第三次作业

李子龙
上海交通大学
计算机科学与工程系
logcreative@outlook.com

1 SVM 与神经网络

1.1 数据集信息

对两个数据集进行测试，其规模如表 1 所示。其中 madelon 数据集特征维度多，训练集大小大于测试集大小；ijcnn1 数据集特征维度相对较少，但是数据集规模大，测试集大小大于训练集大小。

表 1: 数据集信息

数据集	训练集大小	测试集大小	特征维度
madelon	2000	600	500
ijcnn1	49990	91701	22

1.2 与 MLP 的比较

数据读取实现于 `src/utils.py`，特征将会被首先归一化再进行训练。SVM 实现于 `src/svm.py`，具体参数为

Listing 1: `src/svm.py`

```
12      model = svm.SVC(kernel=kernel, C=C)
```

MLP 实现于 `src/mlp.py`，具体参数为

Listing 2: `src/mlp.py`

```
12      model = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(int(
        feat_size*5)), random_state=1, max_iter=feat_size*30)
```

表 2 展示了默认参数下 SVM 与 MLP 的效果。从准确率来看，SVM 的准确率在 madelon 多特征维度数据集上略高于 MLP，在 ijcnn1 向本较多的数据集上低于 MLP。训练时间

上 SVM 收敛需要的时间也偏长，当然从后文可以看到这个时间可以通过调节参数的方式缩短。

表 2: SVM 与 MLP

数据集	准确率		训练时间 (s)	
	SVM	MLP	SVM	MLP
madelon	**0.585**	0.583	71	19
ijcnn1	0.919	**0.961**	130	91

1.3 不同的核函数

表 3 展示了使用不同核函数的结果。其中在多维度的 madelon 上 linear 核的表现最好，但是训练时间较长，使用其它核会略微降低一点准确率，但是时间可以减少一个数量级。在少一些维度的 ijcnn1 上，rbf 的准确率最高，可以超过表 2 的 MLP，此时的 poly 核是更好的性价比选择。

表 3: SVM 不同核函数，$C = 1$

数据集	准确率				训练时间 (s)			
	linear	poly	rbf	sigmoid	linear	poly	rbf	sigmoid
madelon	**0.585**	0.578	0.582	0.583	76	5	7	5
ijcnn1	0.919	0.948	**0.968**	0.867	150	70	141	127

1.4 不同的 C

表 4 展示了不同的 C 对 linear SVM 的影响，C 将控制对软间隔的容忍度。可见其对准确率不会有特别大的影响，但是在训练时间上会有差异。同等准确率的情况下，对 madelon 而言，$C = 0.1$ 最好；对 ijcnn1 而言，$C = 0.1$ 最好。

表 4: 不同的 C，linear

数据集	准确率				训练时间 (s)			
	0.01	0.1	0.5	1	0.01	0.1	0.5	1
madelon	0.568	**0.585**	0.57	**0.585**	5	9	29	71
ijcnn1	0.918	**0.919**	**0.919**	**0.919**	87	104	121	188

2 因果发现算法

2.1 数据集

导致肺癌的因素很多，本题将采用 LUCAS (LUng CAncer Simple set) 数据集[1]，来计算这个数据集 12 个因素（如表 5）之间的因果关系图。这 12 个因素都被编码为 0/1 值，共 2000 行。

表 5: 因素

序号	英语名	中文名
0	Lung Cancer	肺癌
1	Smoking	吸烟
2	Yellow_Fingers	黄手指
3	Anxiety	焦虑
4	Peer_Pressure	同辈压力
5	Genetics	基因
6	Attention_Disorder	注意力紊乱
7	Born_an_Even_Day	在偶数日出生
8	Car_Accident	车祸
9	Fatigue	疲劳
10	Allergy	过敏
11	Coughing	咳嗽

2.2 算法

这里采用 LiNGAM（Linear Non-Gaussian Acyclic Model）算法[2]，该算法假设变量满足

$$x_i = \sum_{j:\text{parents of}\, i} b_{ij} x_j + e_i \tag{2.1}$$

并有如下假设：

1. 因果图为有向无环图（DAG）
2. 外部影响 e_i 都是非零方差的、并且独立非高斯
3. 假设不存在 latent confounders

假设 1 与 3 在图 3 中也可以看到都是正确的。假设 2 根据数据描述也被认为是正确的。

ICA 对于输入的数据矩阵 $\mathbf{X}$，每一列为一个样本向量 $\mathbf{x}$，归一化消去偏置后，将公式 (2.1) 进行转换，使用矩阵形式得到

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \tag{2.2}$$

需要对 $\mathbf{x}$ 进行求解

$$\mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e} = \mathbf{W}^{-1}\mathbf{e} \tag{2.3}$$

这将利用 ICA (Fast-ICA) 求解唯一的 $\mathbf{W}$，然后对排序得到对角线不为零的矩阵 $\tilde{\mathbf{W}}$，归一化对角元素得到 $\tilde{\mathbf{W}}'$。从而估计权重矩阵 $\hat{\mathbf{B}} = \mathbf{I} - \hat{\mathbf{W}}'$。

排序 为了一一对应 x_i 与 e_i 对应，我们需要对 $\hat{\mathbf{B}}$ 进行排序，可以表示为 $\tilde{\mathbf{B}} = \mathbf{P}\hat{\mathbf{B}}\mathbf{P}^T$，$\tilde{\mathbf{B}}$ 接近于一个严格的下三角矩阵。从而得到一个有向无环图。

剪枝 删去不重要的边以得到最终结果。

2.3 结果

参考这篇博客[3]，在 Tetrad[4] 建立如图 1 所示的流程图。

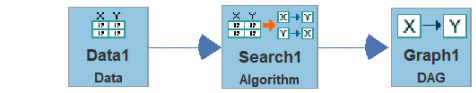

图 1: 流程图

最后推断结果如图 2 所示。这个结果可以与生成数据的标准答案图 3 做对比。可以发现有些联系方向依然有一些问题，比如 Car Accident→Genetics 和 Car Accident→Attention Disorder 关系明显可以用已有知识判定为错误。在 Lung Cancer→Smoking，Lung Cancer→Genetics，Lung Cancer→Fatigue，Smoking→Anxiety，Smoking→Peer Pressure 等为相反方向，Genetics→Smoking，Fatigue→Allergy，Fatigue→Attention Disorder 为新增边。

大部分的框架仍然是很明晰的，而部分错误结果与 ICA-LiNGAM 算法的缺陷有一定的关系，使用 FastICA 算法可能收敛到局部最优解，变量排序可能因尺度发生变化。此算法的进阶版本为 Direct LiNGAM，可以接受先验知识也可以得到更加可靠的结果，但也会导致更高的算法复杂度。由于 Tetrad 并没有实现 Direct LiNGAM，此处无法引入先验知识，只采用了 ICA-LiNGAM 求解。

References

[1] CAUSALITY WORKBENCH PROJECT T. Lucas and lucap are lung cancer toy datasets [EB/OL]. http://www.causality.inf.ethz.ch/data/LUCAS.html.

[2] 王淼皓. 非时序线性非高斯模型—— LiNGAM[EB/OL]. 2022. https://zhuanlan.zhihu.com/p/369720949.

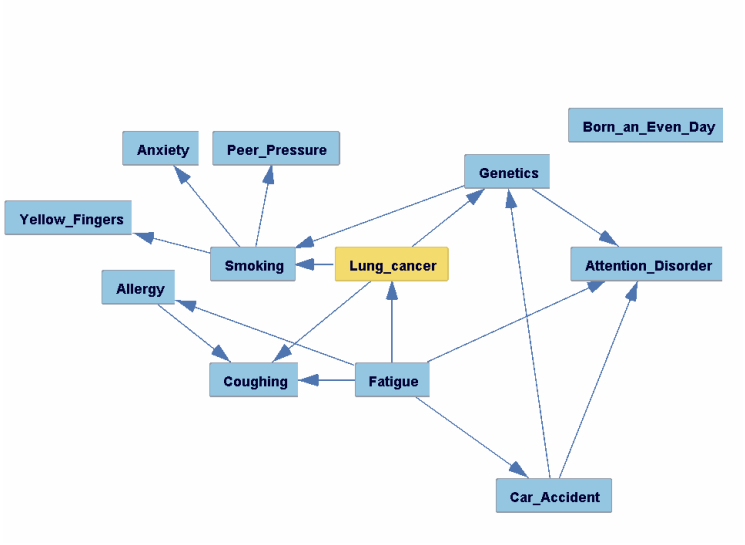

图 2: 因果推断结果

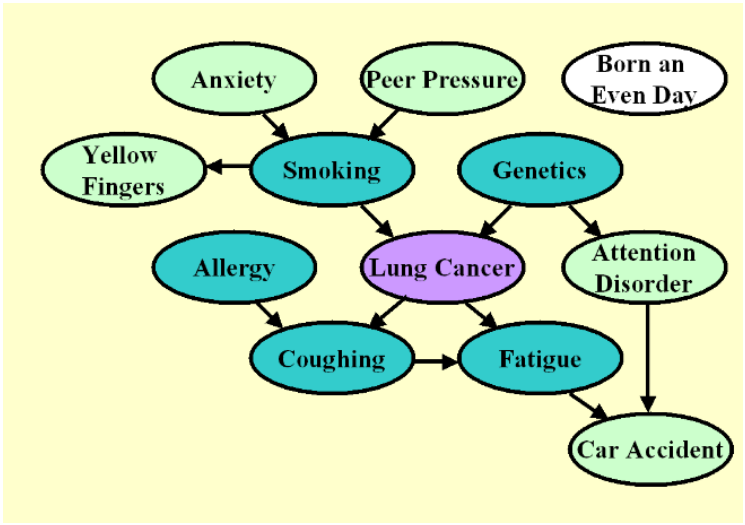

图 3: 生成数据的依赖图

[3] LIN Y. 因果推断学习笔记（五）：画画因果图[EB/OL]. 2019. https://www.dango.rocks/blog/2019/09/24/Causality5-Drawing-Causal-Diagram/.

[4] RAMSEY J D, KUN Z, MADELYN G, et al. TETRAD – a toolbox for causal discovery [C/OL]//8th International Workshop on Climate Informatics. 2018. https://github.com/cmu-phil/tetrad.