5. ベイズ分類機

電気通信大学大
情報理工学研究科
数理情報工学プログラム
植野　真臣

本日の目標

- ベイジアンネットワークの下位モデルである離散ベイズ分類機について学ぶ

分類機 Classifier

- 入力： 属性データ
- （連続データか離散データ）
- 出力： 分類 （離散）

分類機の種類

1. 入力データが連続データの場合

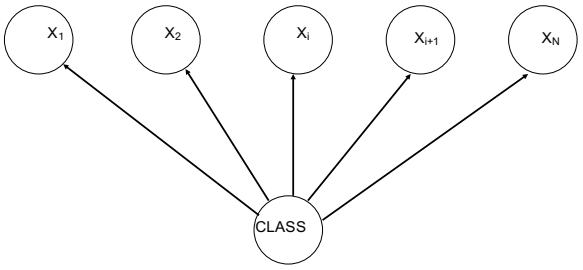
ロジスティック回帰、SVM(サポートベクターマシン)

2. 入力データが離散データの場合

Naïve Bayes, Tree Augmented Network classifier(TAN), Bayesian network Augmented Network classifier (BAN)

2. Naïve Bayes

G. Graham, “A plan for spam”, (2002)



Naïve Bayesモデル

$$p(class \mid x_1, \cdots, x_N) = \frac{p(x_1, \cdots, x_N \mid class)\, p(class)}{\sum_{class} \left[p(x_1, \cdots, x_N \mid class)\, p(class) \right]}$$

$$= \frac{p(x_1, \cdots, x_N \mid class)\, p(class)}{p(x_1, \cdots, x_N)}$$

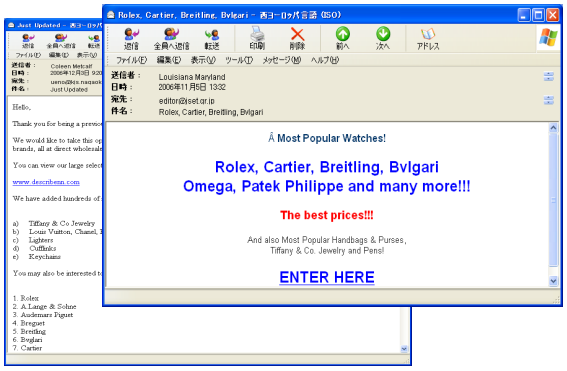
$$\approx \frac{p(class)}{p(x_1, \cdots, x_N)} \prod_{i=1}^{N} p(x_i \mid class)$$

$p(x_i \mid class)$ は、classでx_iが出現する文書数

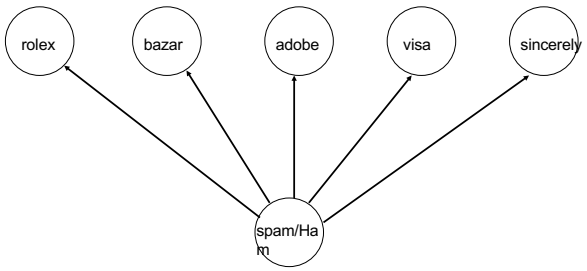
識別関数

$$g_{class} = \log p(class) + \sum_{i=1}^{N} \log p(x_i \mid class)$$

例：ベイジアン・フィルタリング



例：ベイジアン・フィルタリング



識別関数の比較判断

$$g_{spam} = \log p(spam) + \sum_{i=1}^{N} \log p(x_i \mid spam)$$

$$g_{ham} = \log p(ham) + \sum_{i=1}^{N} \log p(x_i \mid ham)$$

例題

- $P(spam) = 0.3$
- $P(Rolex|spam) = 0.9$, $P(\text{Rolex}|ham) = 0.1$
- $P(Cartier|spam) = 0.9$, $P(Cartier|ham) = 0.1$

のとき，　Rolex, Cartierの文字列の入ったメールが来た。これがspamである確率を求めよ。

解答

- $P(spam) = 0.3$
- $P(Rolex|spam) = 0.9$, $P(\text{Rolex}|ham) = 0.1$
- $P(Cartier|spam) = 0.9$, $P(Cartier|ham) = 0.1$

$$g_{spam} = \log p(spam) + \sum_{i=1}^{N} \log p(x_i \mid spam)$$
$$= \log 0.3 + \log 0.9 + \log 0.9 = -0.614$$

$$g_{ham} = \log p(ham) + \sum_{i=1}^{N} \log p(x_i \mid ham)$$
$$= \log 0.7 + \log 0.1 + \log 0.1 = -2.155$$

$$g_{spam} > g_{ham}$$

問題

- 自然言語処理における文書処理では、同じ単語が出現する頻度が重要になる。
- そこで単語の出現分布を考慮したモデルが考えられる。

3．多項モデル

$$P(class \mid x_1, \cdots, x_N) = \frac{p(x_1, \cdots, x_N \mid class)\, p(class)}{p(x_1, \cdots, x_N)}$$

$$\propto P(class) \prod_{i}^{N} \prod_{v \in x_i} p(v \mid class)^{n(x_i, v)}$$

ここで、vは単語で、x_iはi番目の文書、$n(x_i, v)$ は文書x_i 中のvの出現頻度、

$p(v \mid class)$ はclass中に出てくる単語vの出現確率（vの出現頻度/全単語出現頻度の和）

（一般には文章の長さのパラメタを考慮するが、後の流れのために文書の長さを考慮しないモデルを考える）

例題

- Webページを「政治」「それ以外」に分類したい。
- P(”首相”|政治) = 0.7, P(”首相”|政治でない) = 0.1, P(政治) = 0.1とする。
- あるWebページでは、“首相”という単語が20回出ていた。このページが政治に関するものである確率を求めよ。

解答

$$g_{政治} = \log p(政治) + n(x,v) \log p(首相 \mid 政治)$$
$$= \log 0.1 + 20 \log 0.7 = -4.098$$
$$g_{政治でない} = \log p(政治でない) + n(x,v) \log p(首相 \mid 政治でない)$$
$$= \log 0.9 + 20 \log 0.1 = -20.046$$
$$p(政治 \mid データ) \approx 1.0$$

例題

- 学生レポートのトピックを「政治」「それ以外」に分類したい。
- P(”首相”|政治) = 0.7, P(”首相”|政治でない) = 0.1, P(政治) = 0.1とする。
- 同じ課題の3つのレポートでは、1つめのレポートで“首相”が40回、2つめのレポートでは“首相”が0回、3つ目のレポートでは“首相”が25回出ていた。これらのレポートが政治関係かそうでないかをあてよ。

解答

$$g_{政治} = \log p(政治) + \sum_{x} n(x,v) \log p(首相 \mid 政治)$$
$$= \log 0.1 + 40 \log 0.7 + 0 \log 0.7 + 25 \log 0.7 = -11.069$$
$$g_{政治でない} = \log p(政治でない) + \sum_{x} n(x,v) \log p(首相 \mid 政治でない)$$
$$= \log 0.9 + 40 \log 0.1 + 0 \log 0.1 + 25 \log 0.1 = -65.046$$
$$p(政治 \mid データ) \approx 1.0$$

問題

- 一つでも訓練データ中でclassの文章に表れなかったvがあるとp(v|class)=0となり、式全体が0となってしまう。
- 例 $p(\text{首相}|\text{政治}) = 0.0$

と推定されてしまうような場合

4．ディレクレ・モデル

- P(v|class)を確率変数として扱い、θ(v,c)とする。さらにα_{cv}を持つディレクレ分布に従うとすると

$$P(class \mid x) \propto p(class)p(x \mid class) = p(class)\int p(x \mid \theta)p(\theta \mid class)d\theta$$

$$= P(class)\frac{\Gamma(\sum_v \alpha_{cv})}{\Gamma(\sum_v \alpha_{cv}+n)}\prod_{v\in x_i}\frac{\Gamma(\sum_v \alpha_{cv}+n(x,v))}{\Gamma(\alpha_{cv})}$$

5.潜在クラスディレクレモデル

- Hofmann,T,"Probablistic latent Semantic analysis", Proc. of UAI99, pp.289-296 原型モデル、EMを用いて推定されている
- (D.M. Blei, A.Y.Ng and M.I.Jordan:Latent Dirichret Allocation", Journal of Machine Learning Research, 3. pp.993-1022,2003)
- クラスがさらに潜在クラスψによって分類されるモデルP(class|ψ)

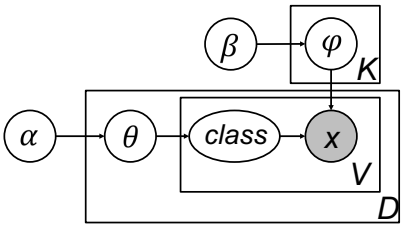
$$P(x \mid \alpha, \theta) = \int p(x \mid \psi)p(\psi \mid \alpha)d\psi$$

一般にはMCMC法を用いて推定することが多い

$$= \int\prod_{n=1}^{N}\sum_{c=1}^{K} p(v \mid c)p(c \mid \psi)p(\psi \mid \alpha)d\psi$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)}\int\prod_k \psi_k^{\alpha_k-1}\prod_{n=1}^{N}\prod_{k=1}^{K}\psi_k\theta_{kv}d\psi$$

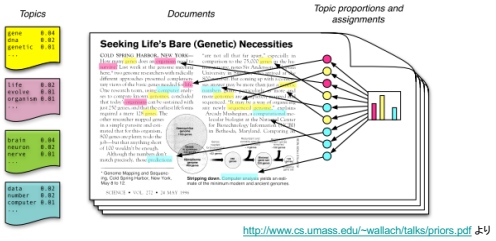
Latent Dirichlet Allocation のグラフィカルモデル



- 観測データxは所属クラス（class）k に対応するディリクレ分布 $p(x|\phi_k,\beta)$に従うと仮定
- クラスはパラメータθ（ハイパーパラメータα）に従うディリクレ分布 $p(class = k|\theta,\alpha)$に従うと仮定

トピックモデルとしての解釈

一つの文書は、複数の話題（潜在クラス；Topicと呼ぶ）から構成されており、話題ごとに単語の分布が異なると仮定



http://www.cs.umass.edu/~wallach/talks/priors.pdf より

単語の分布から各文書中の話題（クラス）の分布を推定

トピックごとの単語分布

トピック1 "Arts"	トピック2 "Budgets"	トピック3 "Children"	トピック4 "Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

http://www.cs.umass.edu/~wallach/talks/priors.pdf より

Naïve Bayes v.s.Bayesian network

Naïve Bayes v.s.Bayesian network
どちらの性能が良いか？

Naïve Bayes v.s.Bayesian network

Naïve Bayes v.s.Bayesian network
どちらの性能が良いか？

- Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classiers. *Machine Learning*, 29(2):131-163, 1997.
- 必ずしもBayesian networkが良いとは限らない。すごく、Naïve Bayesが良いときもある。

理由

- 分類問題ではClass変数をC,その他の変数を $X_1, X_2, \cdots, X_N$ とすると
- ベイジアンネットワークでは、

全変数についての周辺尤度

$$P(X, C \mid G) = P(x_1, x_2, \cdots, x_N, c \mid G)$$

を最大化。ただし、Xはデータセット、小文字はそれぞれの変数のデータの値を示す。

- 分類問題ではClass変数をCとすると

$$P(C \mid X, G) = \frac{P(x_1, x_2, \cdots, x_N, c \mid G)}{P(x_1, x_2, \cdots, x_N \mid G)}$$

- を最大化しなければならない。

問題

$$P(C \mid X, G) = \frac{P(x_1, x_2, \cdots, x_N, c \mid G)}{P(x_1, x_2, \cdots, x_N \mid G)}$$

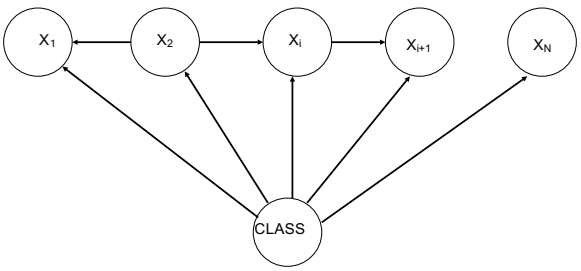
- を計算するための計算量が大きすぎる

近似手法

- Naïve Bayesのように強制的に特徴変数からクラス変数にリンクを引けばいい！！
- (Friedman et al, 1997)

6. Tree Augmanted Naïve-Bayes (TAN)

(Friedman et al, 1997) Chow-Liu algorithm



相互情報量を用いた手法

変数xと変数yの因果の強さとしての相互情報量は

$$I(x,y)=\sum_{x,y}p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}\geq 0$$

相互情報量による木の生成 MWST法(Chow and Liu 68)

1. 与えられたデータより、N(N-1)/2個の枝について、すべての枝の相互情報量I(xi,xj)を求める。
2. 最も大きな値を示す枝を取り出し、木を構成する枝とする。
3. ループができないならば、次に枝を木に加え、ループができるのであればその枝を棄てる。
4. ステップ3をN-1個の枝が選ばれるまで続ける。
5. 一つのルートを選び、そこから外方向へエッジを引く。

Chow & Liu, 1968, Approximating discrete probability distributions with dependence trees.
IEEE Transactions on Information Theory, IT-14,462-467

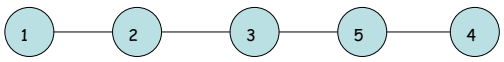
表1

	ノード番号				
	1	2	3	4	5
1	1	0	1	1	1
2	0	0	0	1	0
3	1	1	1	1	1
4	1	1	1	1	0
5	1	1	1	1	0
6	1	1	0	0	0
7	1	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	1	1	1	1
11	0	0	0	0	0
12	1	1	1	1	1
13	0	0	1	1	1
14	1	1	1	1	1
15	1	0	1	1	1
16	1	1	1	0	0
17	1	1	0	1	0
18	1	1	1	0	0
19	1	1	0	1	0
20	1	1	1	0	1
平均	0.70	0.60	0.60	0.60	0.40

表1から計算された5変数間の相互情報量

	ノード番号				
	1	2	3	4	5
1		.0756	.0274	.0038	.0017
2			.0308	.0060	.0004
3				.0308	.1264
4					.0499
5					

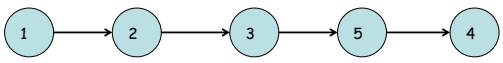
得られた木



変数1をルートとすると

	ノード番号				
	1	2	3	4	5
1		.0756	.0274	.0038	.0017
2			.0308	.0060	.0004
3				.0308	.1264
4					.0499
5					

得られた木



MWST法の性質

利点

- 二次統計量までしか用いないために、データからの演算が容易で信頼できる。
- 計算量が高々$O(n^2)$のオーダーである。

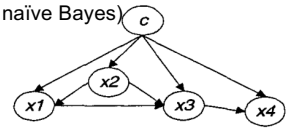
欠点

・ネットワーク構造を表現できない

- 漸近一致性がない

8. BAN

(Bayesian network-augmented naïve Bayes)



- TANよりもさらに条件を緩くし、より学習データにfitした構造。
- 条件：目的変数が親を持たず、説明変数は必ず目的変数を親に持つ
- 利点：NBやTANよりも真の構造に近い構造を学習できる
- 課題：
  データの取り得る組合せ数に対してデータ数が少ないと、過学習を起こしやすい。
  計算量がNBやTANよりも大きい。

K. J. Ezawa and S. W. Norton, "Constructing Bayesian networks to predict uncollectible telecommunications accounts," in IEEE Expert, vol. 11, no. 5, pp. 45-51, 1996.

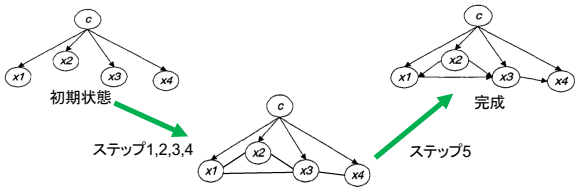
BAN学習のアルゴリズム

- 相互情報量を用いたBAN学習

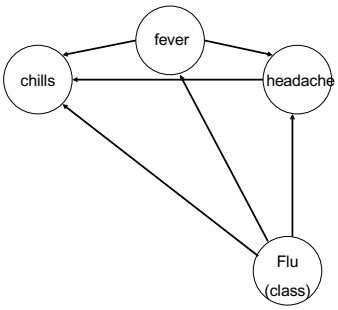
1. 説明変数X_iと目的変数Cの相互情報量$I(X_i, C)$を全てのiに関して求め大きい順にソートする。
2. 下の不等式を満たすようなkの最小値を求める。Nは変数数、kは説明変数間に引かれるエッジの数の許容最大数を表す。
$$\sum_{j=1}^{k} I(X_j, C) \geq t_{CX} \sum_{j=1}^{N} I(X_j, C)$$
ただし、$0 < t_{CX} < 1$　t_{CX}はユーザが決める。
3. Cを所与としたときの説明変数間の相互情報量$I(X_i, X_j|C)$を、ステップ2で選択したk個の変数の全組み合わせについて求め、大きい順にソートする。
4. ステップ3でソートされた相互情報量を順に足していき、許容最大数k個までの総和のt_{XX}パーセント $t_{XX} \sum_{i<j}^{k} I(X_i, X_j|C)$ (t_{XX}はユーザが決める)を超えるまで足していく。このとき足された変数間にリンクを引く。

5. 最後に、ステップ1で求めた$I(X_i, C)$の大きさに従って、エッジの向きを決める。

$$\begin{cases} X_i \rightarrow X_j & if \quad I(X_i, C) > I(X_j, C) \\ X_i \leftarrow X_j & otherwise \end{cases}$$



BAN・TANの推論例



Flu：インフルエンザに罹っている
Fever：熱がある
Chills：寒気がする
Headache:頭痛がある

熱・寒気・頭痛の状態からインフルエンザである確率を推論

BANによる識別

以下の g_{flu}と $g_{\neg flu}$を比較して識別

$$g_{flu} = \log p(flu) + \sum_{i}^{N} \log p(x_i|Pa_i, flu)$$
$$= \log p(flu) + \log p(fever|flu) + \log p(headache|fever, flu) + \log p(chills|fever, flu)$$
$$g_{\neg flu} = \log p(\neg flu) + \sum_{i}^{N} \log p(x_i|Pa_i, \neg flu)$$

ここで、Pa_iはx_iの親変数集合とする

例題1

以下を所与として、寒気・熱・頭痛の全症状が出ている場合にインフルエンザであるといえるかを調べよ。

$p(flu) = 0.2$
$p(fever|flu) = 0.7$
$p(fever|\neg flu) = 0.3$
$p(headache|fever, flu) = 0.6$
$p(headache|fever, \neg flu) = 0.4$
$p(chills|fever, flu) = 0.8$
$p(chills|fever, \neg flu) = 0.5$

例題1の計算結果

$$g_{flu} = \log p(flu) + \log p(fever|flu) + \log p(headache|fever, flu) + \log p(chills|fever, flu)$$
$$= \log 0.2 + \log 0.7 + \log 0.6 + \log 0.8 \fallingdotseq -1.172$$

$$g_{\neg flu} = \log p(\neg flu) + \log p(fever|\neg flu) + \log p(headache|fever, \neg flu) + \log p(chills|fever, \neg flu)$$
$$= \log 0.8 + \log 0.3 + \log 0.4 + \log 0.5 \fallingdotseq -1.319$$

$g_{flu} > g_{\neg flu}$よりインフルエンザであるといえる

例題2

以下を所与として、寒気と頭痛はあるが熱はない場合にインフルエンザであるといえるかを調べよ。

$p(headache|\neg fever, flu) = 0.2,$
$p(headache|\neg fever, \neg flu) = 0.1,$
$p(chills|\neg fever, flu) = 0.2,$
$p(chills|\neg fever, \neg flu) = 0.1,$
上記以外の確率は例題1と同じ.

例題2の計算結果

$$g_{flu} = \log p(flu) + \log p(\neg fever|flu) + \log p(headache|\neg fever, flu) + \log p(chills|\neg fever, flu)$$
$$= \log 0.2 + \log 0.3 + \log 0.2 + \log 0.2 \fallingdotseq -2.619$$

$$g_{\neg flu} = \log p(\neg flu) + \log p(\neg fever|\neg flu) + \log p(headache|\neg fever, \neg flu) + \log p(chills|\neg fever, \neg flu)$$
$$= \log 0.8 + \log 0.7 + \log 0.1 + \log 0.1 \fallingdotseq -2.252$$

$g_{flu} < g_{\neg flu}$よりインフルエンザとはいえない

課題：Tic-Tac-Toe Gameの勝敗予測

以下のリポジトリから取得したデータを用いて、Naïve Bayes, TAN, BANの構造とパラメータを推定し、さらにそれを用いた分類性能の評価を行え

http://archive.ics.uci.edu/ml/machine-learning-databases/tic-tac-toe/tic-tac-toe.data

- ただし、構造とパラメータの学習は総データ数958レコードのうち最初の29レコードと最後の29レコードを除く900レコードで行い、分類性能の評価は除外した58レコードを用いて行うこと。
- 分類性能の評価指標は正答率（分類結果と正解データの一致率）とする。

問題とデータの詳細は次のページ

- 3×3マスを2人のプレーヤ（xと○）が交互に埋めていき、縦横斜めのいずれか一直線にマークが揃えば勝ちというゲームのデータ
- 各レコードは次の変数で定義

1. top-left-square: {x,o,b}
2. top-middle-square: {x,o,b}
3. top-right-square: {x,o,b}
4. middle-left-square: {x,o,b}
5. middle-middle-square: {x,o,b}
6. middle-right-square: {x,o,b}
7. bottom-left-square: {x,o,b}
8. bottom-middle-square: {x,o,b}
9. bottom-right-square: {x,o,b}

マスの位置に対応しているXはプレーヤXが取ったことを、○はプレーヤ○が取ったことを、bはブランクを表す。

10. Class: {positive,negative} ：目的変数（positiveは「Xが勝ち」を表す）

近年の研究

- TAN, BANの学習法があいまい。
- 条件付き周辺尤度

$$P(C \mid X, G) = \frac{P(x_1, x_2, \cdots, x_N, c \mid G)}{P(x_1, x_2, \cdots, x_N \mid G)}$$

- を最大化する、もしくは近似して最大化する研究が主流。

Alexandra M. Carvalho, Teemu Roos, Arlindo L. Oliveira, and Petri Myllym□aki. Discriminative learning of bayesian networks via factorized conditional log-likelihood. *Journal of Machine Learning Research*, 12: 2181-2210, 2011.

条件付き周辺尤度はいらない？

- 定理 (Sugahara, Uto, Ueno 2017)
- 条件付き周辺尤度を最大にするBANと周辺尤度を最大にするBANは同一になる。
- ただし、ベイジアンネットワークでは異なる構造になる。

⇓

- BANでは、

条件付き周辺尤度ではなく、周辺尤度を最大化すればよい。それよりも、周辺尤度を厳密に最大にするアルゴリズムが重要。

BANの厳密学習 Sugahara, Uto, Ueno（2017）

$$\log P(G \mid X) \propto \sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{c=1}^{s}\sum_{k=1}^{r_i}\log P(x_i = k \mid \Pi_i = j, C = c, G)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{c=1}^{s}\left[\log\frac{\Gamma(\alpha_{ijk})}{\Gamma(\alpha_{ij}+n_{ij})} + \log\sum_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+n_{ijk})}{\Gamma(\alpha_{ijk})}\right]$$

- 厳密な周辺尤度を動的計画法で最大にするBAN構造を学習
- 計算量 $O((n+1)\bullet 2^n) \Rightarrow O(n\bullet 2^{n-1})$

スコア計算アルゴリズム

Algorithm 1 : Compute local scores.

```
function MAIN(X, C)
    getLocalScores(X, C)
end function
function GETLOCALSCORES(W, C)
    for all X_i ∈ W do
        Π_i^* ← W \ {X_i}
        LS[X_i][Π_i^*] ← P(X_i|Π_i^*, C)
    end for
    if |W| > 1 then
        for all X_i ∈ W do
            getLocalScores(W \ {X_i}, C)
        end for
    end if
end function
```

親変数探索　アルゴリズム

Algorithm 2 : Find best parent sets for each feature variable.

```
function MAIN(X, C)
    for all X_i ∈ X do
        GetBestParents(X_i, X, C, LS)
    end for
end function
function GET BESTPARENTS(X_i, X, C, LS)
    for all Z ⊂ X \ {X_i} in lexicographic order do
        bps[i][Z] ← Z ∪ C
        bss[i][Z] ← LS[X_i][Z ∪ C]
        for all Z' ⊂ Z which |Z \ Z'| = 1 do
            if bss[Z' ∪ C] > bss[P_c ∪ C] then
                bps[i][Z ∪ C] ← bps[Z' ∪ C]
                bss[i][Z ∪ C] ← bss[Z' ∪ C]
            end if
        end for
    end for
end function
```

データベース

Table 1: Description of data sets used in the experiments

Dataset	Features	Classes	Sample size
Balance	4	3	625
Hayes-Roth	4	3	132
Lenses	4	3	24
banknote authentication	4	2	1372
Car Evaluation	6	4	1728
MONK's Problems	6	2	432
mux6	6	2	64
LED Display Domain	7	10	3200
Nursery	8	5	12960
TicTac	9	2	958
Breast Cancer	9	2	286
threeOf9	9	2	512
Breast Cancer Wisconsin	9	2	699
Solar Flare	10	9	1389
mofn	10	2	1324
parity5+5	10	2	1024
EEG	14	2	14980
Congressional Voting Records	16	2	435
vote	16	2	435
zoo	16	5	101
ClimateModel	18	2	540
ImageSegmentation	18	7	2310

Table 2: The accuracies of each classifier for 22 data sets

Dataset	NB	TAN	TAN-fCLL	gANB-BDeu	eGBN-BDeu	eANB-BDeu	CI-rate
Balance	**0.9152**	0.8656	0.8672	**0.9152**	**0.9152**	**0.9152**	1.00
Hayes-Roth	**0.7879**	0.6667	0.6289	**0.7879**	0.4697	**0.7879**	0.75
Lenses	0.7500	0.6667	0.5833	0.7500	**0.8333**	0.7500	0.75
banknote authentication	0.8432	**0.8819**	0.8819	0.8724	0.8812	0.8812	0.5
Car Evaluation	0.8571	**0.9450**	0.9450	0.8814	0.9416	0.9427	0.67
MONK's Problems	0.7500	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.17
mux6	0.5469	**0.6406**	0.5938	0.5469	0.0313	0.5469	0.00
LED Display Domain	0.7294	0.7319	0.7369	0.7294	**0.7413**	0.7294	1.00
Nursery	0.9033	**0.9348**	**0.9348**	0.9140	0.9340	0.9181	0.375
TicTac	0.6920	0.7610	0.7317	0.7317	0.8340	**0.8497**	0.11
Breast Cancer	0.7168	0.6713	0.6504	0.7098	**0.7448**	0.6958	0.44
threeOf9	0.8164	0.8516	0.8516	0.8086	**0.8887**	0.8730	0.22
Breast Cancer Wisconsin	**0.9742**	0.9628	0.9528	**0.9742**	0.9714	**0.9742**	1.00
Solar Flare	0.7811	0.8200	0.8258	0.8143	**0.8431**	0.8229	0.30
mofn	0.8527	0.9335	0.9290	0.8988	**1.0000**	0.8716	0.70
parity5+5	0.3633	0.3135	0.2998	0.3633	**1.0000**	**1.0000**	0.00
EEG	0.5778	0.6274	0.6305	0.6302	0.6814	**0.6864**	0.00
Congressional Voting Records	0.8989	0.9494	0.9172	0.9517	0.9517	**0.9586**	0.875
vote	0.9034	0.9402	0.9149	**0.9517**	0.9448	0.9494	0.875
zoo	**0.9802**	0.9505	0.9604	0.9604	0.9307	0.9604	1.00
ClimateModel	0.9222	**0.9278**	0.9259	0.9222	0.9000	0.8426	0.11
ImageSegmentation	0.7290	0.7983	0.8017	0.7784	0.8156	**0.8225**	0.89
average	0.7860	0.8114	0.7983	0.8133	0.8320	**0.8637**	-
Z-value	2.637	1.7205	1.7546	1.6636	0.3702	-	-
p-value	$< .01$	$< .05$	$< .05$	$< .05$	.36	-	-

結果

- 1位　提案された厳密に探索された周辺尤度を最大にするBAN…有意に良い分類精度
- 2位　厳密なBayesian net…Hayes Roth,mux6, で極端に悪い。クラス変数と特徴変数の直接のリンクが少ないとき、極端に悪くなる。
- 3位　貪欲法によって近似的に探索された周辺尤度を最大にするBAN
- 4位　条件付き周辺尤度を最大にするTAN
- 5位　TAN(相互情報量によるChow and Liu)
- 6位　Naïve Bayes

最終的にどのように分類機を用いるのか

- BAN モデル平均分類機
- Mすべてのモデル集合、m:m番目のモデル

$$P(C \mid X) = \sum_{m \in M} P(C \mid X, G_m) P(G_m \mid X)$$

まとめ

- 入力、出力ともに離散データの場合はベイズ分類機が有効。
- 厳密に周辺尤度を最大にするBayesian network Augmented Network classifier (BAN)が最も精度が高い。
- 現実には、BANをさらに可能なモデルのモデル平均でクラスの確率を求めればより精度が高くなる。