第三次作业

Problem 1. SVM vs Neural Network

比较 SVM 和 Neural Network 在不同数据集上的判别能力。按照要求从两个网站选定了三个数据集：iris，colon-cancer 和 mnist[3, 4]。

(1) 首先是 iris 数据集，该数据集提供了鸢尾花的多个特征，并根据他们对鸢尾花进行分类。表1中展示的是一个两个隐藏层的使用 relu 激活的多层感知机和核函数为多项式，C 为 0.01 的 SVM，他们在测试集上的准确率均为 100%。在这样一个特征维度并不高的数据集上，训练集充分，两种模型均可以学到极好的分类能力。图1和图2展示了各个参数下的 SVM 和 MLP 的测试结果。

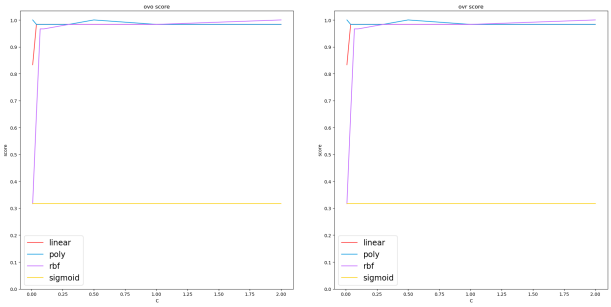


图 1: Iris 数据集上的 SVM 表现。

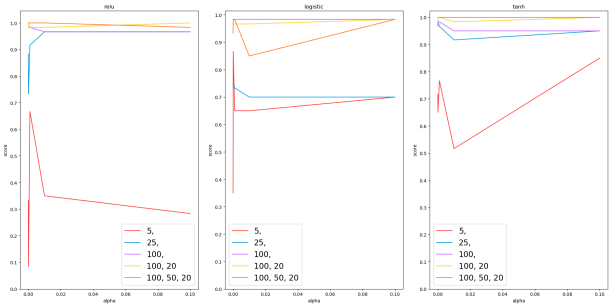


图 2: Iris 数据集上的 MLP 表现。

表 1: Iris 上的准确率比较。

方法	准确率
MLP(relu,0.001,100×20)	99.77%
SVM(polynomial kernel with C=0.01)	100%

(2) Colon cancer 数据集提供了结肠癌患者和普通患者的数据，希望二分类问题，即患者是否患有结肠癌症。表??中核函数为 sigmoid 函数，C 为 2 时，准确率可以达到 84%；而多层感知机，由于该数据集上训练集数量较小，由于我们设置的维度较大，都未能收敛，仅有 10% 的准确度，成了反向的分类器。图3和图4展示了各个参数下的 SVM 和 MLP 的测试结果。

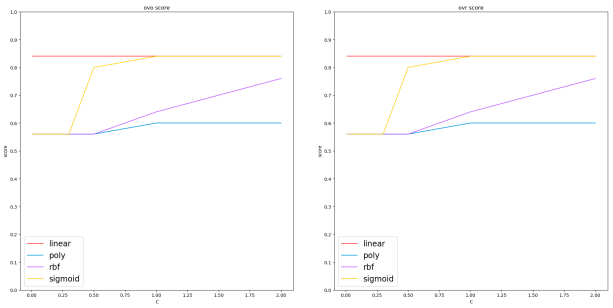


图 3: Colon Cancer 数据集上的 SVM 表现。

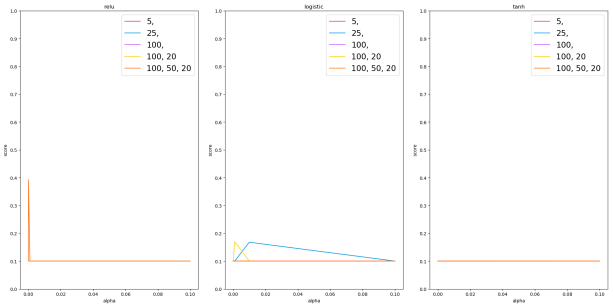


图 4: Colon Cancer 数据集上的 MLP 表现。

表 2: Colon-cancer 上的准确率比较。

方法	准确率
MLP	10%
SVM(sigmoid kernel with C=2)	84%

(3) 最后的 mnist 是著名的手写数字数据集，它是一个被大量测试过的多分类任务数据集。测试了多组 SVM 参数，最终得到了表3，其中 SVM 最好的表现是 C=0.07 时的线性 SVM，准确率可以达到 94.56%；进行对比的是一个 35 个卷积层的神经网络 [7]。图5展示了各个参数下的 SVM 的测试结果。

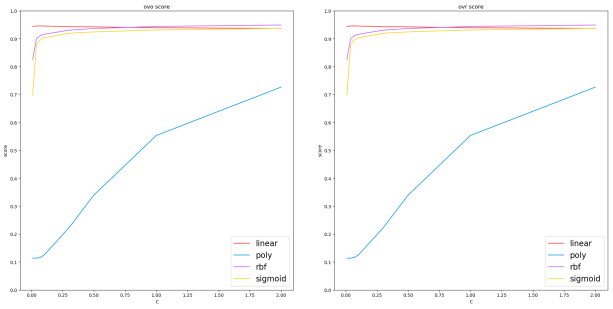


图 5: MNIST 数据集上的 SVM 表现。

表 3: MNIST 上的准确率比较。

方法	准确率
Ciresan et al. CVPR 2012	99.77%
linear-SVM(C=0.07)	94.56%

以上数据集实际上都并不大，因此除了这些数据集外，我选取了文档中没有的 Animals with Attributes2（AwA2）数据集作为拓展实验 [2]。该数据集也是 zero-shot learning 任务中比较经典的一个数据集，提供了 50 类动物的图片。这里用来做简单的分类学习，虽然该数据集的发布方提供了 Resnet50 抽取的深度学习特征，为了与 Resnet50 抽取的特征进行比较，我首先使用 Scale-invariant feature transform（SIFT）对图片抽取特征，结果如图6。图6中的绿色点为选取的尺度不变 local descriptor，矩形中的绿色向量附近各个区域的梯度方向。之后我使用了 Vector of aggregated descriptors（VLAD）方法对这些 local descriptor 进行编码。由于一个 local descriptor 是 128 维的，VLAD 中只保留一阶信息，我设置了 20 个聚类中心，把一张图编码成了 2560 维的向量，作为输入到 SVM 分类器的特征。SVM 使用的是 C 为 0.01 的线性 SVM，虽然由于这个数据量大只做了一组，但是传统方法是明显落后于深度神经网络的。表4中展示了分类准确率，我们可以看到，深度神经网络的抽取出的 feature 做分类，分类的准确率要远高于传统的方法的组合。同时，更为重要的是，深度学习的方法是端到端的过程，无需进行特征的抽取和编码，而使用传统的 SVM 我们需要人为设计出如 SIFT 结合 VLAD 这样的特征抽取系统，显然整体的效果是无法与深度神经网络相比的。

表 4: 对 AwA2 数据集的分类准确率。

方法	准确率
Resnet50	92.6%
SIFT+VLAD+SVM	26.1%

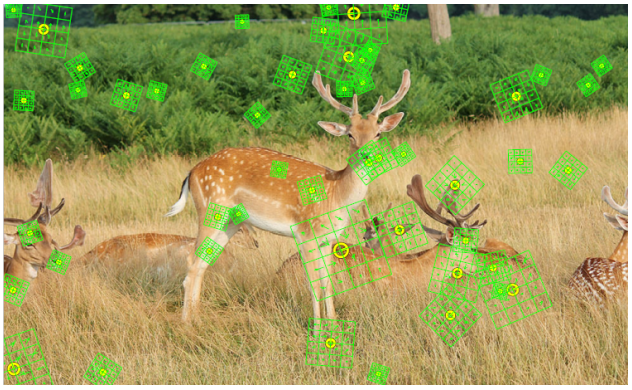
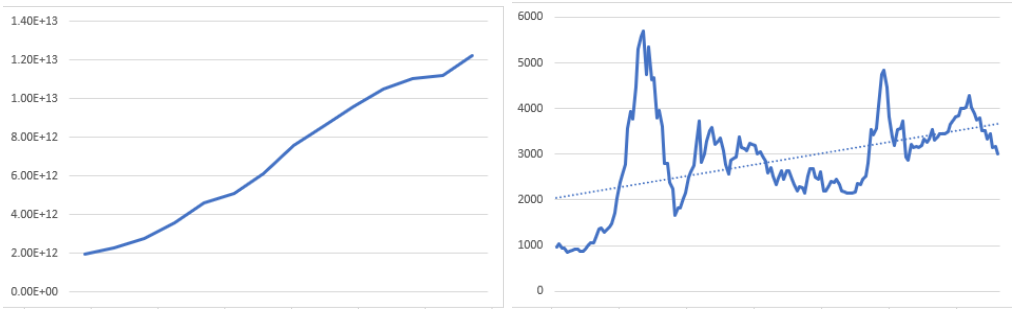


图 6: 使用 SIFT 抽取出的特征。

Solution.

Problem 2. Apply one causal discovery algorithm on a real world problem. You need to specify the details of the problem, collect the data by yourself or from a public website, briefly summarize what algorithm you use, and explain the results.

在现代经济学研究中，一个共识是“国家的股票指数与 GDP 有着极大的相关性”。如下图7所示，二者的变化存在一定时延，但是趋势相同。



(a) GDP。 (b) 沪深 300 指数。

图 7: 中国 GDP 和沪深 300 指数变化曲线。

股指在一定程度上折射出了企业盈利状况和投资人对市场的信心，因而股指可以与 GDP 共同反映出一个国家经济的基本情况。美国股市自 08 年经济危机以来，总体上保持了近十年的牛市，带给了特朗普对国内经济的信心，这也是他敢于打响中美贸易战的一个重要因素。

然而，虽然二者间的相关性是公认的，但是二者的因果关系却不是很明朗。虽然从图7中也可以看到，二者的变化有明显的先后关系，但是经济学家仍然无法确定是 GDP 上升导致了股指上升，还是股指上升导致了 GDP 增长。我在本次作业中选取了全球主要国家从 2005 年到 2017 年共 13 年间的 GDP 和股指的数据。

下表5为选取的国家及股指名称，数据来源于 [6, 5]：

表 5: 国家及对应选取的股指。

国家	股指
中国（China）	沪深 300
美国（United State）	标普 500
澳大利亚（Australia）	澳大利亚标普 200
德国（Germany）	德国 DAX
巴西（Brazil）	巴西 BOVESPA
印度（India）	印度孟买 30
英国（United Kingdom）	英国富时 100
加拿大（Canada）	加拿大标普

由于数据的规模并不大（共八个国家，每个国家 13 年），我尝试了多种分析方法以提高准确性，同时可以比较各个方法。

首先是经典的线性非高斯因果模型（Lingam），在因果推断中，一个难点在于如何在模型中建立有向性，一般的线性模型自变量与因变量可以交换地位，二者均可以互相表示，无法区分因果，Lingam 借鉴了独立成分分析（ICA）的思路，使用了非高斯分布的变量进行线性组合，从而与 ICA 类似确定唯一的表示 [9]。在给定数据集上，分析结果如图??，可以看到，对任意一个国家，均为股票市场变动是因，GDP 是果。

之后我又尝试了 igci 模型，igci 模型添加一个假设来使得因果关系是单向的，它假设输入的分布和因果机制之间是独立的，利用 information space 的正交性来区分因果 [8]。结果如图??，此时分析结果与 Lingam 模型出现了不一致之处，对于一些国家，GDP 变动是因，而一些国家 GDP 变动为果。这可能与国家的经济政策有关，也可能是样本集合太小，但是历史的股票指数数据很难收集，不便于扩充数据集。

最后我尝试了文档中推荐的 Tetrad 软件，该软件集成了大量因果推断算法，且拥有友好的图形界面。在图8中，我首先绘制了一幅图来表示 Structure equation model（SEM），同时使用了两种算法 FCI 和 FASK。FCI 算法歧视很类似著名的 PC 算法，增加了一些限制条件对可以表示因果关系的有向无环图进行搜索。它有两个阶段，第一个阶段先建立一个完整的无向图，执行一系列条件独立性测试，以消除无向图中两个相邻变量之间的边。在删除一些无关的边后，在该无向图的基础上，使用算法添加的条件，进行定向操作，尽可能多得定向。FASK 也是一个 PC 算法的变种，它搜索图中要用于定向的邻接节点 [1]。推断结果在图9和10中。可以看到 FCI 中的结果认为二者有公共的未观测到的干扰因子使得无法得到明确的结论，而 FASK 中则认为股票是因，GDP 是果。

综合多种算法的结果，我们可以初步得出结论，GDP 与股票指数是具有很强的相关性的，甚至因果关系也很明显，可以认为股指的涨跌可以导致 GDP 的涨跌，即可以根据股票市场对 GDP 作出预测。事实上，从经济学的角度来分析，这也是合理的，当股票指数增长时，可以看出企业盈利能力明显增强，投资人信心日益增长，同时政府也可能采取了宽松的货币政策导致市场上的流动的货币数量增加，这一切都可以视作 GDP 增长的信号。反之，当市场低迷时，国民生产总值的增长也就如空中楼阁，无从谈起。

Solution.

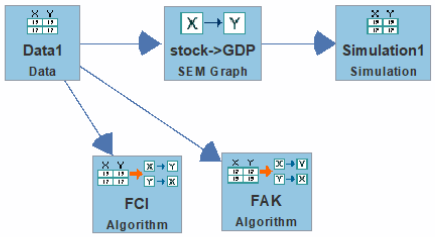


图 8: Tetrad 绘图。

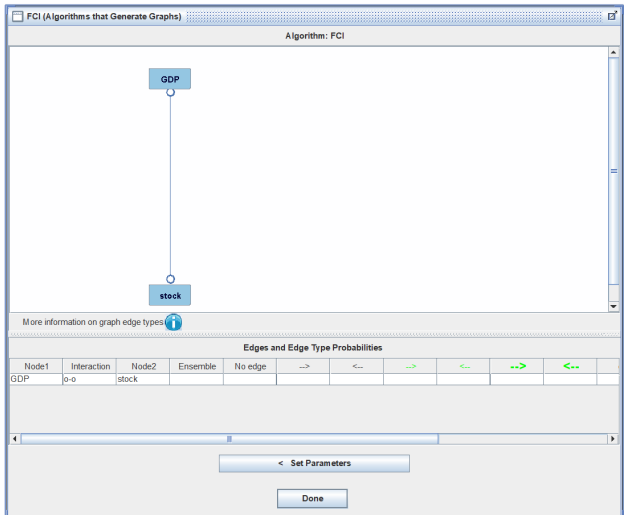


图 9: FCI 结果。

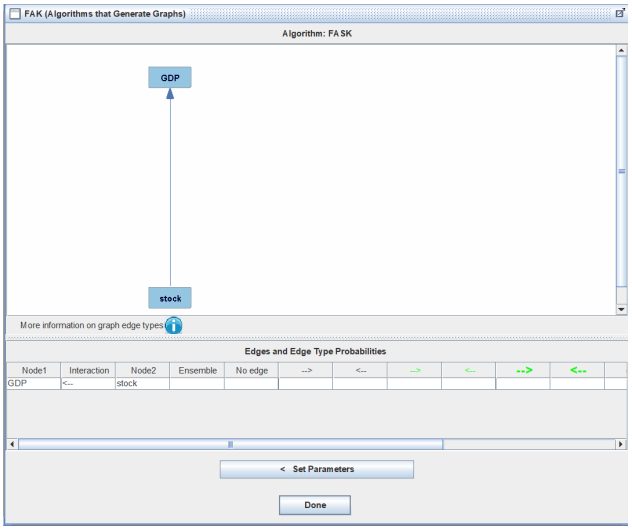


图 10: FASK 结果。

参考文献

[1] http://www.phil.cmu.edu/tetrad/.

[2] Awa2. https://cvml.ist.ac.at/AwA2/, 2019. Accessed June, 2019.

[3] Iris, colon cancer, mnist. https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/, 2019. Accessed June, 2019.

[4] Mnist. http://deeplearning.net/datasets/, 2019. Accessed June, 2019.

[5] World bank. https://databank.worldbank.org/data, 2019. Accessed June, 2019.

[6] 英为财情. https://cn.investing.com, 2019. Accessed June, 2019.

[7] Dan Cireşan, Ueli Meier, and Juergen Schmidhuber. Multi-column deep neural networks for image classification. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 02 2012.

[8] Dominik Janzing, Bastian Steudel, Naji Shajarisales, and Bernhard Schölkopf. Justifying Information-Geometric Causal Inference, pages 253–265. 08 2015.

[9] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. Applied Informatics, 3, 12 2016.