从 Python 入手因果发现

『社会系统中的因果发现』讲习班

计算技术研究所 李奉治

lifengzhi20z@ict.ac.cn

2021 年 12 月 11 日

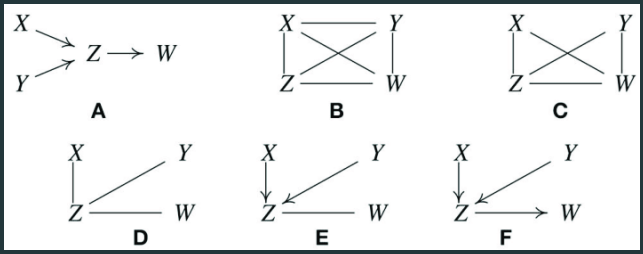
大纲

1. 实现一个基础 PC 算法

2. 使用 causal-learn 算法包

实现一个基础 PC 算法

实现一个基础的 PC 算法

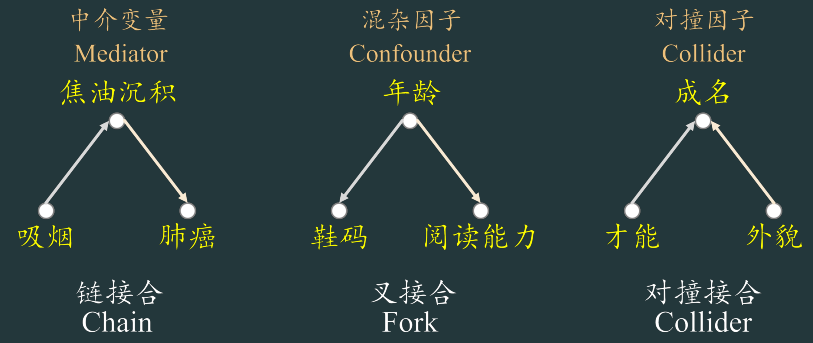


1. 构建一个无向完全图
2. 学习因果骨架（使用条件独立性检验）
3. 标注 V-结构（使用条件独立性检验）
4. 根据限制补充方向（不引入新的 V-结构，且不引入环）

学习因果骨架

1. $n = 0$
2. **重复循环**
   - **重复循环**
     - 选择一对相邻结点 X 和 Y，若 **Adjacencies**$(G, X) \setminus \{Y\}$ 中的元素数量大于等于 n，则选择其中一个大小为 n 的子集：若 X 和 Y 在以此子集为条件时独立，则删除 X 和 Y 之间的边，并将这个子集记录到 **Sepset**(X, Y) 和 **Sepset**(Y, X)
   - **直到**对于每一对相邻结点 X 和 Y，若则其中大小为 n 的子集都已经通过条件独立性检验。
   - $n = n + 1$
3. **直到**对于每一对相邻结点 X 和 Y，**Adjacencies**$(G, X) \setminus \{Y\}$ 中的元素数量都少于 n

标注 V-结构



1. 对于每一对结点 X 和 Y，如果他们相同的邻居结点，记这个邻居结合为 K_common
   - 对于 K_common 中的每个结点 K，如果 K 不会使得 X 和 Y 满足条件独立（即 K 不在 **Sepset**(X, Y) 中），则标注 V-结构：$X \to K \leftarrow Y$

根据限制补充方向

- 限制一：不引入新的 V-结构（因为上一步应该已经标出了所有正确的 V-结构）
- 限制二：不引入环结构（因为我们假设因果图的有向无环图）

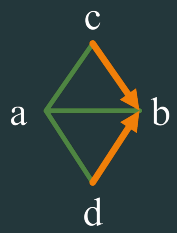
补充方向规则：

- R1: 将 $a \to b - c$ 标为 $a \to b \to c$，若 a 和 c 不相邻
- R2: 将 $a - b$ 标为 $a \to b$，若存在链结构 $a \to c \to b$.
- R3: 将 $a - b$ 标为 $a \to b$，若存在链结构 $a - c \to b$ 和 $a - d \to b$，且 c 和 d 不相邻

补充方向规则

- R1: 将 $a \to b - c$ 标为 $a \to b \to c$，若 a 和 c 不相邻
- R2: 将 $a - b$ 标为 $a \to b$，若存在链结构 $a \to c \to b$.
- R3: 将 $a - b$ 标为 $a \to b$，若存在链结构 $a - c \to b$ 和 $a - d \to b$，且 c 和 d 不相邻

完备性:

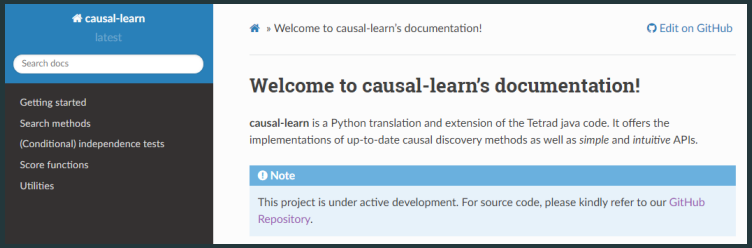


Christopher Meek, 1995, Causal inference and causal explanation with background knowledge.

使用 causal-learn 算法包

causal-learn 简介

> Causal-learn，由 CMU 张坤老师主导，多个团队（CMU 因果研究团队、DMIR 实验室、宫明明老师团队和 Shohei Shimizu 老师团队）联合开发出品的因果发现算法平台。



算法支持

- 基于约束的因果发现方法（PC、FCI、CD-NOD）
- 基于评分的因果发现方法（GES、Exact Search）
- 基于因果函数模型的因果发现方法（LiNGAM、ANM、Post-nonlinear causal models Additive noise models）
- 基于梯度的因果发现方法
- Hidden causal representation learning

威斯康星大学医院-乳腺癌数据集

数据来源：

archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

O. L. Mangasarian and W. H. Wolberg: ”Cancer diagnosis via linear programming”, SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

数据采集于威斯康星大学医院，涉及 699 个乳腺癌样本，每个样本包含以下 11 个属性

- 样本编码（整数 id）
- 块丛厚度、细胞大小的均匀性、细胞形状的均匀性、边缘附着力、单上皮细胞大小、裸核、染色质、正常核仁、有丝分裂（均为 1-10 的整数）
- 癌细胞分类：2 表示良性,4 表示恶性

其中包含了 16 个缺失的数据点，以? 来表示

感谢倾听！

Q&A