ベイジアンネットワークによる予測モデルの構築

小坪琢人

平成 29 年 10 月 12 日

目 次

図 目 次

表 目 次

1　はじめに

現在, 様々な分野で機械学習やディープラーニングが用いられている. その中で自動化というものが大きなテーマであり, 「今後自動化される仕事」のようなニュースも増えている. しかし, マシンの進化により失われる仕事があるのなら, 生まれる新しい仕事も当然存在する. これから生まれる仕事, 必要になる仕事を考えるためには, まず機械学習やディープラーニングでどのようなことができるのかを知る必要がある.

自動化の問題点としては, 分析やシュミレーションの実態が見えにくくなるという点である. 機械がある種のブラックボックスとなり, 都合よく利用されてしまう可能性がある. 人間が生活していくうえで, 重要視されるのは効率性よりも人間らしい判断である. その中で機械による効率的な自動化は, 意思決定に不信感をもたらす危険性を持っている.

それらを解決する方法として, 可視化というものが存在する. グラフやプロットにより数値的根拠を明らかにすることで, ブラックボックス化を防ぐことができる. 一般に可視化は得られた結果の解釈の手助けとして用いられているが, モデル自体が視覚的に解釈できるものも存在する. 例としては決定木やグラフィカルモデルなどである. 本論文ではこのグラフィカルモデルについて取り上げる.

グラフィカルモデルは以下のような特徴を持つ.(Bishop, 2007)

1. 確率モデルの構造を視覚化する簡単な方法を提供し, 新しいモデルの設計方針を決めるのに役立つ.
2. グラフの構造を調べることにより, 条件付独立性などのモデルの性質に関する知見が得られる.
3. 精巧なモデルにおいて推論や学習を実行するためには複雑な計算が必要となるが, これを数学的な表現で暗に伴うグラフ上の操作として表現することができる.

グラフはリンク (link) によって接続されたノード (node) の集合からなる. 確率的グラフィカルモデルでは, 各ノードが確率変数を, リンクがこれらの変数間の確率的関係を表現する. 本論では有向グラフィカルとも呼ばれる, ベイジアンネットワーク (Bayesian network) を用いた分類モデルについて議論する.

本研究ではベイジアンネットワークの構造学習における問題点を改良し, 確率変数間の因果関係を正しく推論した上で, ベイジアンネットワークを構築する.

2　理論

2.1　ベイジアンネットワーク

ベイジアンネットワークは, 確率変数間の条件付依存関係を表した非循環型有向グラフ (DAG: Directed Acyclic Graph) で, 各種推論などに用いられる. 各ノードは確率変数を表し, 有向辺は変数間の直接の依存関係を表す. 一般には各変数は離散値を取るが, 閾値を定めて離散値に変換すれば連続値にも適用可能である. ベイジアンネットワークでは, リンクの先にあるノードを子ノード (X_j), リンクの元にあるノードを親

ノード (X_i) と呼ぶ. 親ノードが複数あるとき子ノード X_j の親ノードの集合を $Pa(X_j)$ と書くことにする. X_j と $Pa(X_j)$ の間の依存関係は次の条件付確率によって表される. ただし $Pa(X_j)$ が空集合の時, X_j はベイジアンネットワークの始点であるので, 事前確率分布 ($P(X_j)$) となる.

$$P(X_j|Pa(X_j)) = \frac{P(X_j, Pa(X_j))}{P(Pa(X_j))} \tag{2.1}$$

さらに n 個の確率変数 $X_1, \ldots, X_n$ のそれぞれを子ノードとして同様に考えると, 全ての確率変数の同時確率分布は式 (2.2) のように表せる.

$$P(X_1, \ldots, X_n) = \prod_{j=1}^{n} P(X_j|Pa(X_j)) \tag{2.2}$$

上記の式 (2.1) に基づいて, 各ノードについて条件付確率を求めた表を条件付確率表 (CPT) といい, 各行の和は必ず 1 になる. 任意の子ノードはそのノードの親ノードのいずれかに起因するので総和は 1 となる (親ノード集合に含まれていないノードを条件とした条件付確率は 0 となる). 条件付確率表の例を表 (2.1) に示す.

表 2.1: 条件付確率表（CPT）

	$Pa(X_j) = x_1$	$\ldots$	$Pa(X_j) = x_m$
$X_j = y_1$	$p(y_1\|Pa(X_j) = x_1)$	$\ldots$	$p(y_1\|Pa(X_j) = x_m)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$X_j = y_n$	$p(y_n\|Pa(X_j) = x_1)$	$\ldots$	$p(y_n\|Pa(X_j) = x_m)$

ベイジアンネットワークモデルでは親ノードに対応する変数の値が与えられたとき, 逐次的に各ノードの値も計算される. つまり親ノード以外のノードの値が与えられているとき, ベイズの定理を用いて, 尤もらしい親ノードの値を予測することができる. ベイジアンネットワークによる予測モデルとして, Naive Bayes と Tree Augmented Naive Bayes について説明する.

2.1.1 Naive Bayes

Naive Bayes 型のベイジアンネットワークは, 親ノード以外のノードがすべて葉であるような木構造である. ナイーブベイズモデルにおける重要な仮定は, 親ノード以外のノードは, 親ノードに対応する変数の値が与えられた条件の下で条件付独立であるということである. ナイーブベイズモデルの例を図 (2.1) に示す.

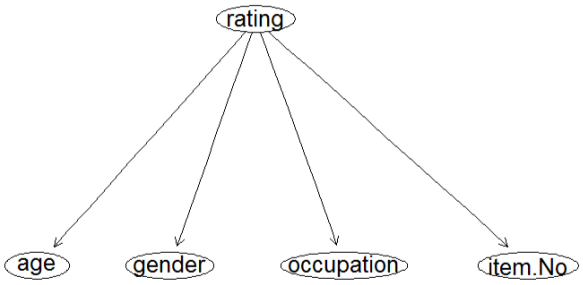

図 2.1: ナイーブベイズの例

構造は常に同様なので, パラメータ学習のみ行う. 訓練データから親ノードと各ノードの値を用いて条件付き確率を求める. 予測を行う際には, 各変数に値を与えてそれぞれ親ノードの値の確率を求め, それらの確率の総和で親ノードの値を決定する.

2.1.2 Tree Augmented Naive Bayes (TAN)

TAN は, 親ノード以外のノード間が条件付独立であるという Naive Bayes の制約を少し緩めたもので, 子ノード間にも以下に示す条件の下でリンクを与えることができる.(N.Friedman et al., 1997)

- 木構造であること
- 全ての点の入次数は 1 以下であること
- n 個の説明変数に対して、$n-1$ 本の有向辺が存在すること

TAN 型のベイジアンネットワークを用いた分類器は, 各変数間の依存関係をある程度反映できるため, グラフ構造から視覚的に情報が得られる. TAN モデル の例を図 (2.2) に示す.

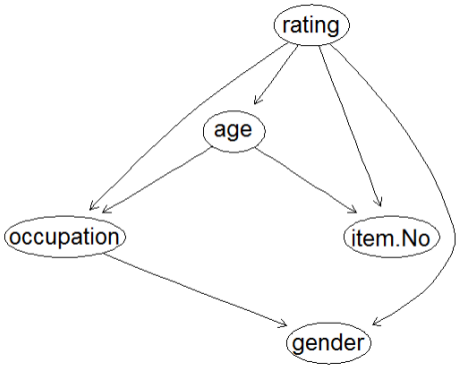

図 2.2: Tree Augmented Naive Bayes

TAN 型のベイジアンネットワークは Naive Bayes 型のベイジアンネットワークを元に, 関係性の強い説明変数について相互情報量をもとにリンクを与える. その際, 元のクラスラベルの値を条件とした条件付相互情報量を用いる. ここでは条件付相互情報量を TAN 構造の仮定の下で最大化する.

まず, エントロピーについて説明する. エントロピーは式 (2.3) のように表され, 不確かさ (乱雑さ) の指標とされる. $p(x)$ は各変数の取りうる値の出現確率を表している. 全ての確率が等しいとき $(1/N)$, エントロピーは最大となり, $\log N$ となる. あるひとつの値の確率が 1 となり, 他の全ての確率が 0 となるとき, エントロピーは最小となり, 0 である. 同様に結合エントロピー (式 (2.4)), 条件付エントロピー, (式 (2.5)) についても示しておく.

$$H(X) = -\int_x p(x)\log p(x) \tag{2.3}$$

$$H(X,Y) = -\int_x\int_y p(x)p(y)\log p(x)\log p(y) \tag{2.4}$$

$$\begin{aligned} H(X|Y) &= -\int_x\int_y p(x,y)\log p(x|y)dxdy \\ &= -\int_x\int_y p(x,y)\log p(x,y)dxdy + \int_x\int_y p(x,y)\log p(x)dxdy \\ &= H(X,Y) - H(Y) \end{aligned} \tag{2.5}$$

次に KL 情報量 (相対エントロピー)(Meyer, 2008) について説明する. KL 情報量は式 (2.6) に示したように, 確率分布の近似に用いられる. KL 情報量が小さいほど二つの分布が近似していることを表し, $p(x) = q(x)$ となるとき KL 情報量は 0 となる.

$$\begin{aligned} KL(p(x)|q(x)) &= -\int p(x)\log q(x)dx - \left(-\int p(x)\log p(x)dx\right) \\ &= -\int p(x)\log\frac{q(x)}{p(x)}dx \end{aligned} \tag{2.6}$$

相互情報量と条件付相互情報量について以下のように定義する.

$$\begin{aligned} \text{相互情報量} &= I(X,Y) \\ \text{条件付相互情報量} &= I(X,Y|C) \end{aligned}$$

ここで相互情報量と条件付相互情報量を KL 情報量の式で表すと以下のように表せる. 相互情報量は X と Y の依存度を表している. 式 (2.7) より X と Y が独立のとき $p(x,y) = p(x)p(y)$ となり, KL 情報量の定義より $I(X,Y) = 0$ となる. すなわち依存関係がある場合に相互情報量は大きくなる. 条件付相互情報量についても同様である.

$$
\begin{aligned}
I(X,Y) &= KL(p(x,y)|p(x)q(x)) \\
&= -\int\int p(x,y)\log\frac{p(x,y)}{p(x)q(x)}dxdy
\end{aligned} \tag{2.7}
$$

$$
\begin{aligned}
I(X,Y|Z) &= KL(p(x,y|z)|p(x|z)q(x|z)) \\
&= -\int\int\int p(x,y,z)\log\frac{p(x,y|z)}{p(x|z)q(x|z)}dxdydz \\
&= -\int\int\int p(x,y,z)\log p(x,y,z)dxdydz \\
&\quad +\int\int\int p(x,y,z)\log p(x|z)dxdydz
\end{aligned} \tag{2.8}
$$

相互情報量と条件付相互情報量は, エントロピーの定義式 (式 (2.3), 式 (2.4), 式 (2.5)) を用いて置き換えることができる.

$$
\begin{aligned}
I(X,Y) &= H(X) - H(X|Y) \\
&= H(X) - H(X,Y) + H(Y)
\end{aligned} \tag{2.9}
$$

$$
\begin{aligned}
I(X,Y|Z) &= H(X|Z) - H(X|Y,Z) \\
&= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z)
\end{aligned} \tag{2.10}
$$

TAN 型のベイジアンネットワークでは式 (2.10) に示した, 条件付相互情報量の式を元に親ノード以外のノード間にリンクを与えている. パラメータ学習はナイーブベイズモデルと同様に, 訓練データから親ノードと各ノードの値を用いて条件付き確率を求める. 予測を行う際には, 各変数に値を与えてそれぞれ親ノードの値の確率を求め, それらの確率の総和で親ノードの値を決定する.

2.2 統計的因果推論

統計的因果推論は, 因果関係をデータから推測する方法論である.(清水, 2017)

3 関連手法

3.1 Matrix Factorization

この手法は協調フィルタリングにおいて次元削減を実現手法である. 協調フィルタリングとはユーザーのレビューをもとに, 同様の評価パターンを持つユーザー同士のデータをもとにまだ評価していない（本質的に

はまだ知らない）アイテムに対しても同様の評価をするだろうと推定するものである. Matrix Factorization の手法ではより少ない次元で特徴を抽出し, 評価の推定を行う.(Koren et al., 2009)

m 人のユーザーと n 個のアイテムを考える. ユーザの評価値を表す $m \times n$ の行列 R に対して, ユーザの特徴を表す $m \times k$ の行列 P と, アイテムの特徴を表す $n \times k$ の行列 Q を考えて以下のように近似できる.

$$R \approx PQ^T \tag{3.1}$$

ここでユーザ u が評価したアイテム i の評価値を $\vec{p_u}^T \vec{q_i}$ として表現する. この各ユーザ, 各アイテムに対する $\vec{p_u}, \vec{q_i}$ を既知の評価値から学習する手法が Matrix Factorization である. 以下の式を満たす P, Q を訓練データから導く.

$$\min_{p,q} = \sum_{(u,i)\in R} (r_{u,i} - \vec{p_u}^T \vec{q_i})^2 + (||\vec{p_u}||_F^2 + ||\vec{q_i}||_F^2) \tag{3.2}$$

式 (3.2) を最適化する更新式を以下に示す.

$$e_{u,i} = r_{u,i} - \vec{p_u}^T \vec{q_i} \tag{3.3}$$

$$p'_{u,k} = p_{u,k} + 2 * \alpha e_{u,v} q_{k,i} \tag{3.4}$$

$$q'_{i,k} = q_{k,i} + 2 * \alpha e_{u,v} p_{u,k} \tag{3.5}$$

具体的な方法は行列 P, Q を適当な乱数を発生させて, それを初期状態として式 (3.3) を用いて, 更新していく確率的勾配降下を用いた最適化手法である. 実際の評価値と近似式より計算される推定値の誤差を最も小さくすることで, 実際の評価がない部分についても推定を行うことができる.

4 評価実験

本章では 2 章で提案したベイジアンネットワークモデルのレーティング予測性能を, 3 章で示した協調フィルタリング法による結果と比較する.

4.1 解析データ

本研究では, GroupLens プロジェクトによる公開データセット MovieLens の一部分を用いた. MovieLens データセットは映画評価サイト”movielens.com” において 1997 年 9 月から 1998 年 4 月までの 7ヶ月間の間に集められた 943 人のユーザ, 1682 個の映画についての 10 万個のレーティングデータ, 簡単なユーザ情報, コンテンツ情報から構成されている. レーティングデータは, 1 から 5 までの 5 段階評価で数字が大きいほ

ど高い評価である. 各ユーザは最低 20 個のレーティングを持っている. ユーザ情報, コンテンツ情報について表 (4.1) にまとめる. またデータについて一部を示す.

表 4.1: ユーザ情報およびコンテンツ情報

映画ジャンル	unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western
職業	administrator, artist, doctor, educator, engineer, entertainment, executive, healthcare, homemaker, lawyer, librarian, marketing, none, other, programmer, retired, salesman, scientist, student, technician, writer
年齢	5 歳ごとに分割 or 10 歳ごとに分割
性別	male, female

表 4.2: MovieLens データセット

user-No	age	gender	occupation	item-No	rating
1	24	M	technician	1	5
159	23	F	student	274	3
535	45	F	educator	511	3
655	50	F	healthcare	393	2
943	22	M	student	1330	3

4.2 評価基準

性能評価の基準として, 予測されたレーティング $\hat{r}_{i,u}$ と真のレーティング $r_{i,u}$ との平均二乗誤差 (Mean Squared Error: MSE):

$$\frac{1}{|\mathcal{W}|} = \sum_{(i,u)\in\mathcal{W}} (\hat{r}_{i,u} - r_{i,u})^2 \tag{4.1}$$

を用いた. ここで, $\mathcal{W}$ は評価に用いるデータに含まれるレーティングのインデックス集合であり, $|\mathcal{W}|$ は $\mathcal{W}$ に含まれる要素の数, すなわちレーティングの総数を表す. MSE が小さいほど予測の精度が高いことになる.

4.3 実験結果

第 2 章で示した, Naive Bayes 型, TAN 型のベイジアンネットワークモデルを用いて, 予測を行う. TAN 型のベイジアンネットワークについては式 (2.10) を用いて, 関係性の強い変数を決定する. 変数として, age, gender, occupation, item.No を選んだ場合の条件付相互情報量を表 (4.3) に示す.

表 4.3: 条件付相互情報量

	age	gender	occupation	item.No
age	3.666	0.0618	**0.873**	**0.981**
gender		0.573	**0.0988**	0.0644
occupation			2.583	0.517
item.No				6.564

実際にリンクを与える部分を太字で強調した. 表からわかるように (ocupation, item.No) 間のほうが値が大きいがここに線を引くと閉路ができてしまうのでその次に大きい値を用いていることがわかる.

これらの手順により相互情報量をもとに無向辺を与える. ここで相互情報量をもとに有向辺を与えることができないことに注意する. 有向辺を決定するのは TAN 型のベイジアンネットワークにおける条件を用いてベイジアンネットワークを構成する.

前項で定めたグラフモデルをもとに, 全変数の取りうる値について条件付確率を求める. ここでも一例を表 (4.3) に示す. これらの条件付確率表 (CPT) をもとにテストデータについて最も確率の高いレイティングを出力とする.

表 4.4: 条件付確率表の例 (occupatin = executive)

gender/rating	1	2	3	4	5
F	0.113	0.089	0.242	0.121	0.147
M	0.886	0.910	0.757	0.878	0.852

4.3.1 性能評価

第 2 章で示した, Naive Bayes 型, TAN 型のベイジアンネットワークモデルを用いて, 予測を行う. 性能評価は MSE を用いた. 結果を表 (4.5) に示す.

表 4.5: 各手法における性能評価

	MSE
Matrix Factorization	0.8832
Naive Bayes model	1.3027
TAN model	2.0558

5 まとめと今後の課題

AdaBoost, バギングとの組み合わせを実装する. 以前読んだ論文でも Boosting の手法を取り入れたベイジアンネットワークを構築していたので, その論文を参考にする. また, 性能評価の方法が適切であるかを

検討する.

6 謝辞

参考文献

Bishop, C.M. (2007) 『パターン認識と機械学習　下』，シュプリンガー・ジャパン.

Koren, Y., R. Bell, and C. Volinsky (2009) “Matrix Factorization Techniques for Recommender Systems,” *Computer*, Vol. 42, No. 8.

Meyer, Patrick Emmanuel (2008) “Information-Theoretic Variable Selection and Network Inference from Microarray Data,” Ph.D. dissertation, Universite libre de Bruxelles.

N.Friedman, D.Geniger, and M.Goldszmidt (1997) “Bayesian Network Classifiers,” *Machine Learning*, Vol. 29, pp. 131–163.

清水昌平 (2017) 『統計的因果探索』，講談社.