

データからの ベイジアンネットワーク 構造推定

2016/04/09

@mabonki0725

自己紹介

≡ データ分析を15年以上経験

実用上は精度より頑健性のあるモデル作りが大切

≡ 統計数理研究所 研究生 機械学習

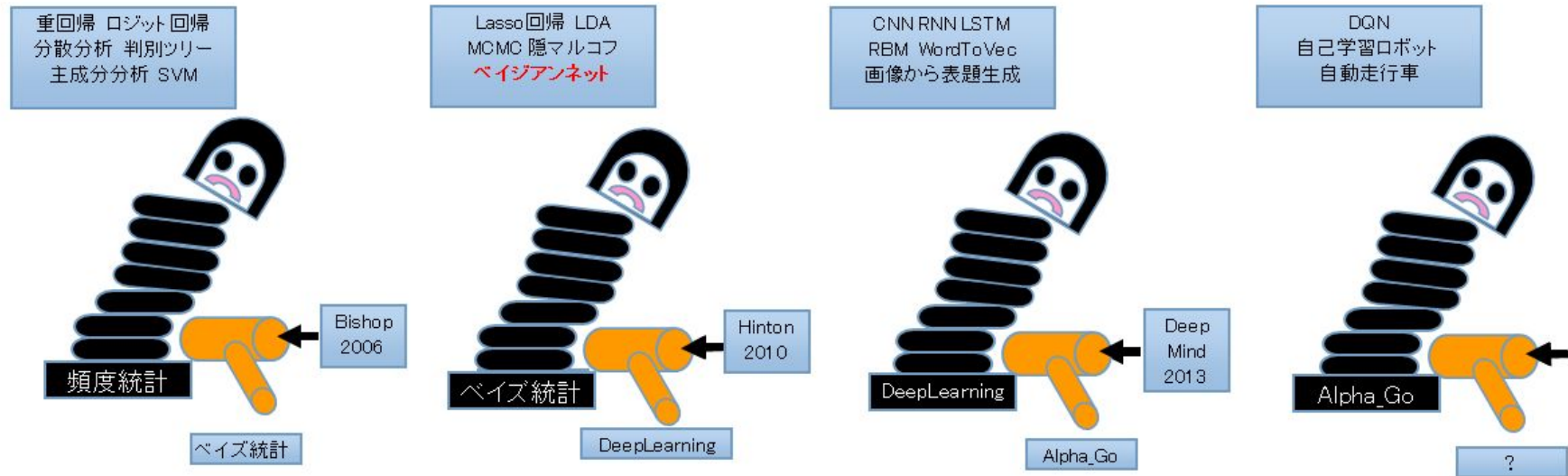
≡ 都立技術大学院 学生 ロボット製作したい

≡ 確率モデルのプログラムに興味(実用的でない)

■ PRMLの主要な確率モデルは殆ど作成

■ 最近はDeepLearningモデルCNN RNN LSTMを作成

機械学習の革新過程



革新的な発表によって、既存の技術が破綻していく(達磨落としの過程)

新しい技術をCatchUpして、その延長上に既存技術を生かすしかない

データから疎な有向グラフ推定

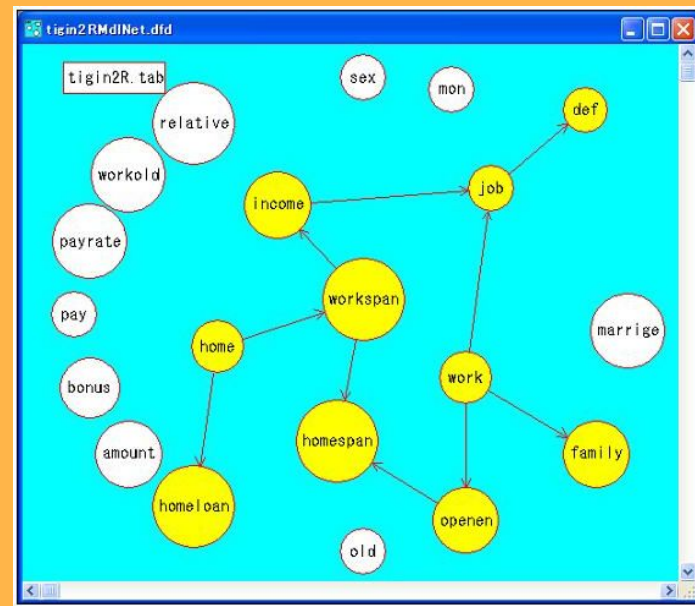
Data

Inference

Bayesian Net

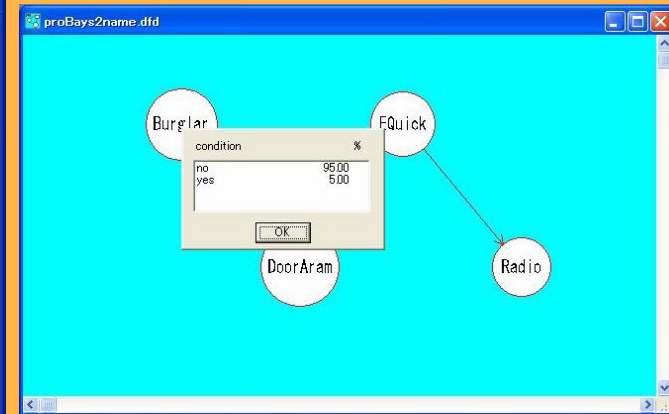
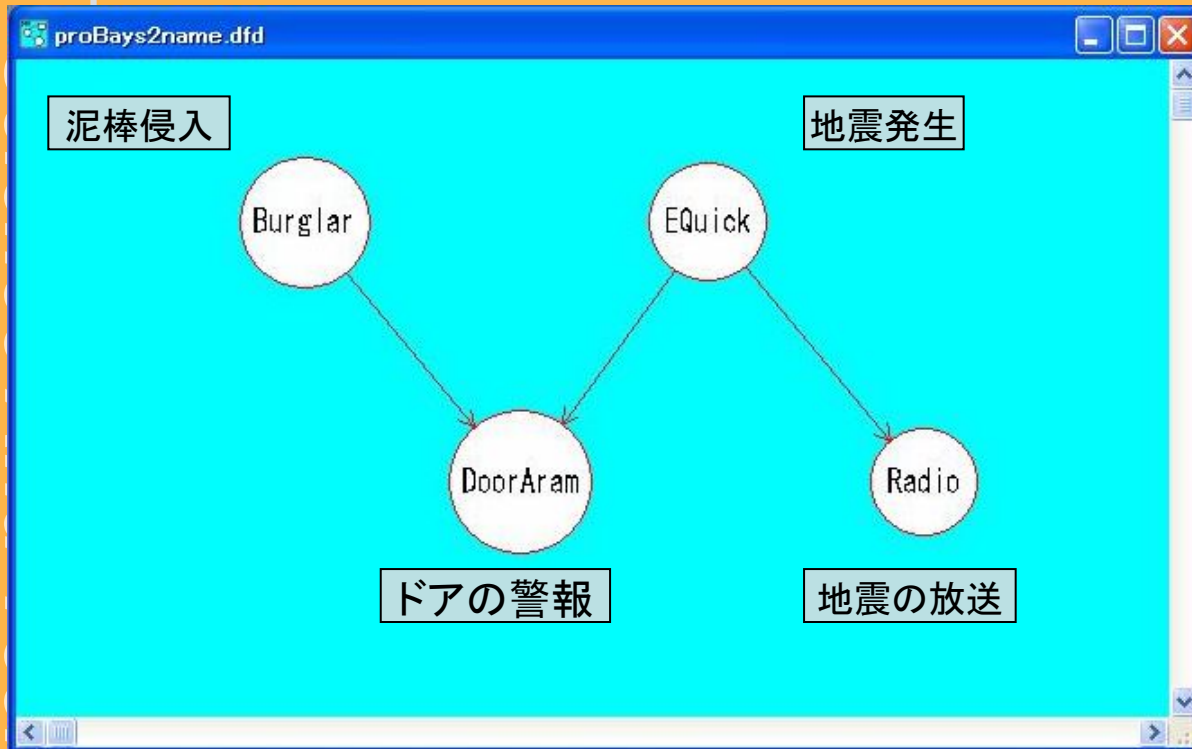


	*work	*job	workspan	income	*marriage	family	*home	homespan
1	会社員	その他	0	1680000	既婚	?	家族持家	5
2	会社員	小売卸売	0	1440000	既婚	?	その他	1
3	会社員	その他	2	2020000	既婚	?	その他	7
4	地方公務員	その他	2	4760000	未婚	?	その他	1
5	会社員	その他	3	3340000	既婚	?	家族持家	23
6	法人役員	飲食	2	2900000	既婚	?	家族持家	22
7	会社員	サービス業	4	2870000	既婚	5	一戸建借家	11
8	会社員	サービス業	1	2670000	既婚	4	公寓アパート	14
9	会社員	金融	1	3060000	既婚	?	家族持家	21
10	会社員	その他	4	2800000	既婚	3	家族持家	20
11	会社員	製造業	2	3400000	既婚	1	民間アパート	2
12	会社員	その他	0	3070000	既婚	?	家族持家	22
13	会社員	サービス業	1	1300000	既婚	?	家族持家	0
14	会社員	その他	2	2410000	既婚	4	賃貸マンション	20
15	会社員	その他	1	1910000	既婚	?	家族持家	10
16	会社員	小売卸売	2	2960000	既婚	1	家族持家	15
17	会社員	理容美容	6	1990000	既婚	?	賃貸マンション	4
18	会社員	サービス業	3	2590000	既婚	1	家族持家	15
19	会社員	金融	2	3390000	既婚	?	家族持家	10
20	会社員	製造業	2	2910000	既婚	?	家族持家	8
21	会社員	飲食	0	2400000	既婚	5	家族持家	6
22	会社員	建設土木	4	3130000	既婚	6	本人持家	20
23	会社員	小売卸売	6	3290000	既婚	2	家族持家	10
24	会社員	その他	6	2710000	既婚	?	家族持家	10



- データ項目間の影響関係を一目で把握できる
- ベイジアンネットで各ノードの確率が計算ができる
風が吹けば桶屋が儲かる式の推論ができる

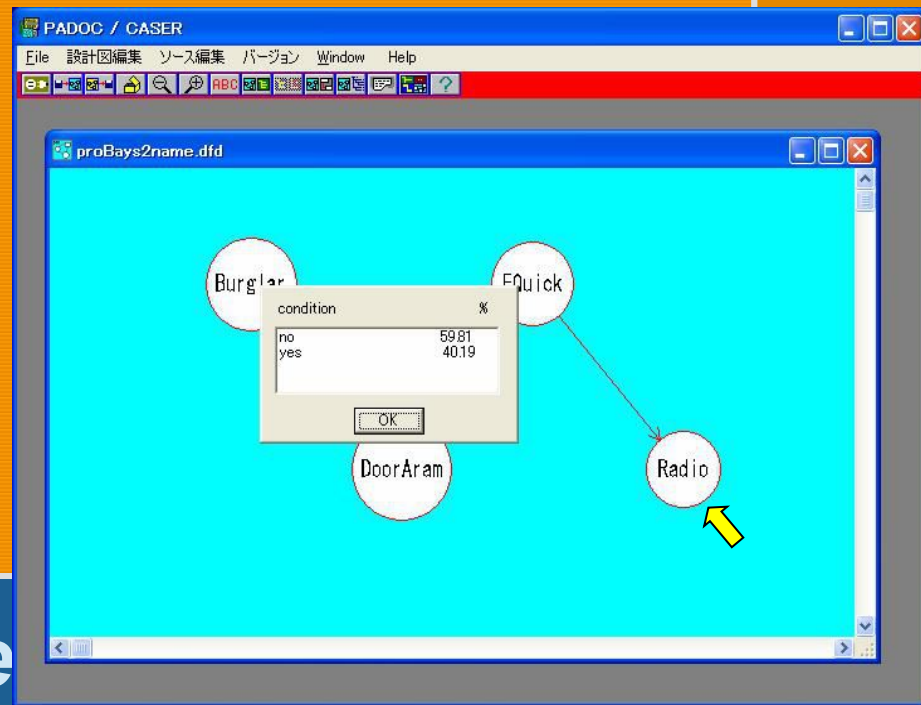
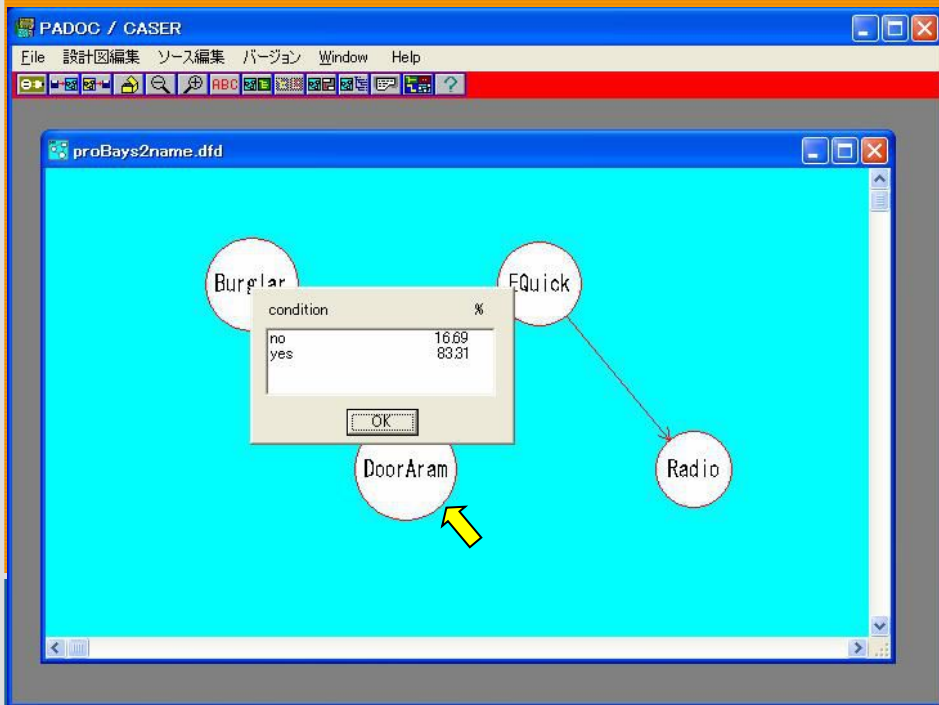
ベイジアンネット確率伝播とは



泥棒侵入(事前確率)
5%に設定

ドアの警報はドアの振動で鳴るモデル

グラフ上の矢印で確率が伝搬する



ドア警報

泥棒侵入は83%に上昇

地震放送

泥棒侵入は40%に減少

データからグラフを自動生成できれば、データに基づく確率になる

項目間の関係で有向グラフ生成

1. 制約ベース(大規模)

条件付独立検定によるV字連結で有向結線
ベ이지アンネットの考案者J.PearlのICアルゴリズム

2. スコアベース(小規模)

MDL(最小記述長)のスコアで有向結線

※MDL(Minimum Description Length)

3. 統合ベース(大規模を小規模にする)

疎な無方向グラフを作ってスコアベースで有向結線

WPS Office

制約ベースの有向グラフ

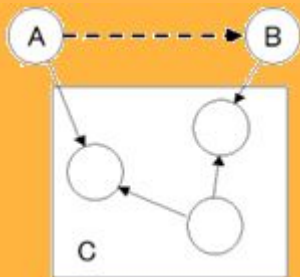
- 任意の2点が有意な関係か調べる(相互情報量)

$$I(A, B) = \sum_{a,b} P(a, b) \log \frac{P(a, b)}{P(a)P(b)} > \delta$$

有意な関係なら2点を**無**方向で連結する



- 次に2点とその他の点について、
条件付独立を検定する

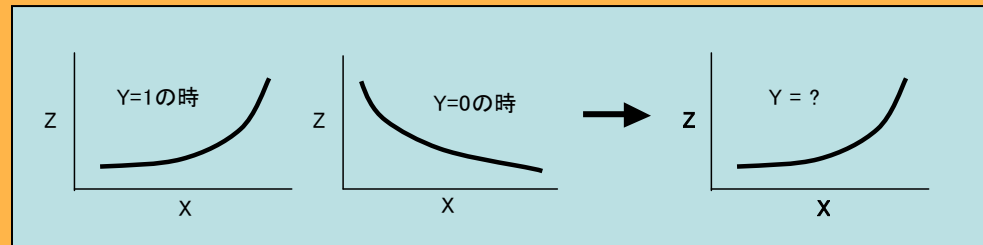
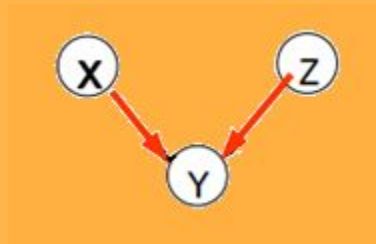


$$I(A, B | C) = \sum_{a,b,c} P(a, b, c) \log \frac{P(a, b | c)}{P(a | c)P(b | c)} > \varepsilon$$

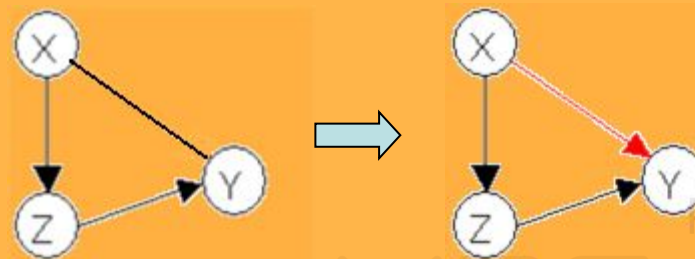
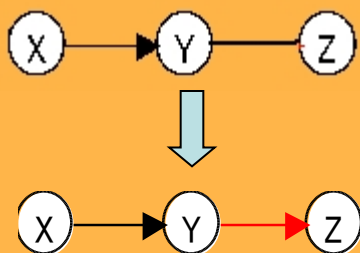
条件付独立で**無い**なら2点を分離して、**V字連結**できる

何故V字連結できるか

- XとZが Yで条件付独立では**無い**(条件付**非**独立)とは
Yが観測されていると Zの値がXの値で決まる
→ XとZの両方が観測されるとYが定まる(**V字連結**)



- V字連結後、無方向で残ったものは次のルール適用



最初に全てのV字連結を探し、次に無方向を有向化する

スコアベースの有向グラフ

MDLスコアが高い方向に連結

- 2点間の方向元を条件とする対数尤度値を計算
- 対数尤度が高い方向に連結する

(例) 乗船等級と遭難時の生死



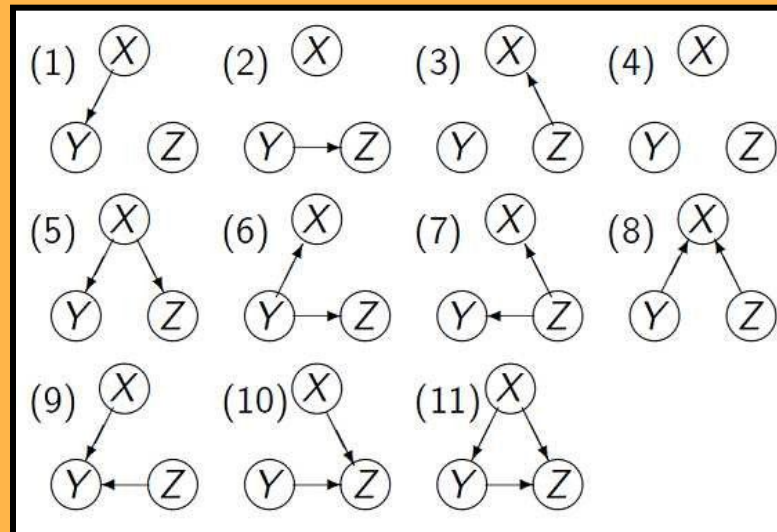
$$MDL0 = -\log \left(\sum_{i=1}^N p(\text{乗船等級} \mid \text{生存有無}) \right) + \frac{k \ln(N)}{2}$$

$$MDL1 = -\log \left(\sum_{i=1}^N p(\text{生存有無} \mid \text{乗船等級}) \right) + \frac{k \ln(N)}{2}$$

生存有無	船室等級	人数	条件確率	pd	log(pd)	条件確率	pd	log(pd)
生存	1等	80	P(1等/生存)	14.6%	-0.83648	P(生存/1等)	37.0%	-0.43136
生存	2等	97	P(2等/生存)	17.7%	-0.7528	P(生存/2等)	52.7%	-0.27805
生存	3等	372	P(3等/生存)	67.8%	-0.16903	P(生存/3等)	75.8%	-0.12054
死亡	1等	136	P(1等/死亡)	39.8%	-0.40049	P(死亡/1等)	63.0%	-0.20091
死亡	2等	87	P(2等/死亡)	25.4%	-0.59451	P(死亡/2等)	47.3%	-0.3253
死亡	3等	119	P(3等/死亡)	34.8%	-0.45848	P(死亡/3等)	24.2%	-0.61553
合計		891		MDL0	-3.21179		MDL1	-1.9717

スコアベースの組合せ爆発

- 全連結全方向でMDLを計算するので組合せの爆発が起こる。3点でも11組ある



ノードが10個の場合 4.2×10^{18} となり、実用的な計算ができない

統合ベースの有向連結

■ 疎な無方向グラフとスコアベースの統合

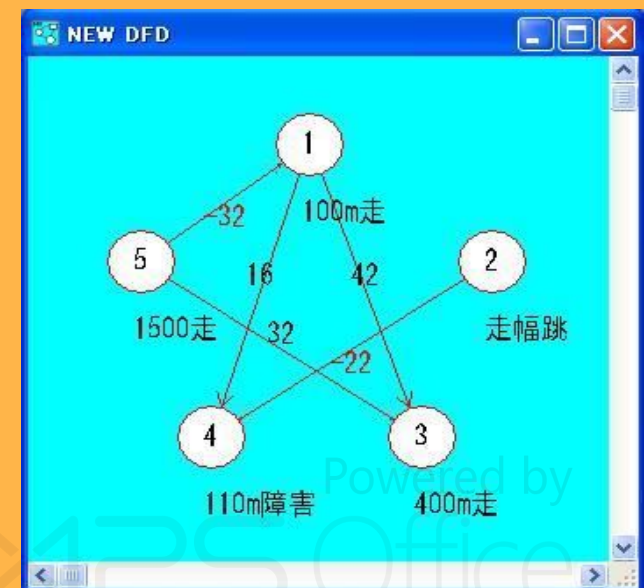
一旦 疎な無方向グラフを生成し、スコアベースで有向化
少ない結線の組合せで爆発を回避する

■ 無方向グラフには

- GGM
- グラフィカルLasso

GGMは全体の情報量を保持
しながら見かけの相関を排除する

※GGM(Gaussian Graphical Mode)



構造推定の適用例

■ 適用データ

1912年4月14日に発生したタイタニック号の乗船名簿と生存者のデータから
ベイジアンネットの構造推定をする。

kaggle データ分析最大コンペ・サイトの課題 <http://www.kaggle.com/c/titanic>

1309名の乗船名簿（欠損データ除く）

氏名	Name
性別	Sex
年齢	Age
乗船等級	Rank
爵位	Position
船室番号	Cabin
船室階	Floor

加工変数として、船室番号から

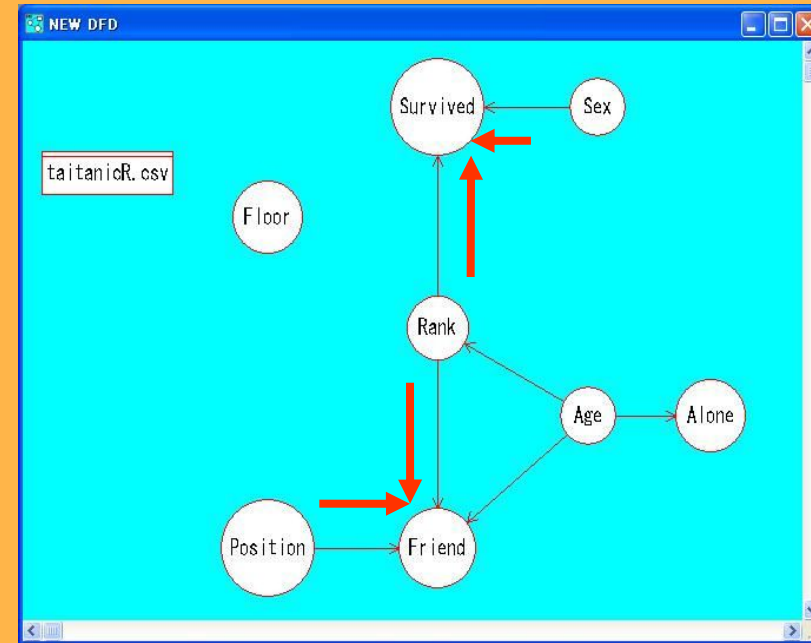
独居	Alone
隣室	Friend

生存有無 Survived



適用例(制約ベース)

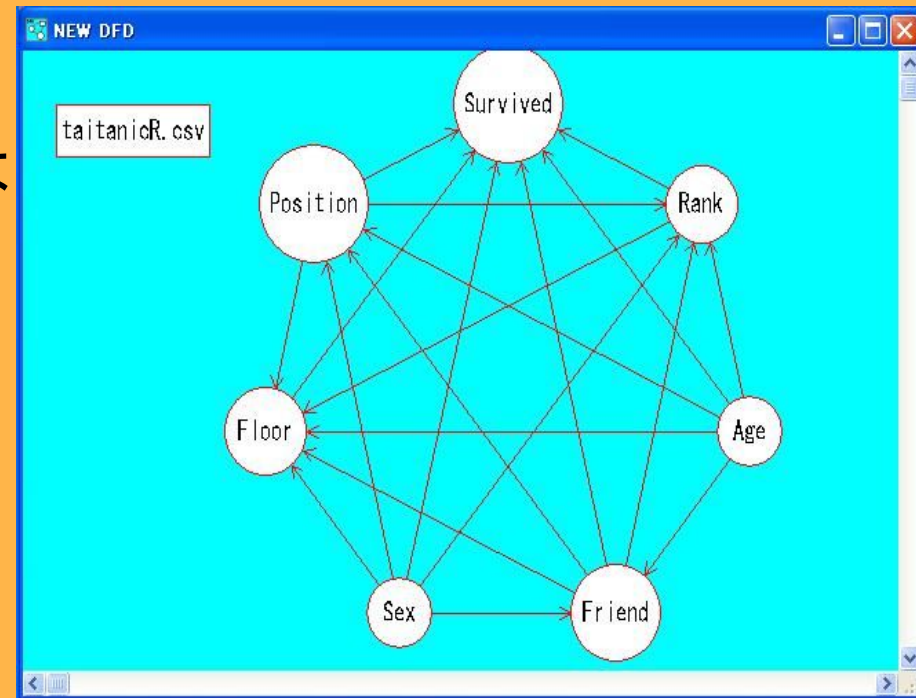
- 有意な2ノード間を無方向連結
 - 条件付**非**独立を検定しV字連結を見つける
 - 閾値 ε (0.01)以下は独立
 - 残る無方向を適応規則で有向化
-
- 条件付**非**独立が検出されV字連結されている
 - SexとRankがSurvivedで条件付非独立
 - RankとPositionがFriendで条件付非独立



Survived	生存有無
Sex	性別
Age	年齢
Friend	同船室
Position	爵位
Rank	乗船等級
Floor	船室階
alone	単独

適用例(スコアベース)

- 7ノードの全リンク方向の組合せは1,138,779,265個
- 全組合せのMDLを計算し最も尤度が高い有向連結は密結合する
ダイヤモンドモデル
- 計算時間はPCで7時間
- 全ノードの矢印がSurvivedを指している。



適用例(統合ベース)

- ノード間の相互情報量より

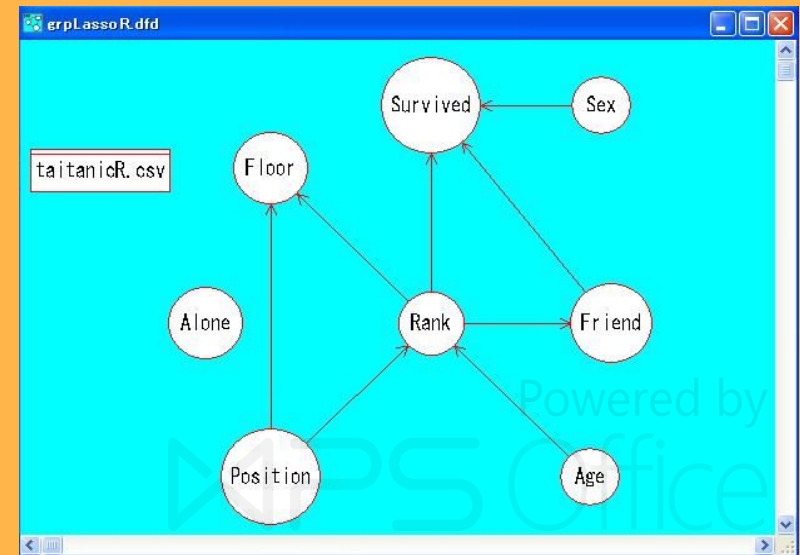
$$I(A, B) = \sum_{a,b} P(a, b) \log \frac{P(a, b)}{P(a)P(b)}$$

疎な関係をグラフィカルLassoで求める。

- グラフィカルLassoの0値の関係は非連結にする
- 連結の方向はMDLの高い方で決定する。

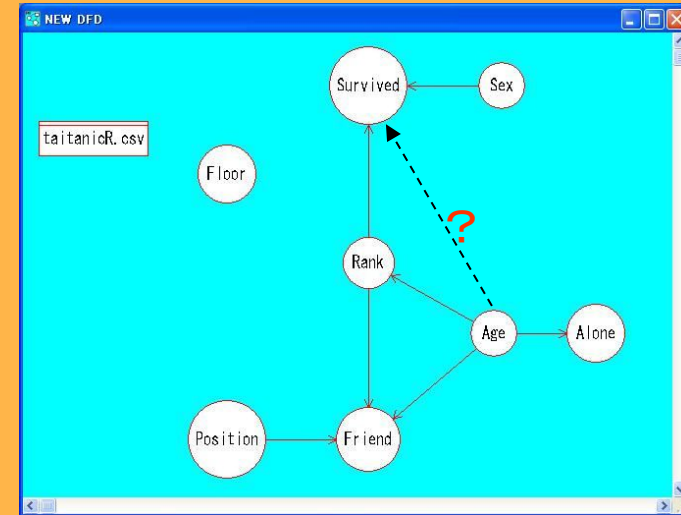
グラフィカルLassoの結果

	Survived	Sex	Rank	Age	Friend	Position	Alone	Floor
Survived	0.706	0.147	0.085	0	0.058	0	0	0
Sex	0.147	0.689	0	0	0	0	0	0
Rank	0.085	0	1.038	0.138	0.040	0.211	0	0.153
Age	0	0	0.138	1.457	0	0	0	0
Friend	0.058	0	0.040	0	0.901	0	0	0
Position	0	0	0.211	0	0	0.735	0	0.657
Alone	0	0	0	0	0	0	0.703	0
Floor	0	0	0.153	0	0	0.657	0	0.979



適用例(考察)

- 適用例で共通している関係
 - 性別と乗船等級が生存に関わっている
 - 年齢は生存と直接関わっていない
- 史実からの説明
 - 座礁から沈没まで3時間かかっており、最初は秩序だった事故対策が採られた
 - 最初に女性を救命ボートに降ろし、次に1等船客を乗せている
 - 生存率(女性は75% 男性は19%)
- データからの説明
 - 年齢各層の生存率をみると差がない
年齢は生存と関係していない



年齢層別生存率

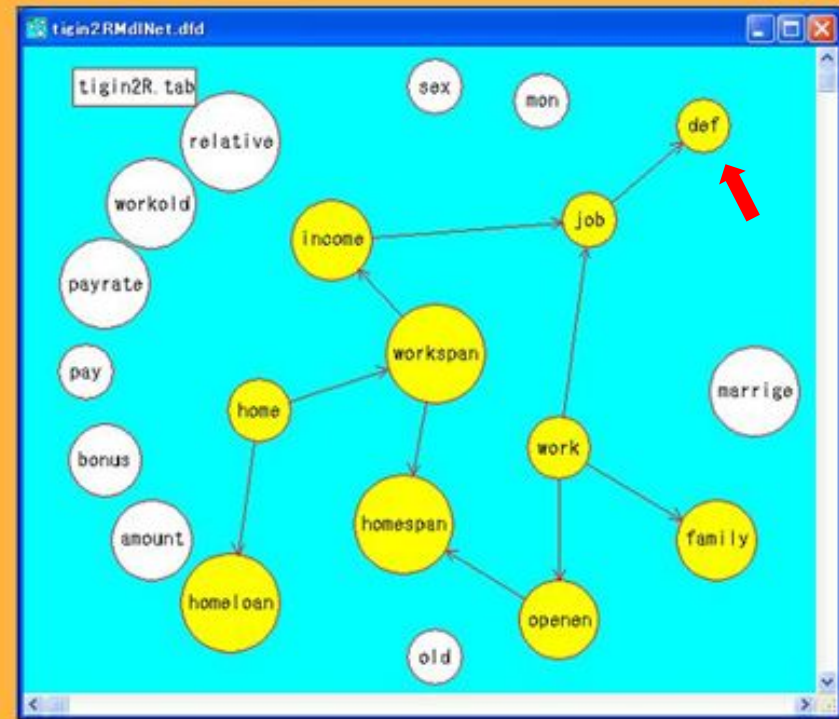
実用的な観点から

実用的な観点から

- 結線数を指定した方が構造図が作成し易い
- 時間軸上の生起する順番が自明なものは有方向にする

結線数10本の構造推定(例)

- 図は6000件、20項目の債務者データから生成（計算時間 38秒）
- ベイジアンネットなので、業種(job)や収入(income)や勤務形態(work)で未払い確率が計算できる



def	債務不履行
job	業種
work	勤務形態
openen	口座開設期間
family	家族人数
homespan	居住年数
home	住居種類
homeioan	住宅ローン有無
income	年収
workspan	勤続年数

ロボットへの応用

■ ベイジアンネットとロボテックスの結合

- ロボットの設計では報酬付き状態遷移図を作成する
- ロボットの行動ログデータから状態遷移図をベイジアンネットに変換し、行動の因果関係を確率に把握する
- ロボットに望ましい動きをさせるため、報酬と因果の関係より、報酬を調整する(逆強化学習理論?)

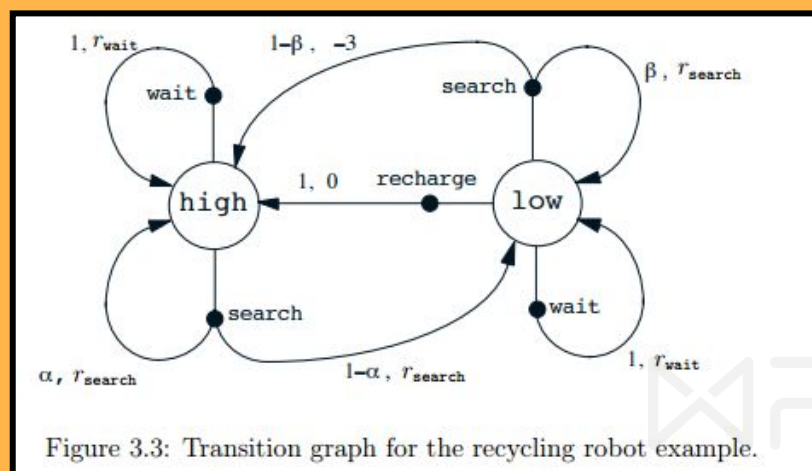


Figure 3.3: Transition graph for the recycling robot example.

ロボットの状態遷移図(Sutton)

まとめ

- ベイジアンネットはグラフ上で確率的な推測ができるがデータから直接ベイジアンネットを生成できると実用的
- データから有向グラフを構造推定には主に3方法ある統合モデル型で、有意な連結数を指定してMDLによる方向付けが判り易いグラフを作る
- ロボットの制御で、報酬をベイジアンネットから逆学習する手法が今後の課題

謝辞

- 大規模ベイジアンネットワークの構造学習の研究者Dr.森下民平(きざしカンパニー)さんに懇切な解説を頂いた。
- Y.Mitsuiさんのタイタニック号の構造学習の解説と公開プログラムが大変参考になった。

データ解析の備忘録

<http://mitsui725.blogspot.jp/2014/02/c.html>

Powered by

WPS Office

参考文献

- [1] Kevin Murphy *Machine Learning § 20* (2012)
- [2] Judea .Pearl 黒木学訳 統計的因果推論 (2009)
- [3] Jie Cheng *Learning Bayesian network form Data* (2000)
- [4] 鈴木讓, 記述長最小規準と状態分割の立場から見た確率的規則の学習、電子情報通信学会論文誌A, Vol. J75-A, No. 8
- [5] 植野真臣 ベイジアンネットワーク (2013)
- [6] Narry Wermuth, Eberhard Scheidt *Fitting Covariance Selection Model to a Matirix*
- [7] 鹿島久嗣 数理情報工学特論第一
- [8] Y.Mitsui データ解析の備忘録
<http://mitsui725.blogspot.jp/2014/02/c.html>
- [9] 中井真人 独立成分分析を用いたベイジアンネット構造推定 SASユーザ会 (2015)
- [10] C.M.Bishop *Pattern Recognition and Mchine Learning § 8* (2006)
- [11] Daphne.Kollar, Nir Friedman *Probabilistic Graphical Model § 3* (2009)
- [12] Padoc/stat ガウシアン・モデル (2013)
http://www1.mjcnet.jp/mabonki/sub/ex_ggm.htm
- [13] 稲邑哲也 ロボティクスにおけるベイジアンネットの応用 人工知能学会誌17巻5号 (2002)
- [14] Richard S. Sutton and Andrew G. Barto *Reinforcement Learning* (2012)

Powered by

WPS Office