 基于经验知识
  - 基于data自动构造，计算机领域的热门方向
- 相关书籍
  - The Book of Why: The New Science of Cause and Effect
  - Casual inference in Statistics, A Primer



$$P(Y=y|do(X=x)) = P_m(Y=y|X=x)$$

$$= \sum_z P_m(Y=y|X=x,Z=z)P_m(Z=z|X=x)$$

$$= \sum_z P_m(Y=y|X=x,Z=z)P_m(Z=z)$$

adjust for z

$$= \sum_z P(Y=y|X=x,Z=z)P(Z=z)$$

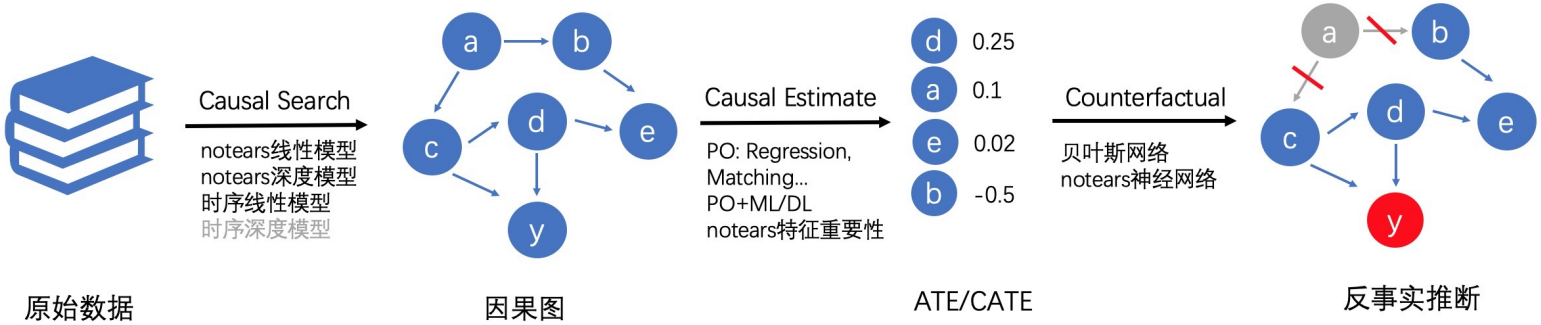
$$= \frac{\sum_z P(Y=y|X=x,Z=z)P(X=x|Z=z)P(Z=z)}{P(X=x|Z=z)}$$

$$= \sum_z \frac{P(X=x,Y=y,Z=z)}{P(X=x|Z=z)}$$

倾向分，IPTW

因果科学

CSD与CEI



- SCM表征因果关系，PO计算因果效应
- 因果科学
  - 因果结构发现（Causal Structural Discovery， CSD）
  - 因果效应推断（Causal Effect Inference，CEI）

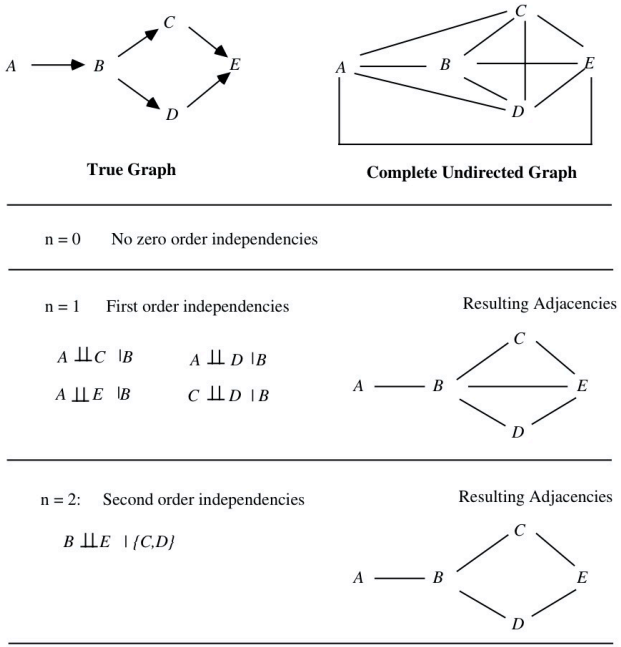
02

因果推断与机器学习

Causal Structural Discovery

Constraint-based Algorithms

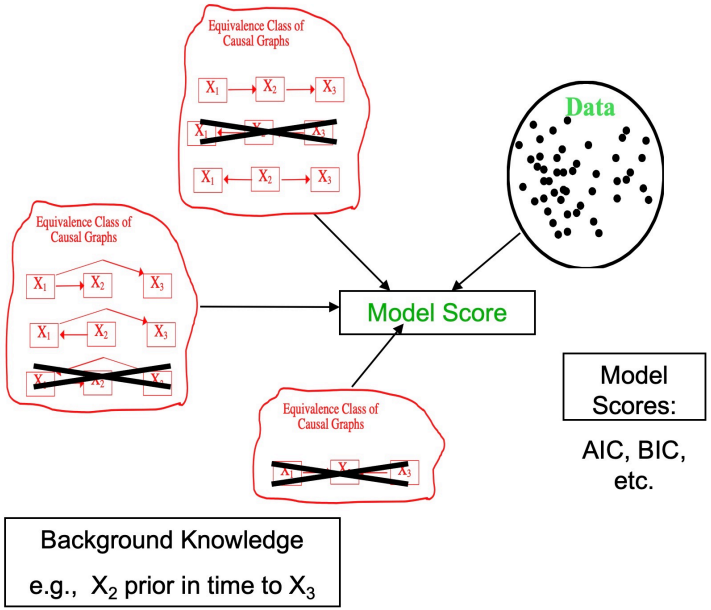
- 方法概述
  - 通过chains, forks, colliders三种结构条件独立的检验，最终构造SCM
  - 优点：方法思路清晰，可解释性强
  - 缺点：算法复杂度非常高，并且对数据质量要求高，不能有unobserved confounder，对chain和fork结构，只能得到马尔可夫等价类
- 相关研究
  - IC algorithm:《Causality: models, reasoning, and inference》[Judea_Pearl], p60
  - PC algorithm: Causation, Prediction, and Search, 2000
  - Fast Causal Inference(FCI): Spirtes et al., 2001



Causal Structural Discovery

Score-based Algorithms and Machine Learning

- 方法概述
  - 通过最优化给图打分的函数来找到最优的图
  - 优点：定义打分函数，优化目标明确，非常容易和机器学习模型结合
  - 缺点：算法复杂度非常高，需要搜索所有的图，NP-hard问题，容易局部最优
- 相关研究
  - 打分函数：BDe(u) (Heckerman et al., 1995), BGe (Kuipers et al., 2014), BIC (Chickering and Heckerman, 1997), and MDL (Bouckaert, 1993)
  - GES: Greedy Equivalence Search , 2003
  - CGNN: Generative Neural Networks, 2017
  - Continuous Optimization for Structure Learning: 2018
  - CAUSAL DISCOVERY WITH REINFORCEMENT LEARNING: 2020
- 其他方法
  - Functional Causal Models: ANM, LiNGAM, CAM
  - Hybrid methods



Causal Effect Inference

CEI and Machine Learning

- ATE
  - Average Treatment Effect，宏观overall的因果效应估计
  - Regression、Matching、Weighting、Stratification
  - LR回归、Nearest Neighbor 、遗传算法等应用于matching
  - Covariate Balancing Method对样本重新赋权，与LR、DNN结合：IPTW、EB、ARB、CBPS
  - IV：两阶段回归、deepIV等
- ITE
  - Individual Treatment Effect，估计个体因果效应
  - 典型的uplift model
- CATE
  - Conditional Average Treatment Effect，异质性群体的因果效应评估
  - Meta-learner、Tree-based algorithms、Deep Learning Method（CEVAE、Balancing Neural Network、TARNet、BART）

$$ITE_i = \tau_i = y_i^1 - y_i^0$$

$$ATE = E_i[\tau_i] = E_i[y_i^1 - y_i^0] = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0)$$

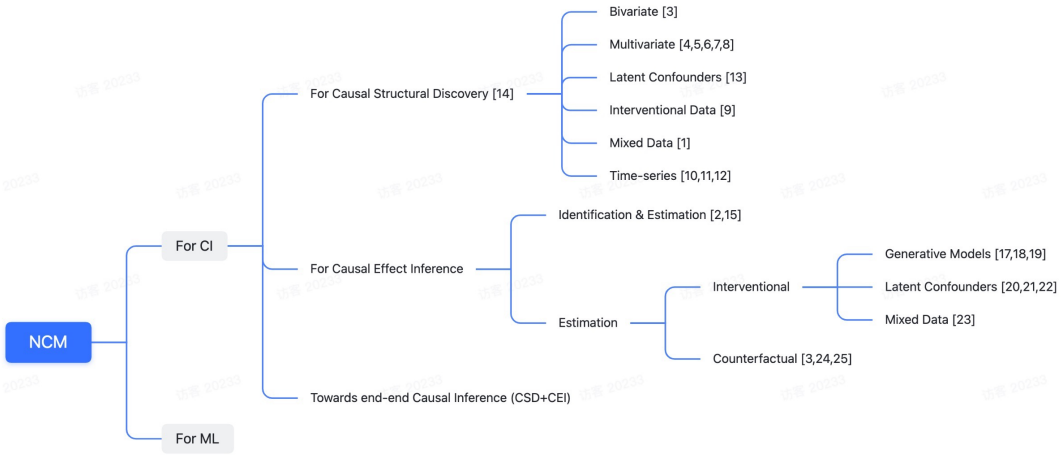
$$CATE = \tau(X) = E_{i:x_i \in X}[\tau_i]$$

Neural Causal Models

NCM与因果推断的应用场景

- 因果推断应用场景
  - 典型的因果问题：uplift model，同质人群对比，A/B Test，反事实推断
  - 分类、推荐等场景，结合因果推断，结合matching、weighting消除数据偏差，或者将因果机制作为约束条件，解决长尾问题
  - 模型的可解释性：贡献分配、关键因素洞察、运营序列分析

- 神经因果模型
  - 2019年提出，使用神经网络建模结构因果模型SCM
  - 神经网络与因果科学的结合：主流MLP、GNN、AutoEncoder等网络与CSD、CEI的结合
  - NCM for Causal Inference：完成因果识别、估计任务
  - NCM for Machine Learning：解决机器学习问题提出的融合因果约束、因果机制



03

因果推断与营销科学

因果推断与营销科学

因果推断在数字营销的应用

- 评测实验
  - 结合因果推断对历史数据洞察，辅助选择实验组、对照组
- 同质人群对比
- 用户增长
  - 典型的uplift model
  - 留存的关键因素洞察，可解释，运营序列抓手
- 多渠道归因、预算分配
  - 贡献分配、结合matching、weighting消除数据偏差，反应渠道的真实价值
  - 建立干预、反事实模型，评估预算分配的全面影响
- 异常分析、可解释性
  - 波动、异常归因与解释
  - 优质视频关键因素挖掘，优化创意素材
- 推荐、lookalike、CTR/CVR预估、优惠券等典型算法应用
  - 结合因果推断，消除数据偏差
  - 将因果机制作为约束条件，解决长尾问题

因果推断与用户增长

因果推断在用户增长的应用

如何助力增长

- 提供特色的投放功能，为客户提供有价值的工具
- 合理的引导体系，帮助客户循序渐进了解营销工具的使用
- 了解客户的核心诉求，指导广告投放，取到满意的效果

核心诉求

- 确立北极星指标，即洞察对于用户留存的关键因素
- 依据北极星指标，进行运营序列拆解

需要解决的问题

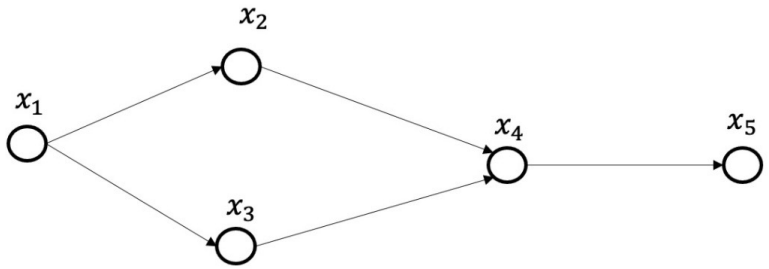
- 哪些投放功能对于留存具有非常大的价值？
- 如何判断流失预警的客户？
- 流失的原因是什么？如果避免这些原因，对于整体留存率有多少提升？
- 定位原因后，如何指导客户去完成目标？

模型该怎么做

- 对于新开通的客户，建立预测模型，预测30天之后的留存状态
- 以留存果，挖掘影响留存的因，基于SCM表征因果关系，求解其因果效应
- 干预因，计算留存的提升率

Neural Causal Models

Learning Sparse Nonparametric DAGs



$$P(x_1, x_2, x_3, x_4, x_5) = \prod_{i=1}^{5} P(x_i | Pa(x_i))$$

$$= P(x_1)P(x_2|x_1)P(x_3|x_1)p(x_4|x_2x_3)P(x_5|x_4)$$

$$\begin{aligned} E_i &\sim \varepsilon \\ x_1 &= f_1(E_1) \\ x_2 &= f_2(x_1, E_2) \\ x_3 &= f_3(x_1, E_3) \\ x_4 &= f_4(x_2, x_3, E_4) \\ x_5 &= f_5(x_4, E_5) \end{aligned}$$

SEM方程表示：
对于i.i.d特征 $X = (X_1, \ldots, X_d)$

与DAG图G=(V, E)， V=X，存在函数 f_j 与 g_j

$\mathbb{E}[X_j \mid X_{\mathrm{pa}(j)}] = g_j(f_j(X)), \quad \mathbb{E}f_j(X) = 0$

如果 $X_k \notin pa(j)$ 那么 $f_j(u_1, \ldots, u_d)$ 与 u_k 相互独立

g_j 对应non-additive errors

找到DAG G(X)，得到 $f = (f_1, \ldots, f_d)$
最优化损失 $\ell(y, y')$

$$\min_f \; L(f) \quad \text{subject to} \quad G(f) \in \text{DAG},$$

$$\text{where} \quad L(f) = \frac{1}{n} \sum_{j=1}^{d} \ell(\mathbf{x}_j, f_j(\mathbf{X})).$$

Neural Causal Models

Learning Sparse Nonparametric DAGs

$$\min_f \; L(f) \;\; \text{subject to} \;\; \mathsf{G}(f) \in \mathsf{DAG},$$

$$\text{where} \quad L(f) = \frac{1}{n}\sum_{j=1}^{d} \ell(\mathbf{x}_j, f_j(\mathbf{X})).$$

如何满足$G(f) \in DAG$

f_j 与X_k在Sobolev Spaces相互独立的条件

$\|\partial_k f_j\|_{L^2} = 0$, where $\|\cdot\|_{L^2}$ is the usual L^2-norm

$[W(f)]_{kj} := \|\partial_k f_j\|_{L^2}$.

$$\min_{f: f_j \in H^1(\mathbb{R}^d), \forall j \in [d]} L(f) \;\; \text{subject to} \;\; h(W(f)) = 0.$$

如何定义f

如何保证$G(f)$稀疏

$$\mathsf{MLP}(u; A^{(1)}, \ldots, A^{(h)}) = \sigma(A^{(h)}\sigma(\cdots A^{(2)}\sigma(A^{(1)}u))),$$

$$A^{(\ell)} \in \mathbb{R}^{m_\ell \times m_{\ell-1}}, \quad m_0 = d.$$

$$\min_\theta \quad \frac{1}{n}\sum_{j=1}^{d} \ell(\mathbf{x}_j, \mathsf{MLP}(\mathbf{X}; \theta_j)) + \lambda \|A_j^{(1)}\|_{1,1}$$

$$\text{subject to} \;\; h(W(\theta)) = 0.$$

如何定义损失，L-BFGS-B算法

$$\min_\theta \; F(\theta) + \lambda\|\theta\|_1,$$

$$F(\theta) = L(\theta) + \frac{\rho}{2}|h(W(\theta))|^2 + \alpha h(W(\theta))$$

因果推断与用户增长

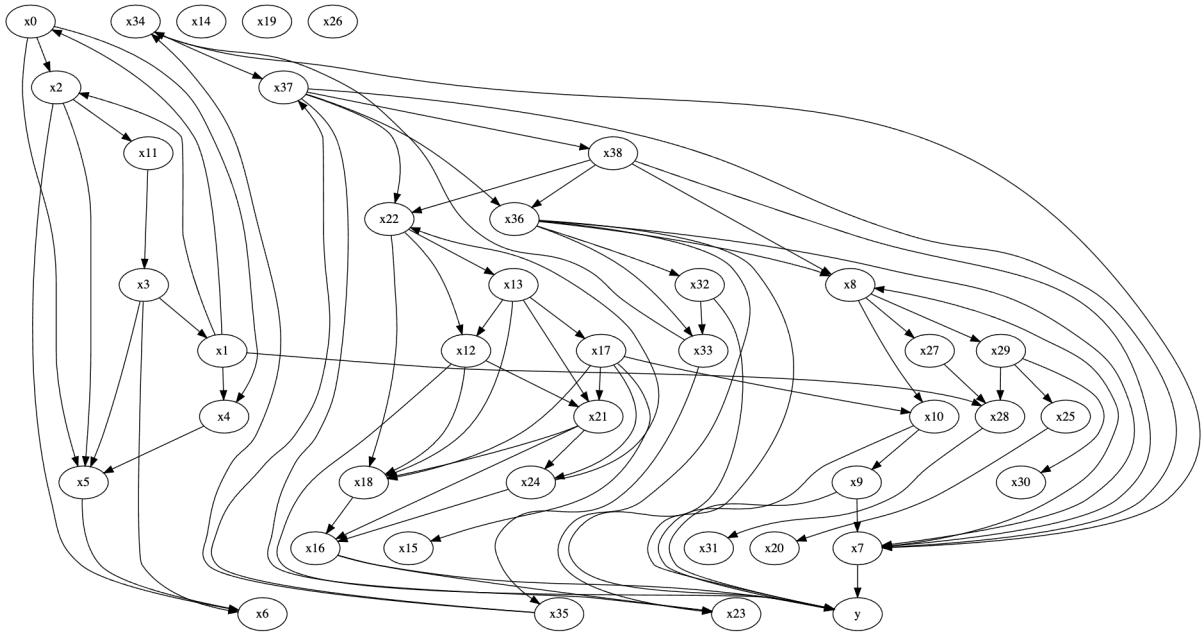
NCM模型应用于用户增长的结论

建模过程

- 将留存作为outcome，各类潜在因素作为treatment，使用NCM建立二分类模型
- 训练NCM，导出SCM以及各条边的权重（因果效应）
- 通过NCM完成干预/反事实任务

结论

- 留存预测模型的AUC达到0.84
- SCM因果关系图，整体符合预期，细节需要微调，引入knowledge
- 与相关性分析相比，模型得到的结论更符合认知
- A/B实验表明，留存率、活跃度提升明显



非常感谢您的观看

DataFunCon 2021

阿里妈妈广告技术-SDS
zl165646@alibaba-inc.com