数据挖掘与信息检索实验室

Data Mining & Information Retrieval Laboratory

数据科学实战

——概貌介绍

蔡瑞初

cairuichu@gmail.com

Outline

- ☑ 机器学习专题
- ☑ 因果专题
- ☑ 大数据平台专题
- ☑ 深度学习专题
- ☑ 自然语言处理专题
- ☑ 社交网络挖掘专题
- ☑ 总结分享专题

机器学习专题

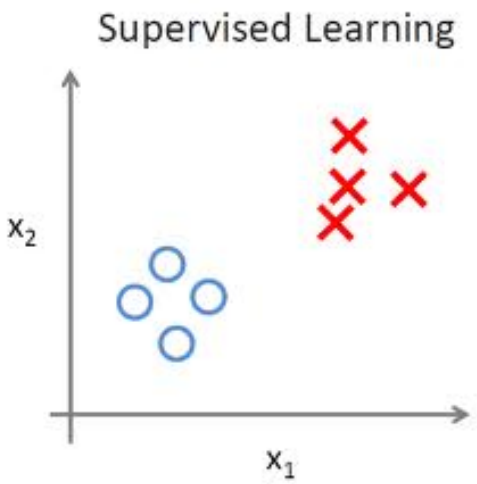
☑ 根据**任务**分类

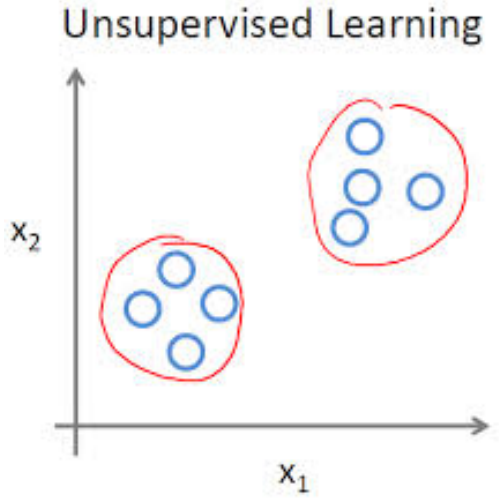
- 有监督学习
- 无监督学习
- 增强学习

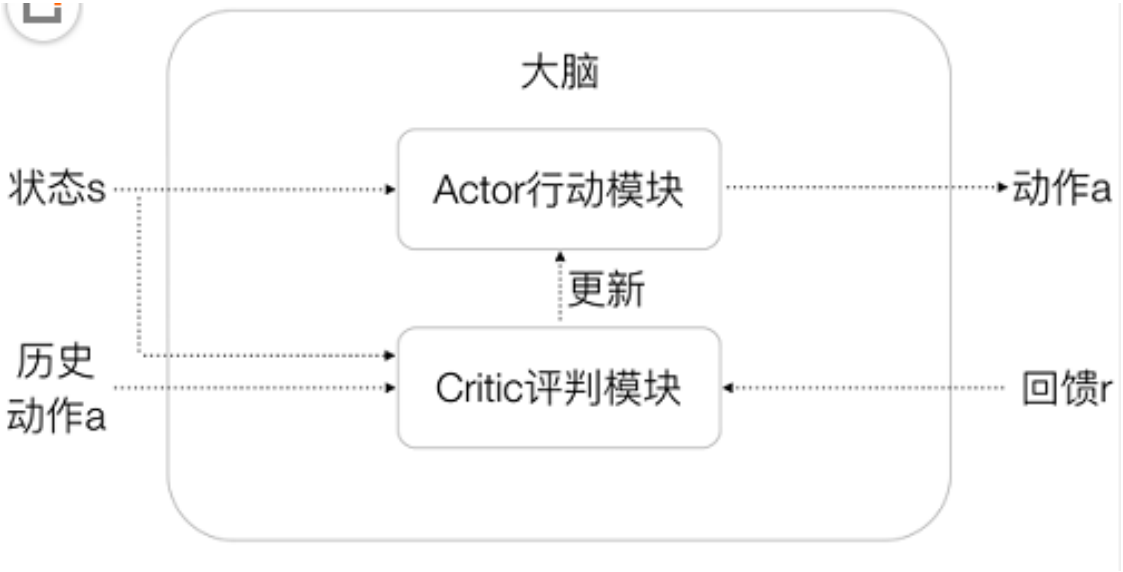
☑ 根据**模型**分类

- 聚类： k-means（快速聚类），SOM（自组织映射）,Ward（最小方差聚类），稀疏表达
- 分类：决策树，神经网络，深度学习，SVM
- 关联： Apriori，fp_growth
- 回归：GLM（广义线性回归），Logistic（逻辑回归）

机器学习专题


Supervised Learning
x_2
x_1


Unsupervised Learning
x_2
x_1


大脑
状态s
Actor行动模块
动作a
更新
历史
动作a
Critic评判模块
回馈r

机器学习专题:分类

☑ Decision Tree

- Information Gain
- Overfitting
- Ref1： https://en.wikipedia.org/wiki/Decision_tree,
- Ref2： Utgoff, P. E. (1989). Incremental induction of decision trees. Machine learning, 4(2), 161-186
- Extension：Deep Forest https://arxiv.org/abs/1702.08835

☑ SVM

- Max Margin
- Kernel Trick
- Ref1：https://en.wikipedia.org/wiki/Support_vector_machine
- Extension： Tensor，Zhifeng Hao, Lifang He, Bingqian Chen, Xiaowei Yang:
- A Linear Support Higher-Order Tensor Machine for Classification. IEEE Trans. Image Processing 22(7): 2911-2920 (2013)

机器学习专题:聚类

☑ Kmeans

- Metric
- Complexity
- ref1: https://en.wikipedia.org/wiki/K-means_clustering

☑ Hierarchical based：

- the meaning of a cluster
- Ref1: https://en.wikipedia.org/wiki/Hierarchical_clustering
- Extension: Ruichu Cai, Zhenjie Zhang, Anthony Tung, Chenyun Dai, Zhifeng Hao. A General Framework of Hierarchical Clustering and Its Applications[J]. Information Science, 2014: 272, 29-48

☑ Density based

- The shape of cluster
- Ref1: https://en.wikipedia.org/wiki/DBSCAN

机器学习专题:其它

☑ Association rule

- Complexity/Lattice
- ref1: https://en.wikipedia.org/wiki/Apriori_algorithm
- Extension: Ruichu Cai, Tung K.H. Anthony, Zhifeng Hao, Zhenjie Zhang. What is Unequal among the Equals? Ranking Equivalent Rules from Gene Expression Data. IEEE Transactions on Knowledge and Data Engineering. 2011;23(11):1735-1747

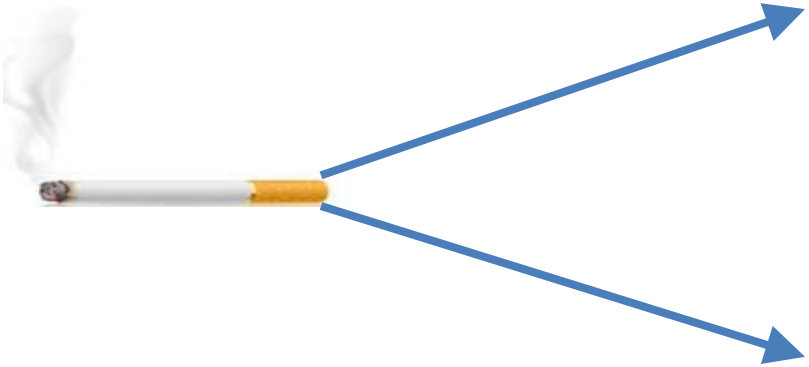
☑ Active Learning/Reinforcement Learning

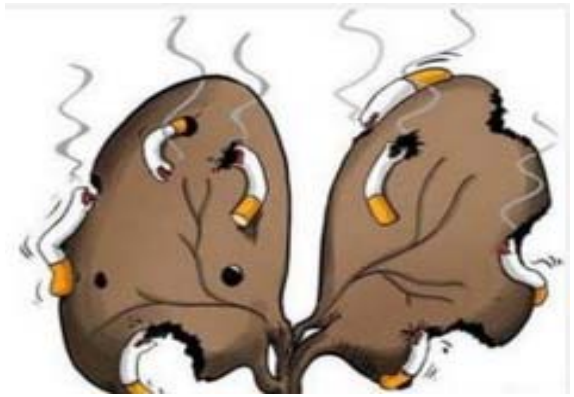
- Ref1: http://www.nature.com/nature/journal/v529/n7587/nature16961/metrics/newsMastering
- the game of Go with deep neural networks and tree search
- Ref2: https://arxiv.org/abs/1809.09095
- On Reinforcement Learning for Full-length Game of StarCraft

因果专题

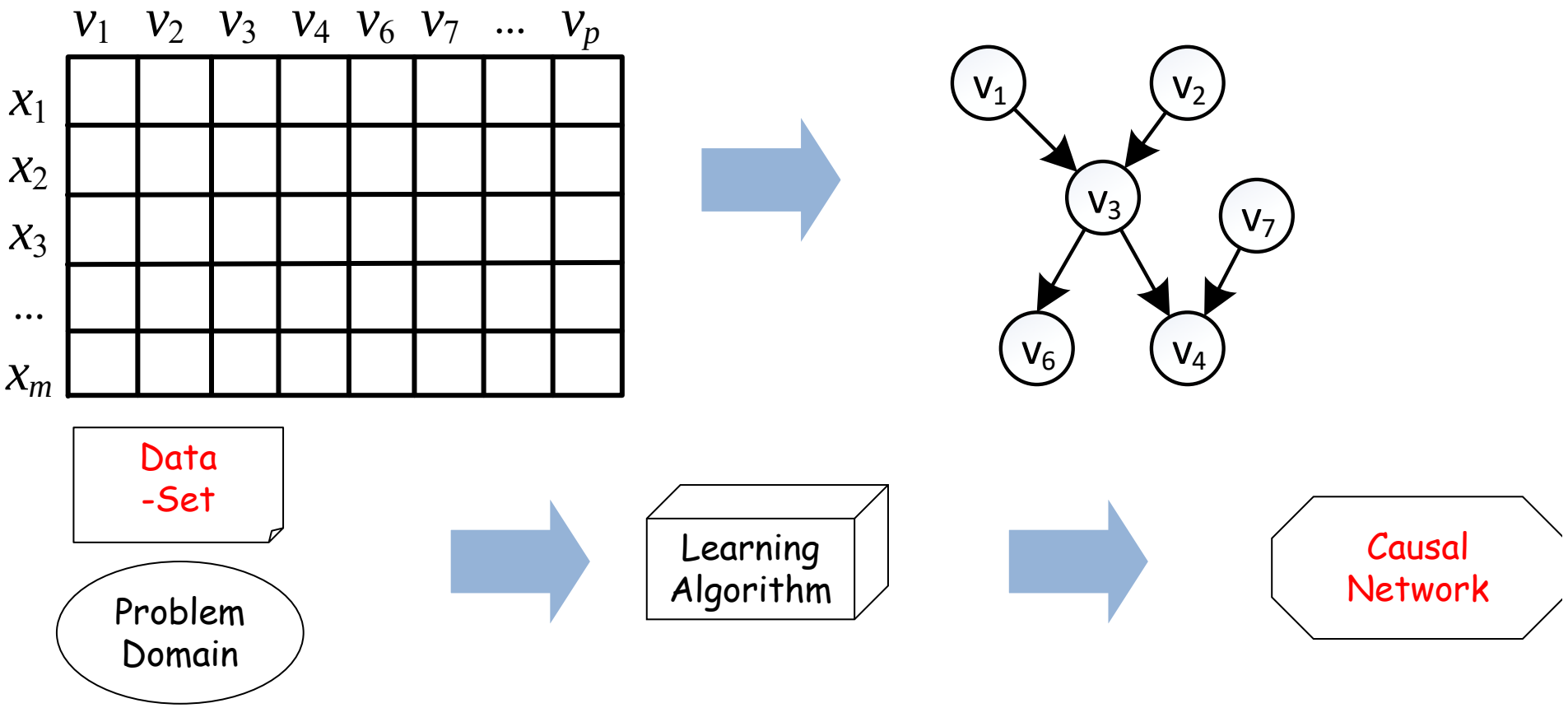
- ☑ 基于约束的方法
- ☑ 基于结构方程模型的方法
- ☑ 似然度的方法

因果专题





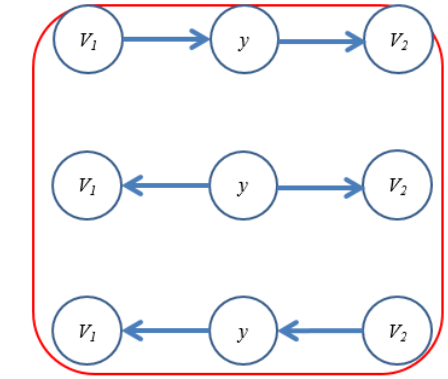
因果专题

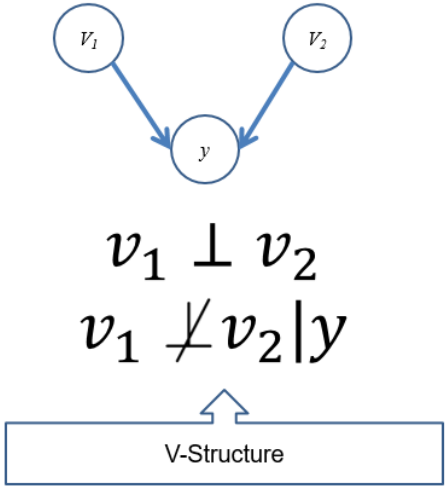

v_1 v_2 v_3 v_4 v_6 v_7 ... v_p
x_1
x_2
x_3
...
x_m
v1
v2
v3
v7
v6
v4
Data -Set
Problem Domain
Learning Algorithm
Causal Network

因果专题:基于约束的方法

☑ IC & PC

- V-structure
- Independence test
- Ref1： Pearl J, Verma T. A theory of inferred causation[M]. San Mateo, Morgan
- Kaufmann, 1991.,
- Ref2: Ruichu Cai, Zhenjie Zhang, Zhifeng Hao. Causal Gene Identification Using Combinatorial V-Structure Search, Neural Networks. 2013;43:63-71(SCI二区)

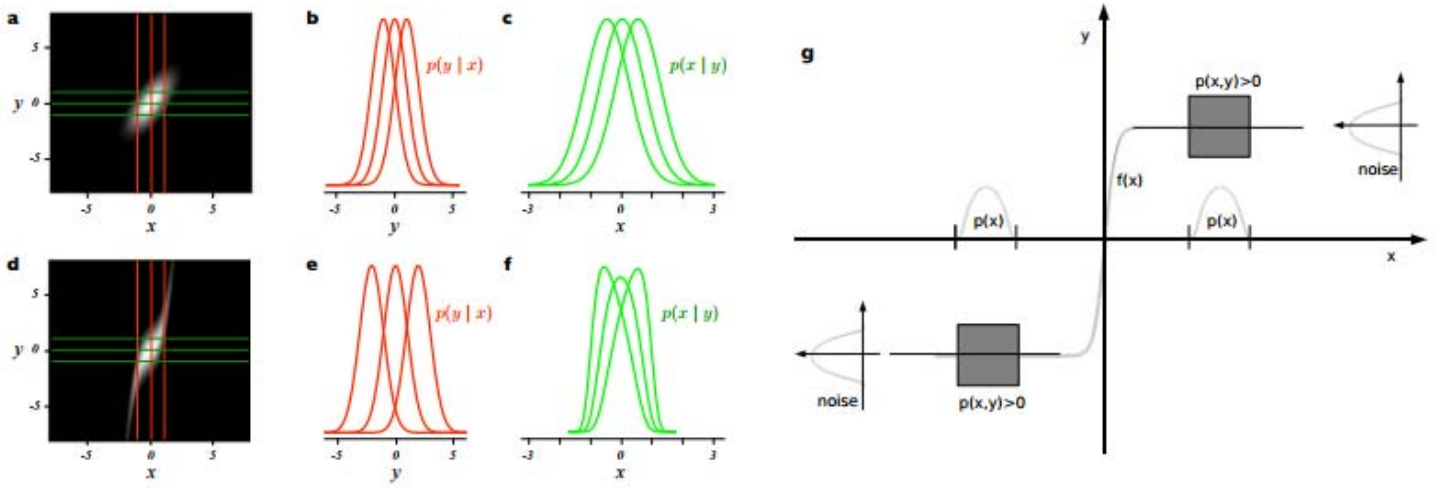




因果专题:基于结构方程的方法

☑ ANM

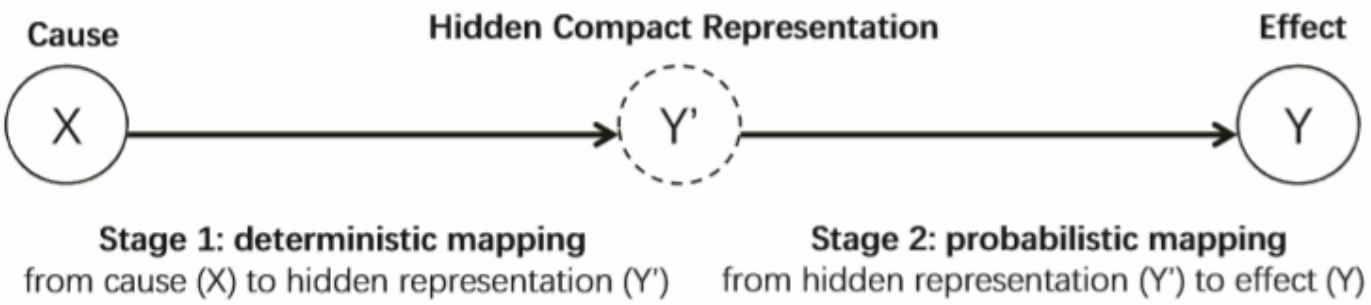
- Direction inference
- Ref1： Hoyer P O, Janzing D, Mooij J M, et al. Nonlinear causal discovery with additive noise models[C]//Advances in neural information processing systems. 2009: 689-696.
- Ref2： Janzing D, Mooij J, Zhang K, et al. Information-geometric approach to inferring causal directions[J]. Artificial Intelligence, 2012, 182: 1-31.
- Extension： Ruichu Cai, Zhenjie Zhang, Zhifeng Hao. SADA: A General Framework to Support Robust Causation Discovery, ICML. 2013(CCF A类会议).



因果专题:似然度的方法

☑ Likelihood & MDL

- Ref1：Tsamardinos I, Brown L E, Aliferis C F. The max-min hill-climbing Bayesian network structure learning algorithm[J]. Machine learning, 2006, 65(1): 31-78.
- Ref2： Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, Zhifeng Hao. Causal Inference on Discrete Data using Hidden Compact Representation. NIPS,2018.



$$L^*(M;D) = \prod_{i=1}^{m} P(X = x_i)P(Y = y_i|Y' = f(x_i)) - \frac{d}{2}\log(m)$$

$$= \sum_{x} n_x \log\left(\frac{n_x}{\sum_x n_x}\right) + \sum_{y'} \sum_{y} n_{y',y} \log\left(\frac{n_{y',y}}{\sum_y n_{y'.y}}\right) - \frac{d}{2}\log(m)$$

因果专题:基于因果关系的机器学习

☑ Semi supervise & causality

- B. Sch¨olkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *ICML 2012*,, 2012.
- Kilbertus N, Parascandolo G, Schölkopf B. Generalization in anti-causal learning[J]. arXiv preprint arXiv:1812.00524, 2018.

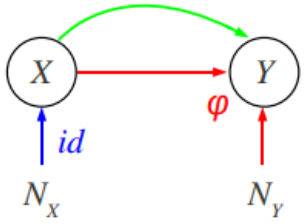


Figure 2. Predicting effect Y from cause X.

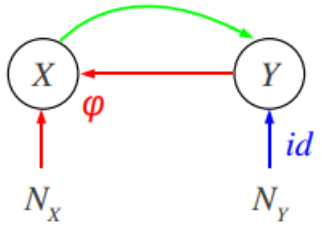
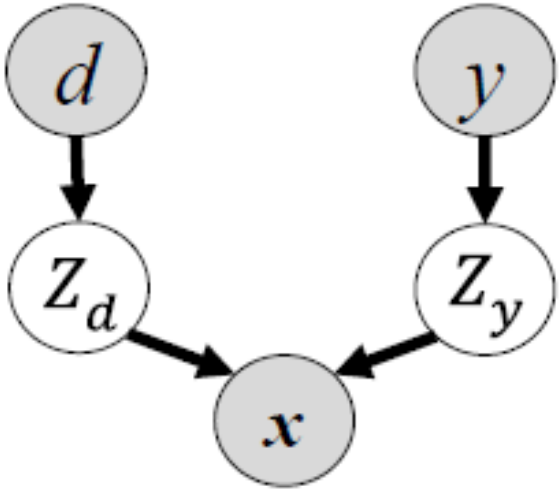


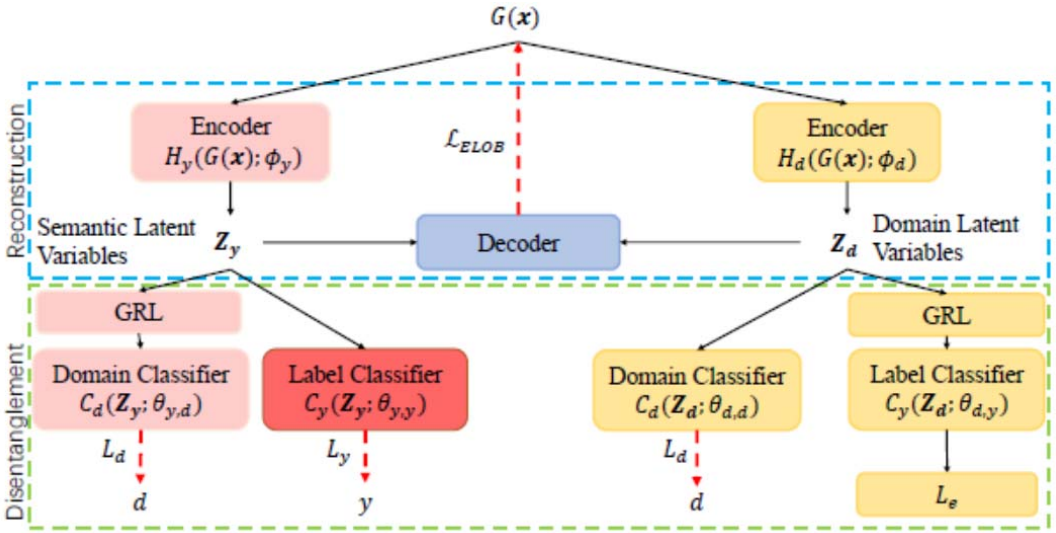
Figure 3. Predicting cause Y from effect X.

因果专题:基于因果关系的机器学习

☑ Domain Adaption & causality

- Zhang K, Schölkopf B, Muandet K, et al. Domain adaptation under target and conditional shift[C]//International Conference on Machine Learning. 2013: 819-827.
- Ruichu Cai, Zijian Li et al. Learning Disentangled Semantic Representation for Domain Adaptation. IJCAI 2019

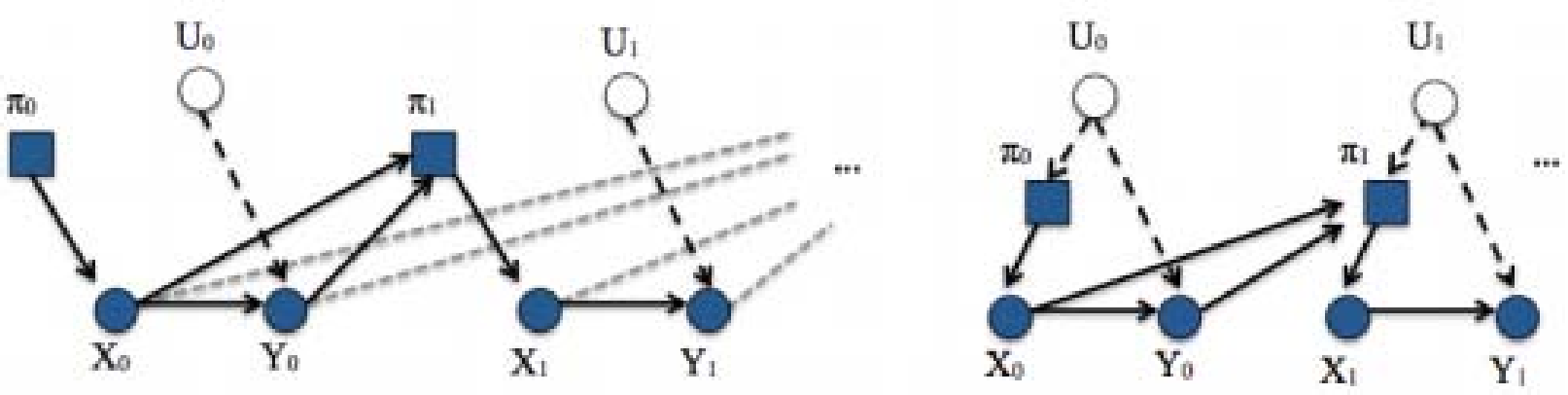




因果专题:基于因果关系的机器学习

☑ Reinforcement Learning & causality

- Lattimore F, Lattimore T, Reid M D. Causal bandits: Learning good interventions via causal inference[C]//NIPS. 2016: 1181-1189.
- Bareinboim E, Forney A, Pearl J. Bandits with unobserved confounders: A causal approach[C]//NIPS. 2015: 1342-1350.



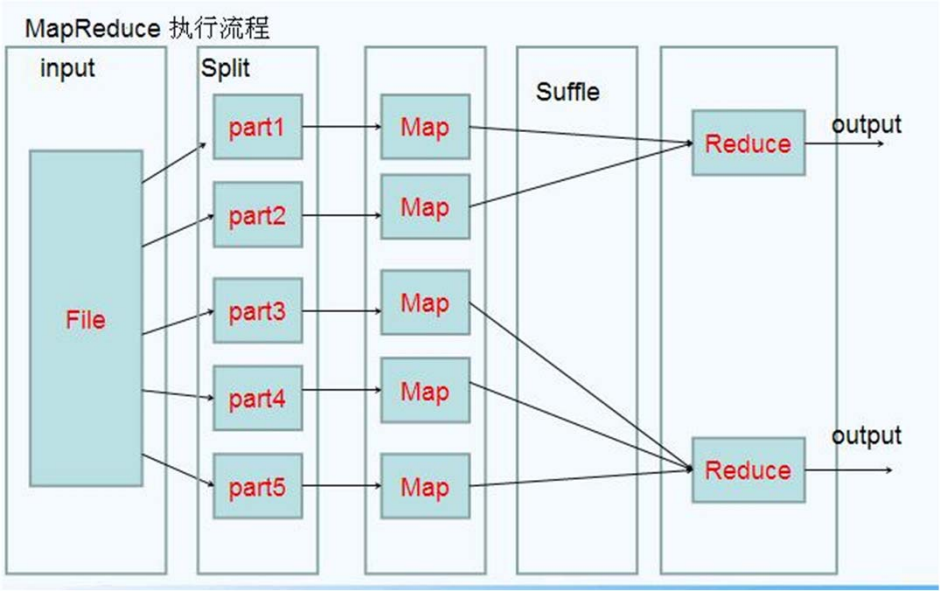
大数据平台专题

- ☑ Batch
- ☑ In Memory
- ☑ Stream

大数据平台:Batch

☑ Hadoop

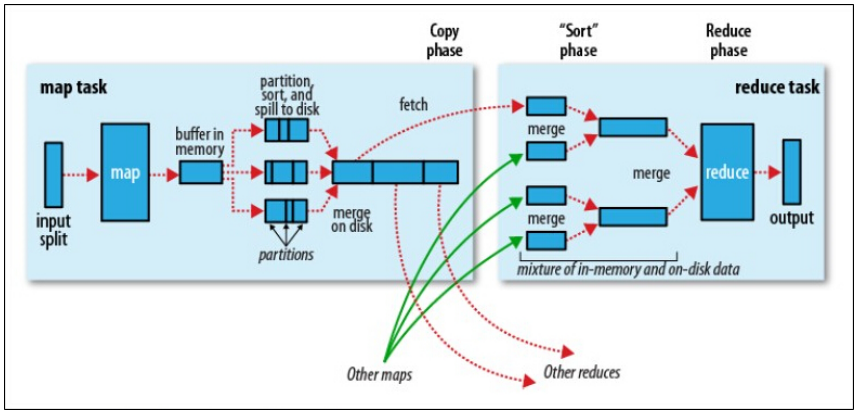
- GFS & Mapreduce
- Ref1： Map-reduce Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.
- Ref2： Ghemawat S, Gobioff H, Leung S T. The Google file system[C]//ACM SIGOPS operating systems review. ACM, 2003, 37(5): 29-43.

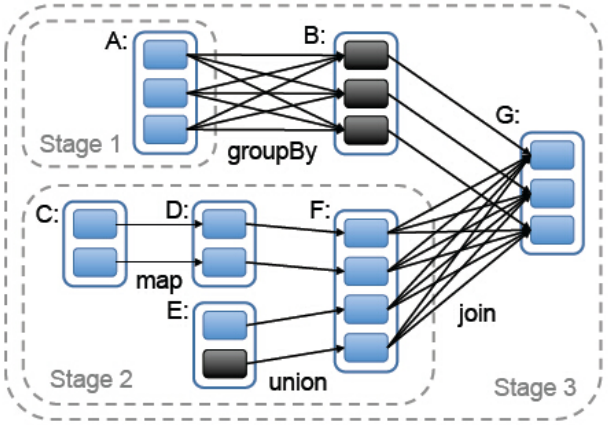


大数据平台: In Memory

☑ Spark

- RDD & Graph
- Ref1： http://spark.apache.org/

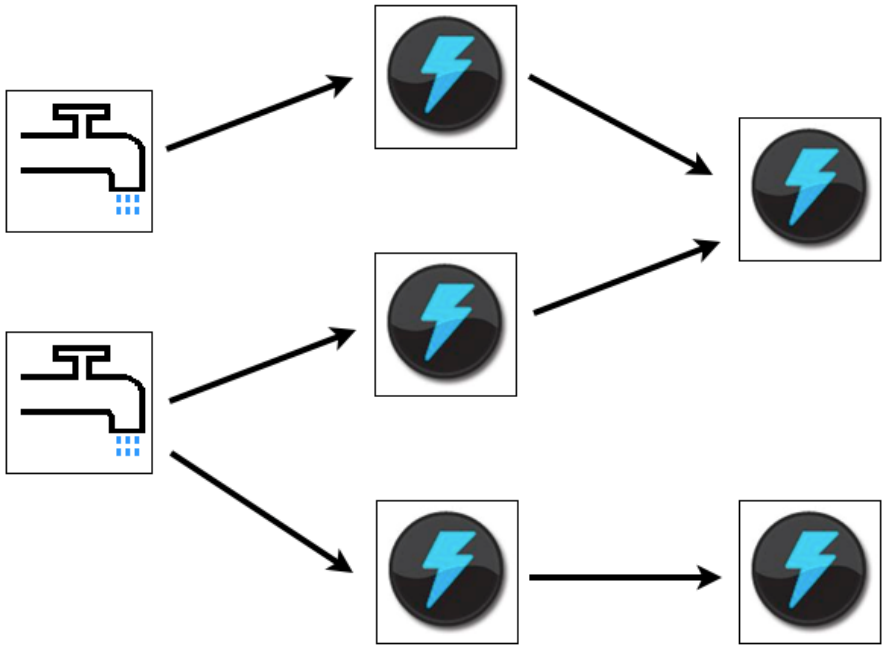




大数据平台: Stream

☑ Storm & Heron

- Bort & Graph & stream
- Ref1： https:// storm.apache.org/
- Ref2： https://twitter.github.io/heron
- Ref3： Toshniwal A, Taneja S, Shukla A, et al. Storm@ twitter[C]//Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014: 147-156.



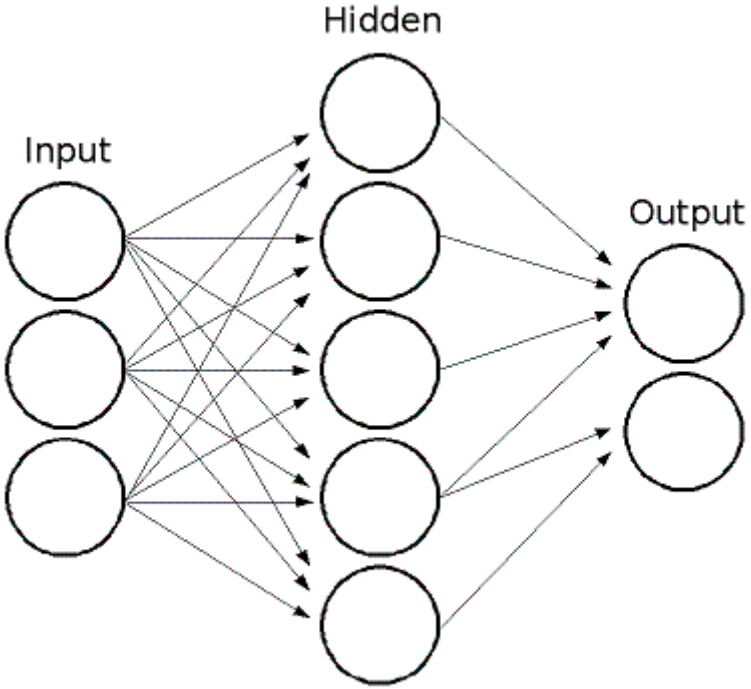
深度学习专题

- ☑ MLP & BP
- ☑ CNN
- ☑ RNN
- ☑ GAN
- ☑ GN

深度学习: MLP & BP

☑ MLP & BP

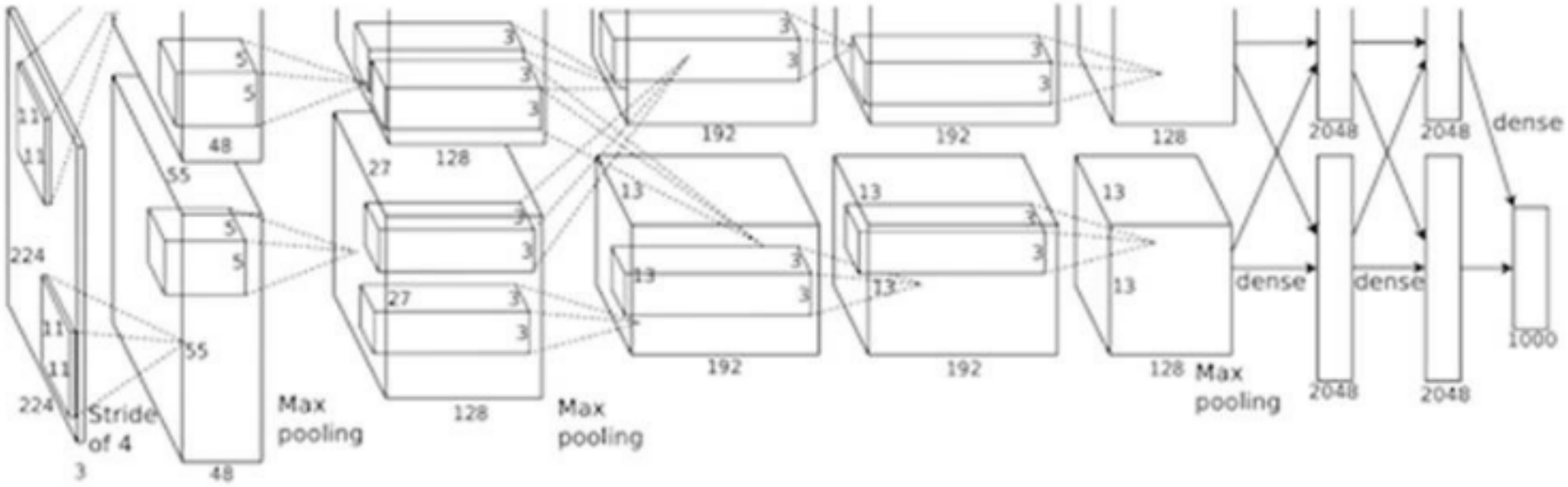
- Ref1：机器学习 by tom



深度学习: CNN

☑ alexnet

- Ref1： Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
- Ref2: tensorflow, caffe, theano



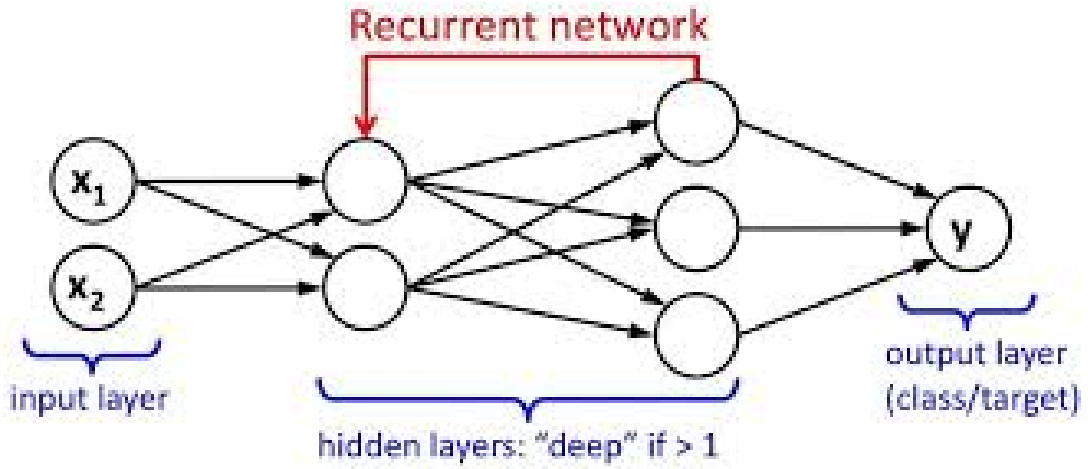
深度学习: RNN

☑ LSTM

- Ref 1： Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

☑ Hessian Free RNN

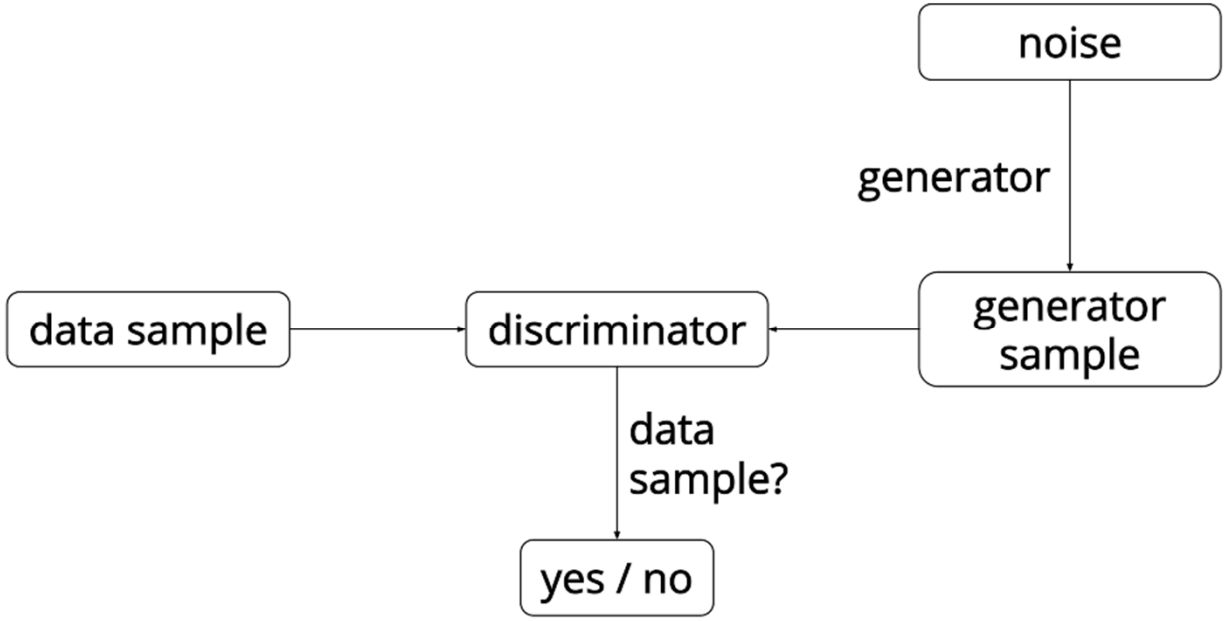
- Ref 1: Martens J, Sutskever I. Learning recurrent neural networks with hessian-free optimization[C]//Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011: 1033-1040.



深度学习: GAN

☑ Generative Adversarial Nets

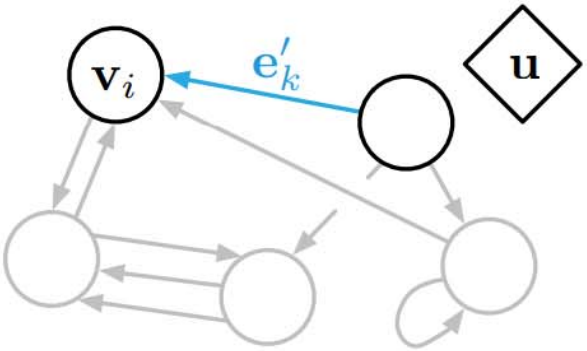
- Ref1： Goodfellow, Ian J.; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014). "Generative Adversarial Networks". arXiv:1406.2661Freely accessible [stat.ML].



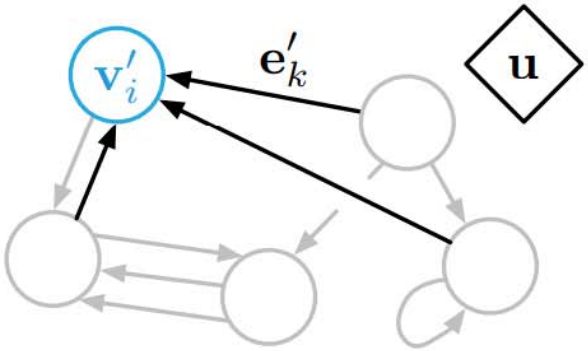
深度学习: GNN

☑ Graph Neural Net

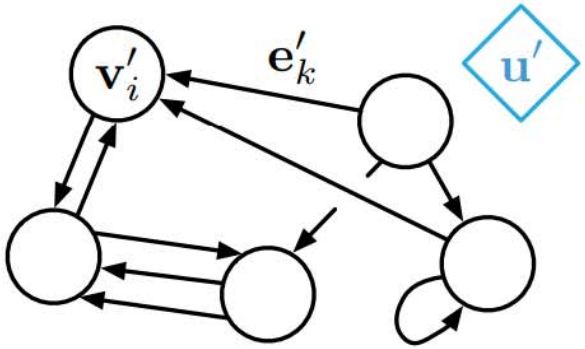
- Ref1：Battaglia P W, Hamrick J B, Bapst V, et al. Relational inductive biases, deep learning, and graph networks[J]. arXiv preprint arXiv:1806.01261,
- Ref2：http://tkipf.github.io/graph-convolutional-networks/



(a) Edge update



(b) Node update



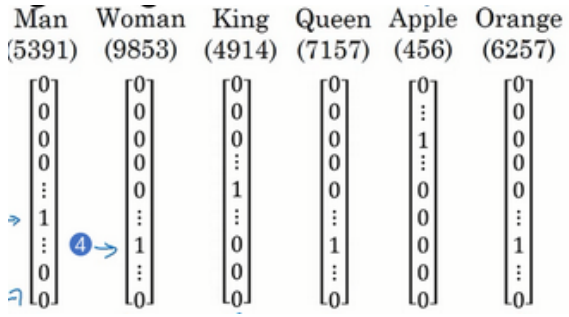
(c) Global update

自然语言处理专题

- ☑ 表示问题
- ☑ 序列标注问题
- ☑ 序列生成问题
- ☑ 序列匹配问题

自然语言处理专题：表示问题

☑ Onehot & TFIDF



☑ Word2vec & doc2vec

- Similarity & presentation
- Ref1： Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
- Ref2： Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
- Ref3: Transformer： Attention Is All You Need

  https://arxiv.org/pdf/1706.03762.pdf

Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
-1	1	-0.95	0.97	0.00	0.01
0.01	0.02	0.93	0.95	-0.01	0.00
0.03	0.02	0.7	0.69	0.03	-0.02
0.09	0.01	0.02	0.01	0.95	0.97
⋮	⋮	⋮	⋮	⋮	⋮

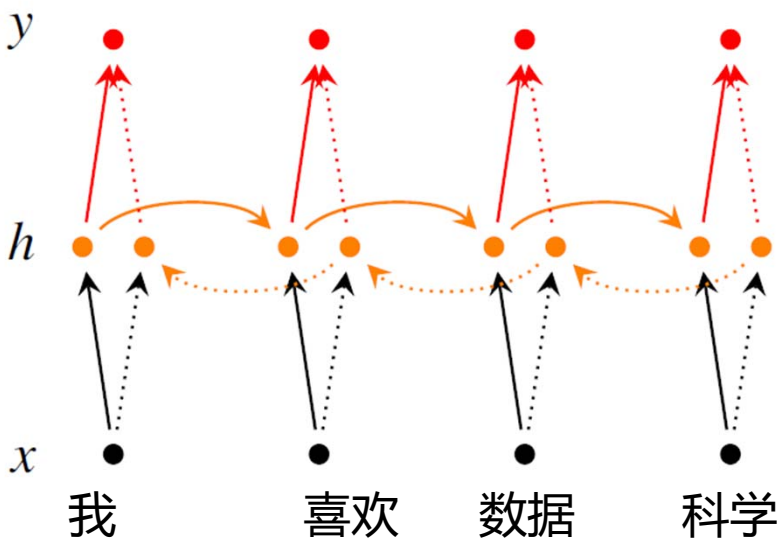
自然语言处理专题：序列标注

☑ CRF

- Ref1: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data
  https://faculty.cs.byu.edu/~ringger/CS479/papers/LaffertyMcCallumPereira-CRF-icml01.pdf

☑ RNN

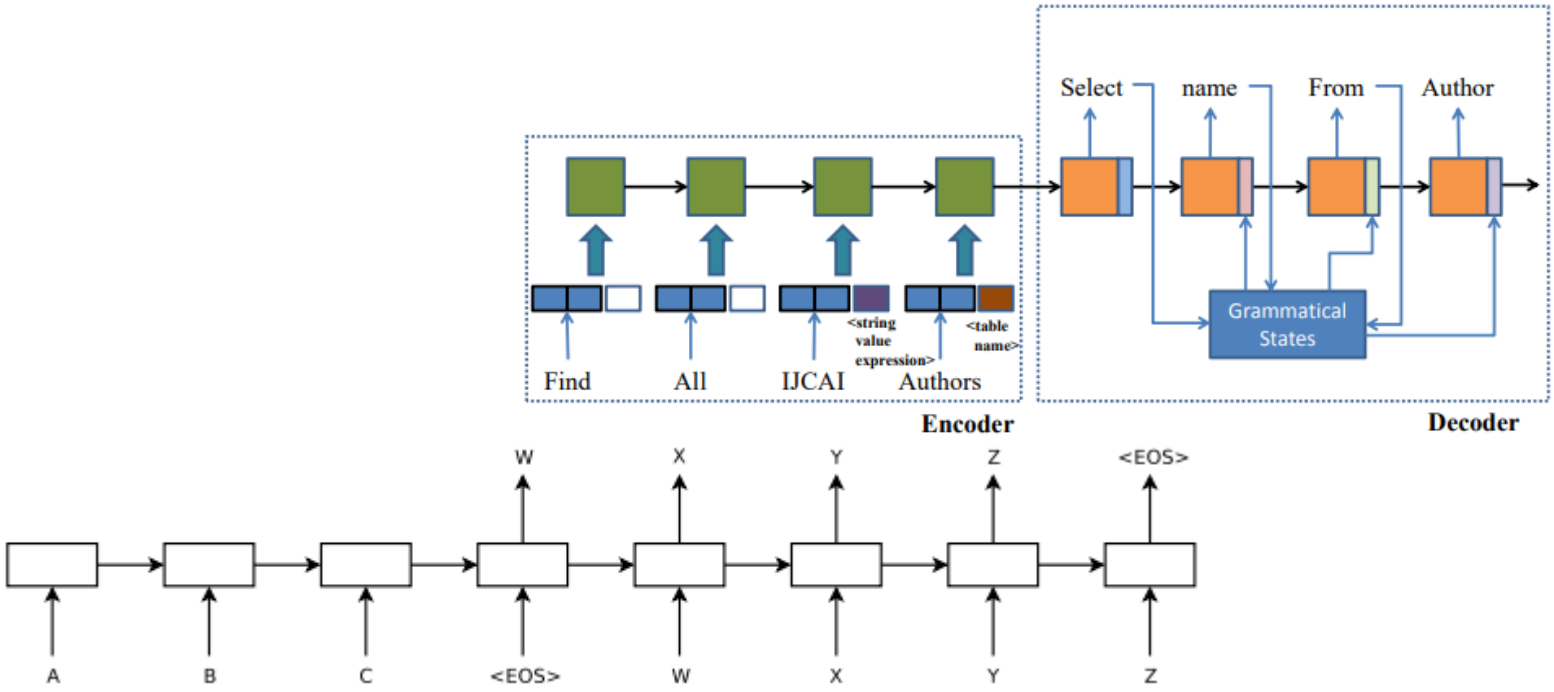
- dependence
- Ref1: Bidirectional LSTM-CRF Models for Sequence Tagging
- https://arxiv.o



自然语言处理专题：翻译问题

☑ Translation

- Prior & Representation
- Sequence to Sequence Learning with Neural Networks
  https://arxiv.org/pdf/1409.3215.pdf
- An encoder-decoder framework translating natural language to database queries
  https://arxiv.org/pdf/1711.06061.pdf



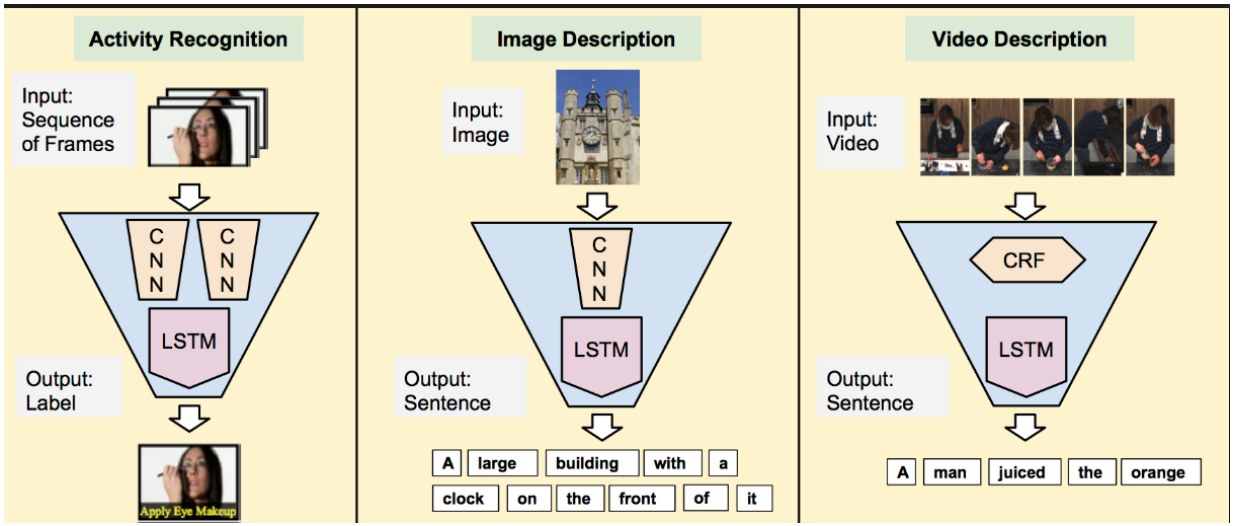
自然语言处理专题：序列生成

☑ Chatting robot

- memory & knowledge
- Ref1： End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning https://arxiv.org/pdf/1606.01269.pdf

☑ Show & tell

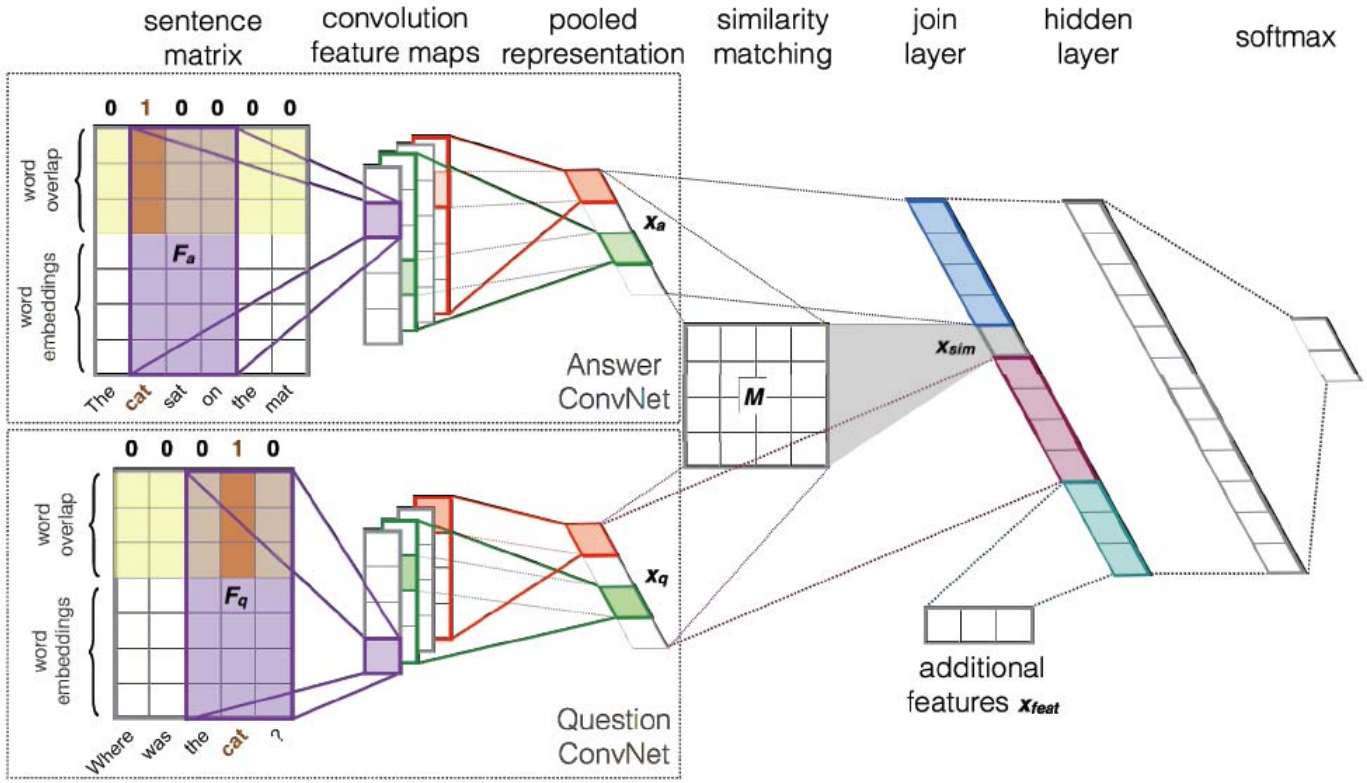
- Ref1： show and tell : Long-term Recurrent Convolutional Networks for Visual Recognition and Description
- http://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Donahue_Long-Term_Recurrent_Convolutional_2015_CVPR_paper.pdf



自然语言处理专题：序列匹配

☑ Q/A

- Metric learning
- Ref1： Severyn A, Moschitti A. Modeling relational information in question-answer pairs with convolutional neural networks[J]. arXiv preprint arXiv:1604.01178, 2016.



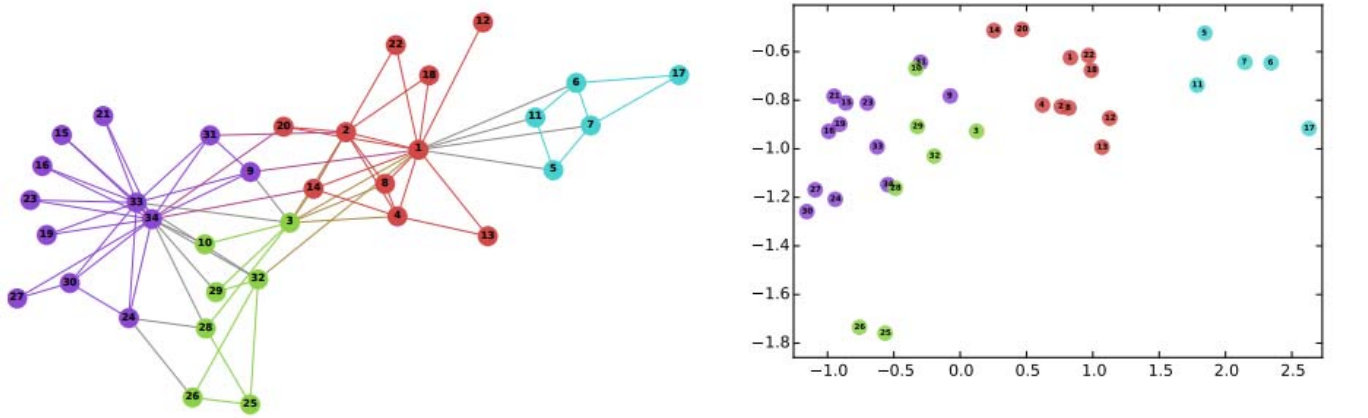
社交网络专题

- ☑ 表示问题
- ☑ 网络结构特性
- ☑ 信息传播问题
- ☑ 用户行为分析

社交网络主题：表示问题

☑ Graph Embbeding

- Metric & sampling
- Ref 1: Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." *SIGKDD*. ACM, 2014.
- Ref 2: Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." SIGKDD ACM, 2016.
- Fu X, Zhang J, Meng Z, et al. MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding[C]//WWW. 2020: 2331-2341.

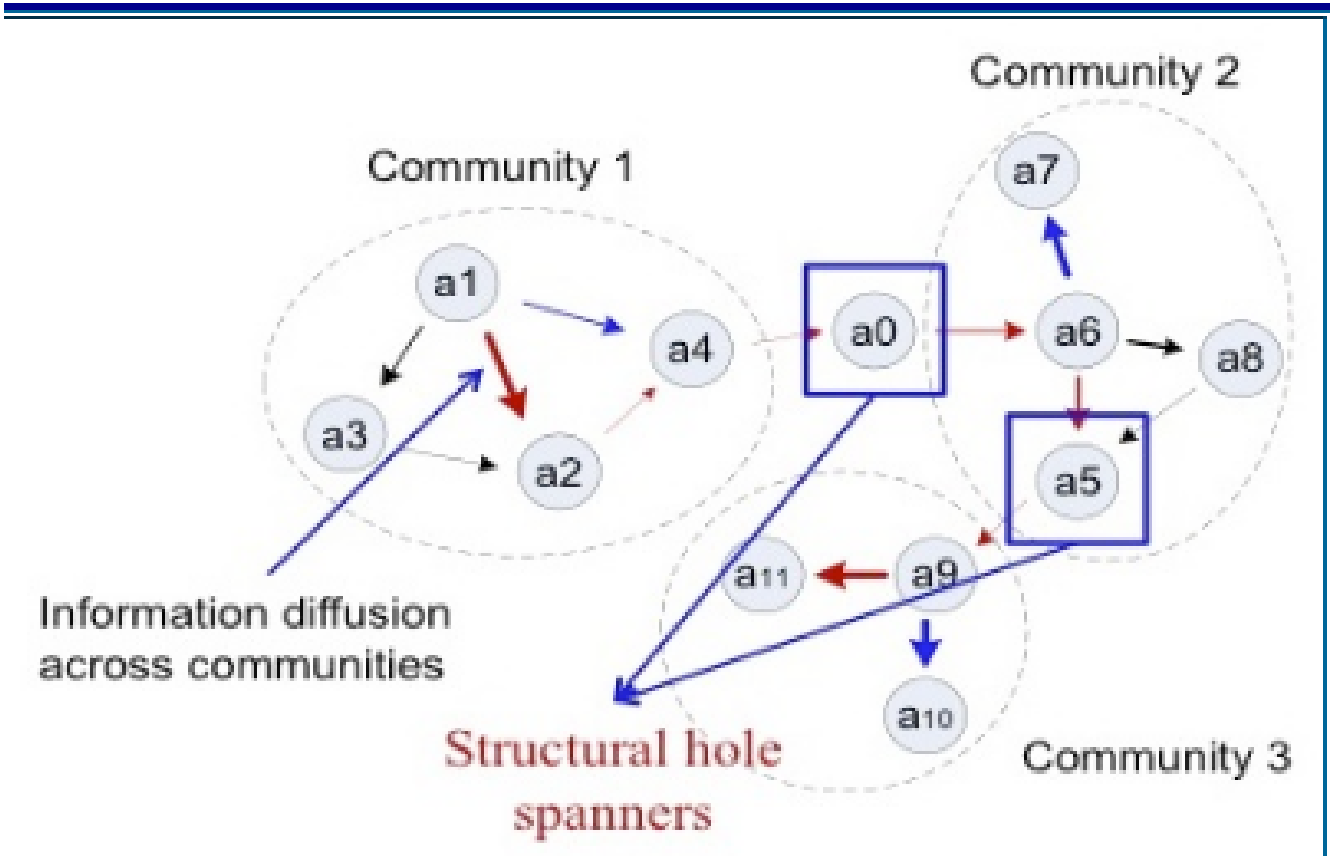


(a) Input: Karate Graph (b) Output: Representation

社交网络主题：网络结构特性

☑ Structure hole

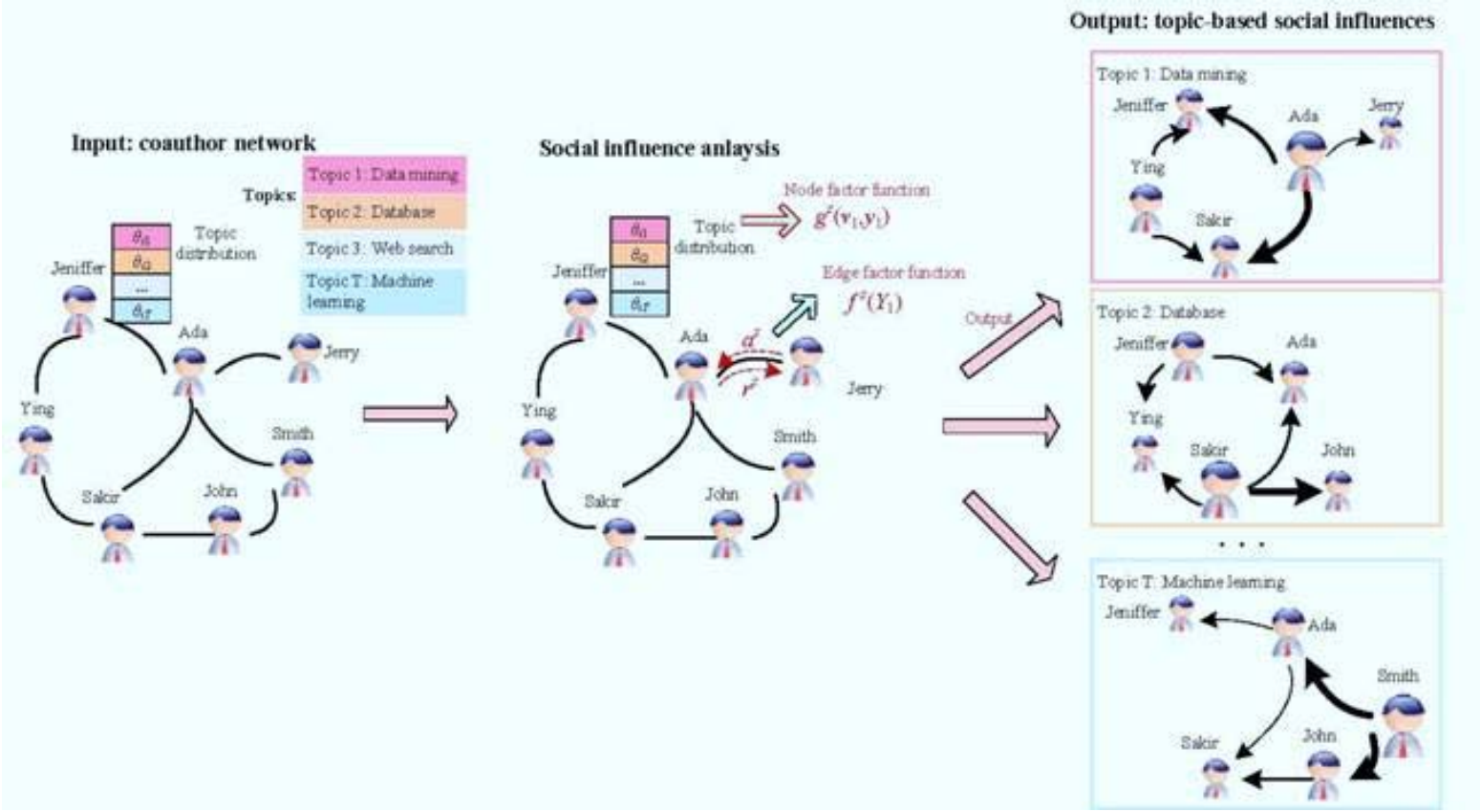
- Ref 1: Lou T, Tang J. Mining structural hole spanners through information diffusion in social networks[C]//Proceedings of the 22nd international conference on World Wide Web. ACM, 2013: 825-836. RBort & Graph



社交网络主题：信息传播问题

☑ Influence

- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2009).



社交网络主题：用户行为分析

☑ Action

- Ref1： Understanding Behaviors that Lead to Purchasing: A Case Study of Pinterest. C. Lo, D. Frankowski, J. Leskovec. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.
- Ref2： Ruichu Cai, Zhenjie Zhang, Zhifeng Hao, Marianne Winslett. Understanding Social Causalities Behind Human Action Sequences. IEEE Transactions on Neural Networks and Learning Systems. 2016.(SCI一区) [2016]

具体要求

☑ 每个专题 1-2次课

- 根据报名情况调整

☑ 每次课程　15*6

- 15 分钟基础知识介绍
- 15*4 四个主要专题
- 15 分钟实践成果

☑ 每个专题设立1组长

- 协调每位同学报告内容
- 小组排练计时
- 小组内可以分工，比如准备ppt，上台报告，算法实现，报告ppt说明分工及贡献

☑ 自愿选分组，课后到维杰处报名，到时会适当调整。

Thank you!

http://www.dmirlab.com/