ベイズ機械学習の基礎

担当教授：植野真臣, TA: 菅原聖太, 加藤弘也

e-mail: kato@ai.lab.uec.ac.jp

0.1 本実験について

- 実験テキストと課題で使うデータセットは,
  http://www.ai.lab.uec.ac.jp/実験/ からダウンロードできる.

- 予定

期日	内容
10 月 5 日（水） 13:00 – 14:30	講義：ベイズとコンピュータサイエンス、ビッグデータ、ＡＩ
10 月 5 日（水） 14:40 – 16:10	講義：ベイズとコンピュータサイエンス、ビッグデータ、ＡＩ
10 月 10 日（月） 13:00 – 14:30	講義：ベイズとコンピュータサイエンス、ビッグデータ、ＡＩ
10 月 10 日（月） 14:40 – 16:10	講義：ベイズとコンピュータサイエンス、ビッグデータ、ＡＩ
10 月 12 日（水） 13:00 – 14:30	実験課題
10 月 12 日（水） 14:40 – 16:10	実験課題
10 月 17 日（月） 13:00 – 14:30	講義：確率、最尤法とベイズ的アプローチ
10 月 17 日（月） 14:40 – 16:10	講義：確率、最尤法とベイズ的アプローチ
10 月 19 日（月） 13:00 – 14:30	講義：Naive Bayes，TAN，ディリクレモデル
10 月 19 日（月） 14:40 – 16:10	講義：Naive Bayes，TAN，ディリクレモデル
10 月 24 日（月） 13:00 – 14:30	実験課題
10 月 24 日（月） 14:40 – 16:10	実験課題
10 月 26 日（水） 13:00 – 14:30	実験課題
10 月 26 日（水） 14:40 – 16:10	実験課題

- 課題について
  - 課題は全部で 14 個ある.
  - 注意：絶対に人のソースをコピーしたりしないこと．発覚した場合は不合格になる.
  - 実装で使うプログラム言語はどれでもよい．ただし，課題 12 以降については java で書いた穴あきのソースコードを用意してあるため，java を推奨する.

- 成績：課題の提出数（＋質）を基準とする.
  ⇒ 最期まですべて実装が正しくできれば優

- 質問や意見があれば kato@ai.lab.uec.ac.jp までお知らせください.

第1章　確率とビリーフ

1.1　確率

本節では，まず，確率 (probability) を定義する．確率を定義するためには，それが対象とする事象 (event) の集合を定義しなければならない．

定義 1 σ 集合体
Ω を標本空間 *(sample space)* とし，$\mathcal{A}$ が以下の条件を満たすならば σ **集合体** *(σ-field)* と呼ぶ．

1. $\Omega \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$ *(ただし，$A^c = \Omega \setminus A$)*
3. $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$

つまり，互いに素な事象の和集合により新しい事象を生み出すことができ，それら全ての事象を含んだ集合を σ 集合体と呼ぶ．
σ 集合体上で確率 (probability) は以下のように定義される．

定義 2　確率測度
今，σ 集合体 $\mathcal{A}$ 上で，次の条件を満たす測度 *(measure)*P を，**確率測度** *(probability measure)* とよぶ *(Kolmogorov 1933)*.

1. $A \in \mathcal{A}$ について, $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. 互いに素な事象列 $\{A_n\}_{n=1}^{\infty}$ に対して，

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

上の定義では,1. は確率が 0 から 1 の値をとること,2. は全事象の確率が 1 になること,3. は互いに素な事象の確率はそれぞれの確率の和で求められること, が示されている．特に 3 の条件を, 確率の加法性 (additivity) と呼ぶ．また，三つ組 $(\Omega, \mathcal{A}, P)$ を確率空間 (probability space) と呼ぶ.
以下，確率の重要な性質を導こう．
定義 1. の 2. より，$P(A^c) = P(\Omega \setminus A) = P(\Omega) - P(A) = 1 - P(A)$ となることがわかる．これより，以下の定理が成り立つ.

定理 1 余事象の確率　事象 A の**余事象** *(complementary event)* の確率は以下のとおりである.

$$P(A^c) = 1 - P(A)$$

また, 定義 1. の 2. より，$P(\phi) = P(\Omega^c) = 1 - P(\Omega) = 1 - 1 = 0$ となる．これより, 以下の定理が成り立つ.

定理 2　境界 $P(\phi) = 0$

さらに，$A \subset B$ のとき,$P(A)$ と $P(B)$ の関係について考えよう.

$A \subset B$ より，$\exists B' : B = A \cup B', A \cap B' = \phi$ が成り立つ．定義 1. の 3. より，$P(B) = P(A) + P(B')$, 定義 1. の 2. より，　$0 \leq P(B') \leq 1$, が成り立ち，結果として，$P(A) \leq P(B)$ が成り立つことがわかる．これを以下のように**単調性** (monotonicity) と呼ぶ.

定理 3　単調性
$A \subset B$ のとき $P(A) \leq P(B)$

定義 2. の 3. では，互いに素な事象の和の確率はそれぞれの事象の確率の和で求めることができた．では，互いに素ではない事象の和はどのように求められるのであろうか？以下のように求められる.

$A \cup B = (A \cap B^c) \cup (A^c \cap B) \cup (A \cap B)$ より，$P(A \cup B) = P(A \cap B^c) + P(A^c \cap B) + P(A \cap B)$ が成り立つ.

ここで，$P(A \cap B^c) = P(A) - P(A \cap B)$，$P(A^c \cap B) = P(B) - P(A \cap B)$ を代入して，$P(A \cap B^c) + P(A^c \cap B) + P(A \cap B) = P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) = P(A) + P(B) - P(A \cap B)$

これを以下のように**確率の和法則** (Additional low of probability) と呼ぶ.

定理 4　確率の和法則
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

1.2　主観確率

前節で確率を数学的に定義した．しかし，確率の実際的な解釈には二つの立場がある．最も一般的な解釈が，ラプラスの**頻度主義**である.

コインを何百回も投げて表が出た回数 (頻度) を数えて，その割合を求めることを考えよう．今，投げる回数を n, とし，表の出た回数を n_1 とすると,$n \to \infty$ のとき,

$$\frac{n_1}{n} \to \frac{1}{2}$$

となることが予想される．このように, 何回も実験を繰り返して n 回中, 事象 A が n_1 回出たとき，n_1/n を A の確率と解釈するのが頻度主義である.

しかし，この定義では真の確率は無限回実験をしなければ得られないので得ることは不可能である．また，科学的実験が可能な場合にのみ確率が定義され，実際の人間が扱う不確かさに比べて極めて限定的になってしまう．

一方, より広く確率を捉える立場として, 人間の個人的な**主観確率** (subjective probability) として解釈する立場がある．

ベイジアン (Bayesian; ベイズ主義者) は, 確率を主観確率として扱う．次節で導出されるベイズの定理を用いる人々をベイジアンだと誤解されているが, ベイズの定理は確率の基本定理で数学的に議論の余地のないものであり, 頻度主義者も用いる．

例えば, 松原 (2010) では以下のような主観確率の例が挙げられている．

1. 第三次世界大戦が 20XX 年までに起こる確率が 0.01
2. 明日, 会社の株式の価格が上がる確率が 0.35
3. 来年の今日, 東京で雨が降る確率が 0.5

ベイズ統計では, これらの主観確率は個人の意思決定のための信念として定義され, **ビリーフ** (belief) と呼ばれる．当然，頻度論的確率を主観確率の一種とみなすことができるが，その逆は成り立たない．本書では, ベイズ統計の立場に立ち, 確率をビリーフの立場で解釈する．ビリーフの具体的な決定の仕方などは厳密な理論に興味のある読者は Bernardo and Smith(1994), Berger (1985) を参照されたい．

1.3 条件付き確率と独立

本節では，条件付き確率と独立を定義する．

定義 3 条件付き確率
$A \in \mathcal{A}$, $B \in \mathcal{A}$ について，事象 B が起こったという条件の下で，事象 A が起こる確率を**条件付き確率** *(conditional probablity)* と呼び,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

で示す．

このとき，$P(A|B) = \frac{P(A \cap B)}{P(B)}$ より以下の乗法公式が成り立つ．

定理 5 乗法公式

$$P(A \cap B) = P(A|B)P(B)$$

このとき,$P(A \cap B)$ を A と B の**同時確率** (joint probability) と呼ぶ.

次に，事象の独立を以下のように定義する.

定義 4　独立

ある事象の生起する確率が，他のある事象が生起する確率に依存しないとき，*2* つの事象は**独立** *(independent)* であるという．すなわち事象 A と事象 B が独立とは $P(A|B) = P(A)$ であり，

$$P(A \cap B) = P(A)P(B)$$

が成り立つことをいう.

さらに乗法公式を一般化すると以下のチェーンルールが導かれる.

$$P(A \cap B \cap C) = P(A|B \cap C)P(B \mid C)P(C)$$

これは,3 個以上の事象にも拡張できるので, **チェーンルール** (Chain rule) は以下のように書ける.

定理 6 チェーンルール N 個の事象 $\{A_1, A_2, \ldots, A_N\}$ について

$$P(A_1 \cap A_2 \cap \cdots \cap A_N) = P(A_1|A_2 \cap A_3 \cap \ldots \cap A_N)P(A_2 \mid A_3 \cap A_4 \cap \ldots \cap A_N) \cdots P(A_N)$$

が成り立つ.

1.4　ベイズの定理

本節では，条件付き確率より，ベイズ統計にとって最も重要なベイズの定理を導出する.

ベイズの定理を導出する前に，互いに背反な事象 $A_1, A_2, \cdots, A_n$，$(A_i \in \mathcal{A})$ が全事象 Ω を分割しているとき，事象 $B \in \mathcal{A}$ について以下が成り立つことがわかる.

$$\sum_{i=1}^{n} P(A_i)P(B|A_i) = \sum_{i=1}^{n} P(A_i)\frac{P(A_i \cap B)}{P(A_i)}$$

$$= \sum_{i=1}^{n} P(A_i \cap B) = P(\Omega \cap B) = P(B)$$

これを以下の全確率の定理と呼ぶ.

定理 7　**全確率の定理** *(total probability theorem)*

互いに背反な事象 $A_1, A_2, \cdots, A_n$，$(A_i \in \mathcal{A})$ が全事象 Ω を分割しているとき，事象 $B \in \mathcal{A}$ について, $P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$ が成り立つ.

全確率の定理より，$P(B)=\sum_{i=1}^{n}P(A_i)P(B|A_i)$，従って

$$\frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n}P(A_i)P(B|A_i)}=\frac{P(A_i)P(B|A_i)}{P(B)}=\frac{P(A_i\cap B)}{P(B)}=P(A_i|B)$$

が成り立つ．これが以下の**ベイズの定理**である．

定理 8　ベイズの定理 *(Bayes' Theorem)*
互いに背反な事象 $A_1, A_2, \cdots, A_n$ が全事象 Ω を分割しているとする．このとき，事象 $B\in\mathcal{A}$ について，

$$P(A_i|B)=\frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n}P(A_i)P(B|A_i)}$$

が成り立つ．

課題 1

1 から 3 の目が赤色で塗られており，4 から 6 の目は青色で塗られているさいころがある．今，このさいころを投げて青色の目が出た時，この目が偶数である確率を求めよ．

課題 2

表と裏の面が赤か青で塗られている 3 枚のカード A, B, C があり，それぞれのカードの面の色は次のようになっている．

- カード A：両面とも青色で塗られている．
- カード B：片面が赤色、もう片面が青色で塗られている．
- カード C：両面とも赤色で塗られている．

このカード 3 枚を袋に入れてよく混ぜて，目をつぶったまま 1 枚を取り出し，机の上に置いて目を開けるとカードは赤色だった．ひっくり返した面も赤色である確率を求めよ．

課題 3

一郎、二郎、三郎、四郎の 4 人がボウリングでストライクを出す確率は 50%, 70%, 90%, 98%である．4 人のうち 1 人が球を投げてストライクを出したときに，それが一郎である確率はいくらか．

課題 4

ある映画の試写会を行い，満足度のアンケート調査を行った．試写会に参加したのは 300 人でそのうち女性が 180 人であり，満足したと回答したのは男性の 50 %，女性の 75 %であった．この映画を見て満足しなかったと答えた人が女性である確率はいくらか．

課題 5

人の「疲れ」を判定する機械が発明された．この機械に人が入ると「疲れている」か「疲れていない」かを判定してくれる画期的なものである．この機械を使うと，疲れている人の 95 %を「疲れている」と判定し，疲れていない人の 98 %を「疲れていない」と判定するということが分かっている．人の 70%は疲れているという研究結果があるとき，ある人がこの機械に入って「疲れている」と判定された場合に実際に疲れている確率はいくらか．

課題 6

キリストの弟子たちはキリストの復活を望んでいました．あまりに望みが強すぎて少し似ているだけの人でもキリストに見えてしまうことがあります．弟子がキリストの復活を見たと証言する事象を A，実際にキリストが復活したという事象を B とする．$P(A \mid B) = 1.0$，$P(A \mid \neg B) = 0.5$ とする．$P(B)$ を入力とし，ある弟子がキリストの復活を見たと証言したとき本当にキリストが復活した確率 $P(B \mid A)$ を出力するプログラムを作成せよ．

課題 6 では，もともとのキリストが復活する確率 $P(B)$ が，弟子の報告により $P(B \mid A)$ にビリーフが更新されていることがわかる．すなわち，弟子の証言によって事前のビリーフが事後のビリーフに更新されたのである．このとき，ベイズ統計では，弟子の証言を「エビデンス」(evidence) と呼び，事前のビリーフを「事前確率」(prior probability)，事後のビリーフを「事後確率」 (posterior probability) と呼ぶ．

課題 7

課題 6 において，B の事前確率を $P(B) = 0.000001$ とする．キリストの復活を見たと証言した弟子の人数 x を引数とし，x 人の弟子たちのエビデンスを所与として本当にキリストが復活した確率を出力するプログラムを作成せよ．また，x を横軸にとり，x 人の弟子たちのエビデンスを所与として本当にキリストが復活した確率を縦軸にとったグラフを作成し，考察せよ．

課題 8

モンティ・ホール問題：三つの扉があり一つは正解で二つは不正解である．挑戦者は三つの中から一つ扉を選ぶ．司会者は答えを知っており，残り二つの扉の中で不正解の扉を一つ選んで開ける．挑戦者は残り二つの扉の中から好きな方を選べる．このとき扉を変えるべきか？変えないべきか？

1.5 確率変数

一つの試行の結果を**標本点** $\omega \in \Omega$ と呼ぶ．この標本点 $\omega \in \Omega$ は，何らかの測定によって観測される．この測定のことを，確率論では**確率変数** (random variable) と呼ぶ．例えば，コインを n 回投げるという試行について，表が出る回数 X は確率変数である．このとき，標本点 ω は表・裏のパターンが n 個あり得るので 2^n 通りあり，X は 0 から n までの値をとる．

定義 5 確率変数 X が，高々加算個の実数の集合 $\mathcal{X} = \{x_1, x_2, \ldots, \}$ の中の値をとるならば，X は離散であるという．離散確率変数 X のとりえる値 $x_k \in \mathcal{X}$ を，その確率 $p = P(X = x_k)$ に対応づける写像 $p : \mathcal{X} \to [0, 1]$ を X の**離散確率分布** *(discrete probability distribution)* とよぶ．

離散確率分布 p は

$$\begin{aligned} p(x) &\geq 0 (x \in \mathcal{X}), \\ \sum_{x \in \mathcal{X}} p(x) &= 1 \end{aligned}$$

を満たす．逆に，これら二つの条件を満たす $\mathcal{X}$ 上の関数 p を確率分布としてもつ確率変数が存在する．

また，複数の確率変数を持つ確率分布について以下のように定義しよう．

定義 6 今，m 個の確率変数を持つ確率分布 $p(x_1, x_2, \cdots, x_m)$ を変数 $x_1, x_2, \cdots, x_m$ の**同時確率分布** *(joint probability distribution)* と呼ぶ．

同時確率分布から特定の変数の分布を以下のように求めることができる．

定義 7 x_i のみに興味がある場合，同時確率分布から x_i の確率分布は，

$$p(x_i) = \sum_{x_1, \ldots, x_{i-1}, x_{i+1} \ldots, x_m} p(x_1, x_2, \cdots, x_m),$$

で求められる．

1.6　尤度原理

本節では，確率分布のパラメータを定義し，データからパラメータを推定するための尤度原理を紹介する．

定義 8 パラメータ空間と確率分布
k 次元パラメータ集合を $\Theta = \{\theta_1, \theta_2, \cdots, \theta_k\}$ と書くとき，確率分布は以下のような関数で示される．

$$f(x|\Theta)$$

すなわち，確率分布 $f(x|\Theta)$ の形状はパラメータ Θ のみによって決定され，パラメータ Θ のみが確率分布 $f(x|\Theta)$ を決定する情報である．

例 1 コインを n 回投げた時，表が出る回数を確率変数 x とした確率分布は以下の二項分布に従う．

$$f(x|\theta, n) = {}_n\mathrm{C}_x \theta^x (1-\theta)^{n-x}$$

ここで，θ は，コインの表が出る確率のパラメータを示す．

あるデータについて，特定の確率分布を仮定した場合，データからそのパラメータを推定することができる．そのひとつの方法では，以下の尤度を用いる．

定義 9 尤度
$X = (X_1, \cdots, X_i, \cdots, X_n)$ が確率分布 $f(X_i|\theta)$ に従う n 個の確率変数とする．n 個の確率変数に対応したデータ $x = (x_1, \cdots, x_n)$ が得られたとき，

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

を**尤度関数** *(Likelihood function)* と定義する *(Fisher,1925)*.

例 2 コインを n 回投げた時，表が出た回数が x 回であったときのコインの表が出るパラメータ θ の尤度は

$$L(\theta|n, x) \propto {}_n\mathrm{C}_x \theta^x (1-\theta)^{n-x},$$

もしくは，

$$L(\theta|n, x) \propto \theta^x (1-\theta)^{n-x}$$

でもよい．

尤度は，データ x パターンが観測される確率に比例する，パラメータ θ の関数である．尤度は確率の定義を満たす保証がないために確率とは呼べないが，これを厳密に確率分布として扱うアプローチが後述するベイズアプローチである．

尤度を最大にするパラメータ θ を求めることは，データ x を生じさせる確率を最大にするパラメータ θ を求めることになり，その方法を**最尤推定法** (Maxmimum Likelihood Estimation; MLE) と呼ぶ．

定義 10 最尤推定量
データ x を所与として，以下の尤度最大となるパラメータを求める時，

$$L(\hat{\theta}|x) = \max\{L(\theta|x) : \Theta \in \mathcal{C}\}$$

$\hat{\theta}$ を**最尤推定量** *(maximum likelihood estimator)* と呼ぶ *(Fisher,1925)*．ただし，$\mathcal{C}$ はコンパクト集合を示す．

推定値の望ましい性質の中で以下の一致性が知られている．

定義 11　強一致性
推定値 $\hat{\theta}$ が真のパラメータ θ^* に概収束する時，$\hat{\theta}$ は**強一致推定値** *(strongly consistent estimator)* であるという．

$$P(\lim_{n\to\infty} \hat{\theta} = \theta^*) = 1.0$$

つまり，データ数が大きくなると推定値が必ず真の値に近づいていくととき，その推定量を強一致推定値と呼ぶ．

このとき，最尤推定値について以下が成り立つ．

定理 9 最尤推定値の一致性
最尤推定値 $\hat{\theta}$ は真のパラメータ θ^* の強一致推定値である．*(Wald, 1949)*

また，一致推定値の漸近的な分布は以下で与えられる．

定義 12
θ^* の推定値 $\hat{\theta}$ が**漸近正規推定量** *(asymptotically normal estimator)* であるとは，$\sqrt{n}(\hat{\theta} - \theta^*)$ の分布が正規分布に分布収束することをいう．すなわち，任意の $\theta^* \in \Theta^*$ と任意の実数に対して

$$\lim_{n\to\infty} P(\frac{\sqrt{n}(\hat{\theta} - \theta^*)}{\sigma(\theta^*)} \leq x) = \Phi(x)$$

このことを，$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{as} N(0, \sigma^2(\theta^*))$ と書く． $\sigma^2(\theta^*)$ を**漸近分散** *(asymptotic variance)* という．

すなわち，一致推定値の漸近的な分布は正規分布になる．また，一致推定値の誤差は以下のように得られる．

定理 10 確率密度関数が**正則条件**（*Regular condition*）の下で，微分可能のとき，最尤推定量は漸近分散 $I(\theta^*)^{-1}$ を持つ漸近正規推定量である． $I(\theta^*) = E_\theta[(\frac{\partial}{\partial\theta}\ln L(\theta|\boldsymbol{x}))^2]$ を *Fischer* の情報行列と呼ぶ.

課題 9

母集団の確率分布がポアソン分布

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad (\lambda > 0, x = 0, 1, \cdots)$$

について n 回の観測を行ったところデータ $\{x_1, x_2, \cdots, x_n\}$ を得た．λ を最尤推定せよ.

1.7 ベイズ推定

前節で述べた尤度原理は古典的な頻度主義であるフィッシャー統計学の流儀である．フィッシャー統計の尤度原理に対して，ベイズ統計では以下の事後分布を用いてパラメータを推定する.

定義 13 事後分布
$X = (X_1, \cdots, X_{\mathrm{n}})$ が独立同一分布 $f(x|\theta)$ に従う n 個の確率変数とする．n 個の確率変数に対応したデータ $x = (x_1, \cdots, x_n)$ が得られた時,

$$p(\theta|x) = \frac{p(\theta)\prod_{i=1}^n f(x_i|\theta)}{\int_\Theta p(\theta)\prod_{i=1}^n f(x_i|\theta)d\theta}$$

を事後分布（*posterior distribution*）と呼び， $p(\theta)$ を事前分布（*prior distribution*）と呼ぶ.

ベイズ統計では，事後分布を最大にするようにパラメータ推定を行う.

定義 14 *MAP* 推定値
データ x を所与として，以下の事後分布最大となるパラメータを求める時,

$$\hat{\theta} = \arg\max\{p(\theta|x) : \theta \in C\}$$

$\hat{\theta}$ を**ベイズ推定値**（*Bayesian estimator*）または，**事後分布最大化推定値:MAP 推定値**（*Maximum A Posterior estimator*）と呼ぶ.
Note:
ベイズ推定は，すべての確率空間で成り立つわけではない．パラメータの事前確率が確率の公理を満たすときにのみ成立する.

また, ベイズ推定値では,MAP 推定値が予測を最適しないことが知られている．予測を最適化するベイズ推定値は, 事後分布を最大化せずに, 事後分布の期待値となる推定値を用いる.

定義 15 *EAP* 推定値
データ x を所与として，以下の事後分布によるパラメータの期待値を求める時,

$$\hat{\theta} = E\{\theta\{p(\theta|x) : \theta \in C\}\}$$

$\hat{\theta}$ を**期待事後推定値:EAP 推定値**（*Expected A Posterior estimator*）と呼ぶ.

ベイズ推定値も強一致性を持つ.

定理 11 ベイズ推定の一致性
ベイズ推定において推定値 $\hat{\theta}$ が真のパラメータ θ^* の強一致推定値となるような事前分布が設定できる

また，ベイズ推定値も漸近的正規性を持ち，誤差を計算できる.

定理 12 ベイズ推定の漸近正規性
事後確率密度関数が正則条件（*Regular condition*）の下で，微分可能のとき，ベイズ推定値が漸近分散 $I(\theta^*)^{-1}$ を持つ漸近正規推定値となる事前分布を設定できる.

ベイズ統計では，どのように事前分布を設定するかが問題となる．事前分布はユーザが知識を十分に持つ場合，自由に決定してよいが，事前に知識を持たない場合にはどのように設定すれば良いのであろうか．このようなときの事前分布を無情報事前分布と呼び，次節のような分布が提案されている.

1.8 無情報事前分布

1.8.1 自然共役事前分布（natural conjugate prior distribution）

ベイズ統計の中で最も一般的で，ベイズ的な有効性を発揮できると考えられるのが，この自然共役事前分布である．データを得る前の事前分布とデータを得た後の事後分布は，データの有無に係らず，分布の形状は同一のほうが自然であろう．そこで，事前分布と事後分布が同一の分布族に属する時，その事前分布を**自然共役事前分布**（natural conjugate prior distribution）と呼ぶ．ここでは，特にこの自然共役事前分布を中心にベイズ的推論を行なうようにする.

例 3 二項分布

$$f(x|\theta, n) = {}_nC_x\theta^x(1-\theta)^{n-x}$$

コインを投げて n 回中 x 回表が出たときの確率 θ をベイズ推定しよう.

尤度関数は，${}_nC_x\theta^x(1-\theta)^{n-x}$ であり，二項分布の自然共役事前分布は，以下の**ベータ分布** $(Beta(\alpha,\beta))$ である．

$$p(\theta|\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

事後分布は，

$$p(\theta|n,x,\alpha,\beta)=\frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

とやはりベータ分布となる．

対数をとり，以下の対数事後分布を最大化すればよい．

$$\begin{aligned}&\log p(\theta|n,x,\alpha,\beta)\\ =\ &\log\frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}+(x+\alpha-1)\log\theta+(n-x+\beta-1)\log(1-\theta)\end{aligned}$$

$\frac{\partial\log p(\theta|n,x,\alpha,\beta)}{\partial\theta}=0$ のとき，対数事後分布は最大となるので，

$$\begin{aligned}&\frac{\partial\log p(\theta|n,x,\alpha,\beta)}{\partial\theta}\\ =\ &\frac{x+\alpha-1}{\theta}-\frac{n-x+\beta-1}{1-\theta}=\frac{x+\alpha-1-x\theta-\alpha\theta+\theta-n\theta+x\theta-\beta\theta+\theta}{\theta(1-\theta)}\\ =\ &\frac{x+\alpha-1-(n+\alpha+\beta-2)\theta}{\theta(1-\theta)}=0\end{aligned}$$

$\theta(1-\theta)\neq 0$ より

$$\theta=\frac{x+\alpha-1}{n+\alpha+\beta-2}$$

がベイズ推定値となる．さて，α,β は事前分布のパラメータであるが，これを**ハイパーパラメータ**（*Hyper parameter*）と呼ぶ．このハイパーパラメータによって，事前分布は様々な形状をとる．

例えば，事前分布が一様となる場合の推定値は，$\widehat{\theta}=\frac{x}{n}$ となり，最尤解に一致する．

例 4 以下のどちらのかけを選ぶと得か？

1. 赤玉と白玉が同じ個数入った壺から一つ玉を取り出し，それが赤玉であったら 1 万円もらえる．白玉であったら 1 万円支払う．
2. 赤玉と白玉が入っている壺から一つ玉を取り出し，それが赤玉であったら 1 万円もらえる．白玉であったら 1 万円支払う．

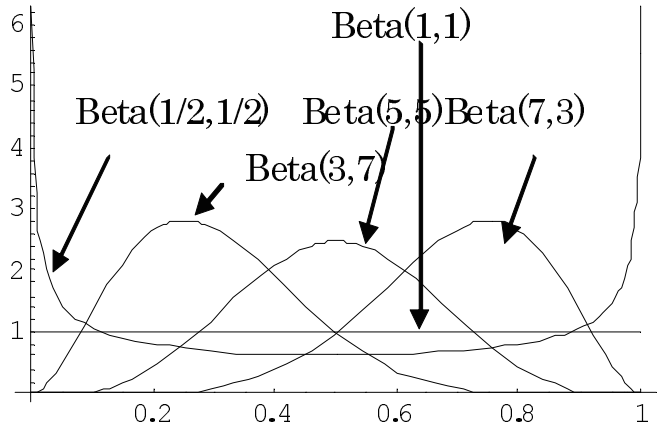

図 1.1: ハイパーパラメータと事前分布の形状

それぞれのかけにおいて赤玉を取り出す事象 A の確率を求める．かけ 1 において赤玉を取り出す確率は

$$p(A \mid \text{かけ } 1) = \frac{1}{2}$$

である．また，赤玉を取り出す確率を $p(A) = \psi$ とすると，かけ 2 において赤玉を取り出す確率は

$$p(A \mid \text{かけ } 2) = \int_0^1 \psi p(\psi) d\psi$$

である．ここで，赤玉を取り出す確率の自然共役事前分布 $p(\psi)$ は以下のベータ分布である．

$$p(\psi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \psi^{\alpha-1}(1-\psi)^{\beta-1}$$

よって，

$$\begin{aligned} p(A \mid \text{かけ } 2) &= \int_0^1 \psi \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \psi^{\alpha-1}(1-\psi)^{\beta-1} d\psi \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \psi^{\alpha}(1-\psi)^{\beta-1} d\psi \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\ &= \frac{\alpha}{\alpha+\beta} \end{aligned}$$

となる．いま，赤玉と白玉の出やすさについての情報が与えられていないため，赤玉と白玉のでる事前確率は等しい $(\alpha = \beta)$ とするのが自然である．このとき，$p(A \mid \text{かけ } 2) = 1/2$ となり，かけ 1 とかけ 2 の赤玉を引く確率は等しいため，どちらのかけを選んでもよい．

課題 10

以下のどちらのかけを選ぶと得か？

1. 赤玉と白玉が同じ個数入った壺から一つ玉を取り出し，それが赤玉であったら 1 万円もらえる．白玉であったら 1 万円支払う．取り出した玉を壺に戻す．これを 10 回繰り返す．

2. 赤玉と白玉が入っている壺から一つ玉を取り出し，それが赤玉であったら 1 万円もらえる．白玉であったら 1 万円支払う．取り出した玉を壺に戻す．これを 10 回繰り返す．

赤玉の出る回数を x，試行回数を n とすると，$p(x \mid \psi, n)$ は以下の二項分布に従う．

$$p(x \mid \psi, n) = \binom{n}{x} \psi^x (1-\psi)^{n-x}$$

したがって，かけ 1 において赤玉を x 回取り出す確率は $p(x \mid \psi = 1/2, n = 10)$ である．かけ 2 において赤玉を x 回取り出す確率は

$$p(x \mid n = 10) = \int_0^1 p(x \mid n = 10, \psi) p(\psi) d\psi$$

である．
横軸に x をとり，$p(x \mid \psi = 1/2, n = 10)$ と $p(x \mid n = 10)$ のグラフを作成せよ．ただし，$p(x \mid n = 10)$ については事前分布 $p(\psi)$ のハイパーパラメータ α と β がともに $1/2, 1, 2, 5, 10$ の場合についてそれぞれグラフを作成せよ．作成したグラフからかけ 1 とかけ 2 の違いを考察せよ．

1.9 マルコフ連鎖モンテカルロ法 (MCMC) 法

ベイズ推定では，以下のパラメータの事後分布 $p(\theta \mid \boldsymbol{x})$ を推定し，得られた分布形に基づいて推定値を求める．

$$p(\theta \mid \boldsymbol{x}) = \frac{p(\theta) \prod_{i=1}^n f(x_i \mid \theta)}{\int_\Theta p(\theta) \prod_{i=1}^n f(x_i \mid \theta) d\theta}$$

具体的には，θ の MAP 推定値 $\underset{\theta}{\operatorname{argmax}}\, p(\theta \mid \boldsymbol{x})$ や EAP 推定値 $E_\theta[p(\theta \mid \boldsymbol{x})]$ を求める．しかし，事後分布 $p(\theta \mid \boldsymbol{x})$ の分母における θ の積分計算は一般に膨大な時間を要する．このため，事後分布は近似的に計算する必要があり，近似手法の一つとしてサンプリング法がしばしば用いられる．今，事後分布 $p(\theta \mid \boldsymbol{x})$ から T 個のサンプリング $\{\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(T)}\}$ が得られたとすると，EAP 推定値は以下のように近似できる．

$$E_\theta[p(\theta \mid \boldsymbol{x})] \approx \frac{1}{T} \sum_{t=1}^T \theta^{(t)}$$

しかし，一般に事後分布は解析的に表現できず，パラメータを直接サンプリングすることはできない．そこで，複雑な事後分布でもサンプリング近似できる手法として MCMC 法が提案された．この章では MCMC 法のうち最も基本的なギブスサンプリング法とメトロポリスヘイスティング法について紹介する．

1.9.1 ギブスサンプリング

事後分布 $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ から直接にはサンプリングできないが，パラメータごとの条件付き分布 $p(\theta_i \mid \boldsymbol{x}, \boldsymbol{\theta}^{\setminus i})$ からはサンプリングができる場合に利用できる手法（ここで，$\boldsymbol{\theta}^{\setminus i} = \boldsymbol{\theta} \setminus \{\theta_i\}$）パラメータごとの条件付き分布から順にサンプリングを繰り返す．いま，θ_i について t 回目の繰り返し時にサンプリングされた値を $\theta_i^{(t)}$ と表すと，事後分布から T 個のサンプルを得るギブスサンプリングのアルゴリズムは以下のように書ける．

ギブスサンプリング

$\{\theta_i^{(0)} : i = 1, \cdots, K\}$ をランダムに初期化.
for $t = 1$ to T:
  $\theta_1^{(t)} \sim p(\theta_1 \mid \boldsymbol{x}, \theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_K^{(t-1)})$ をサンプリングする.
  $\theta_2^{(t)} \sim p(\theta_2 \mid \boldsymbol{x}, \theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_K^{(t-1)})$ をサンプリングする.
  $\vdots$
  $\theta_i^{(t)} \sim p(\theta_i \mid \boldsymbol{x}, \theta_1^{(t)}, \cdots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \cdots, \theta_K^{(t-1)})$ をサンプリングする.
  $\vdots$
  $\theta_K^{(t)} \sim p(\theta_1 \mid \boldsymbol{x}, \theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_{K-1}^{(t)})$ をサンプリングする.
end for
return $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(T)}\}$

例 5 $x_i \sim N(\mu, \sigma^2)$ とする n 個のサンプル $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ を所与としてパラメータ μ, σ^2 を推定．パラメータの同時事後分布はサンプリング可能な既知の分布とならないため，この分布から直接サンプリングすることはできない．

$$p(\mu, \sigma^2 \mid \boldsymbol{x}) = \frac{p(\mu)P(\sigma^2)\prod_{i=1}^n f(x_i \mid \mu, \sigma)}{\int\int p(\mu)P(\sigma^2)\prod_{i=1}^n f(x_i \mid \mu, \sigma)d\mu d\sigma}$$

しかし，条件付き分布 $p(\mu \mid \boldsymbol{x}, \sigma^2), p(\sigma^2 \mid \boldsymbol{x}, \mu)$ はそれぞれ既知の分布になるため，サンプリングが可能．μ, σ^2 の事前分布に一様分布を仮定すると

$$p(\mu \mid \boldsymbol{x}, \sigma^2) = N\left(\frac{1}{N}\sum_{i=1}^n x_i, \frac{\sigma^2}{N}\right),$$

$$p(\sigma^2 \mid \boldsymbol{x}, \mu) = IG\left(\frac{n}{2} + 1, \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}\right).$$

正規分布や逆ガンマ分布 $IG()$ からの乱数生成手法は既知であり，多くのプログラミング言語にはこれらの乱数生成器が実装されている．

1.10 メトロポリスヘイスティングス

条件付き分布からもサンプリングできないときに利用できるのがメトロポリスヘイスティング法である．現在のパラメータ値 $\boldsymbol{\theta}$ の付近の候補値 $\boldsymbol{\theta}^*$ を提案分布（proposal distribution）$q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$ から生成.
一般に $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) = MN(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}, \boldsymbol{I}\sigma)$
MN は多次元正規分布，$\boldsymbol{I}$ は単位行列， σ は微小な値 (0.01 等).
以下の採択確率に基づいて候補値 $\boldsymbol{\theta}^*$ を採択する．

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \min\left\{1, \frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{x})q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta} \mid \boldsymbol{x})q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})}\right\}.$$

特に，$q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) = MN(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}, \boldsymbol{I}\sigma)$ のとき，

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \min\left\{1, \frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{x})}{p(\boldsymbol{\theta} \mid \boldsymbol{x})}\right\}.$$

棄却された場合には $\boldsymbol{\theta}^* = \boldsymbol{\theta}$ とする．採択確率における事後確率の多重積分は以下のように消去できるため，採択確率を高速に計算できる．

$$\frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{x})}{p(\boldsymbol{\theta} \mid \boldsymbol{x})} = \frac{\dfrac{p(\boldsymbol{\theta}^*)\prod_{i=1}^n f(x_i \mid \boldsymbol{\theta}^*)}{\int_\Theta p(\boldsymbol{\theta}^*)\prod_{i=1}^n f(x_i \mid \boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}}{\dfrac{p(\boldsymbol{\theta})\prod_{i=1}^n f(x_i \mid \boldsymbol{\theta})}{\int_\Theta p(\boldsymbol{\theta})\prod_{i=1}^n f(x_i \mid \boldsymbol{\theta})d\boldsymbol{\theta}}} = \frac{p(\boldsymbol{\theta}^*)\prod_{i=1}^n f(x_i \mid \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta})\prod_{i=1}^n f(x_i \mid \boldsymbol{\theta})}.$$

しかし，メトロポリスヘイスティングスでは，パラメータ数が増加すると，パラメータ値が改悪される方向に進むときに，採択確率 $\frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{x})}{p(\boldsymbol{\theta} \mid \boldsymbol{x})}$ が極端に小さくなり，更新が進まなくなることがある．この問題を緩和する手法として，パラメータごとに他のパラメータの条件付分布を求めてメトロポリスヘイスティングスを実行するメトロポリスヘイスティングス with ギブス法が知られている．以下はそのアルゴリズムである．

メトロポリスヘイスティングス with ギブス

$\{\theta_i^{(0)} : i = 1, \cdots, K\}$ をランダムに初期化.
for $t = 1$ to T:
　for $i = 1, \cdots, K$:
　・現在の値を所与として θ_i の候補値 θ_i^* を生成.

$$\theta_i^* \sim N(\theta_i^{(t-1)}, \sigma^2)$$

　・以下の採択確率に基づき θ_i^* を $\theta_i^{(t)}$ として採択または棄却.

$$\alpha(\theta_i^*, \theta_i^{(t-1)}) = \min\left\{1, \frac{p(\theta_i^* \mid \boldsymbol{x}, \theta_1^{(t)}, \cdots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \cdots, \theta_K^{(t-1)})}{p(\theta_i^{(t-1)} \mid \boldsymbol{x}, \theta_1^{(t)}, \cdots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \cdots, \theta_K^{(t-1)})}\right\}$$

　end for
end for
result $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(T)}\}$

アルゴリズム初期のサンプルは，初期値に依存するため，一般に一定回数サンプリングを繰り返した後のサンプルを利用する．初期値に依存しなくなったとみなすまでの時間をバーンイン期間と呼ぶ．また，メトロポリスヘイスティングスはサンプル間の自己相関（隣接するサンプル間の依存性）が高いため，一定区間でサンプルを間引いて用いる必要がある．間引く間隔をインターバル期間と呼ぶ．バーンイン期間を B，インターバル期間を V としてメトロポリスヘイスティングス with ギブスに適用したアルゴリズムは以下である．

メトロポリスヘイスティングス with ギブス（修正版）

$\{\theta_i^{(0)} : i = 1, \cdots, K\}$ をランダムに初期化.
for $t = 1$ to $B + TV$:
　for $i = 1, \cdots, K$:
　・現在の値を所与として θ_i の候補値 θ_i^* を生成.

$$\theta_i^* \sim N(\theta_i^{(t-1)}, \sigma^2)$$

　・以下の採択確率に基づき θ_i^* を $\theta_i^{(t)}$ として採択または棄却.

$$\alpha(\theta_i^*, \theta_i^{(t-1)}) = \min\left\{1, \frac{p(\theta_i^* \mid \boldsymbol{x}, \theta_1^{(t)}, \cdots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \cdots, \theta_K^{(t-1)})}{p(\theta_i^{(t-1)} \mid \boldsymbol{x}, \theta_1^{(t)}, \cdots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \cdots, \theta_K^{(t-1)})}\right\}$$

　end for
end for
result $\{\boldsymbol{\theta}^{(B)}, \boldsymbol{\theta}^{(B+V)}, \boldsymbol{\theta}^{(B+2V)}, \cdots, \boldsymbol{\theta}^{(B+TV)}\}$

第2章　ベイズ分類器

1 章まででベイズの基本を学んできた．本章から本実験のテーマであるベイズ分類器を学ぶ．

ベイズ分類器は次のように定義される．

定義 16 確率変数集合 $\mathbf{X} = \{X_1, \cdots, X_n\}$ の各変数の実現値 $x_1, \cdots, x_n$ を入力とし，離散確率変数 X_0 の値 $\widehat{x}_0$ を出力する以下の関数をベイズ分類器と呼ぶ．

$$\widehat{x}_0 = \underset{c \in \{1, \cdots, r_0\}}{\operatorname{argmax}} \; p(c \mid x_1, \cdots, x_n) \tag{2.1}$$

ここで，各変数 $X_i, (i = 0, 1, \cdots, n)$ は $\{1, \cdots, r_i\}$ から一つの値をとるとする．

X_0 を目的変数，$X_i \in \mathbf{X}, (i = 1, \cdots, n)$ をその説明変数と呼ぶ．式 (2.1) の $p(c \mid x_1, \cdots, x_n)$ はベイズの定理により，以下のように求められる．

$$\begin{aligned} \underset{c \in \{1, \cdots, r_0\}}{\operatorname{argmax}} \; p(c \mid x_1, \cdots, x_n) &= \underset{c \in \{1, \cdots, r_0\}}{\operatorname{argmax}} \; \frac{p(c)p(x_1, \cdots, x_n \mid c)}{p(x_1, \cdots, x_n)} \\ &= \underset{c \in \{1, \cdots, r_0\}}{\operatorname{argmax}} \; p(c)p(x_1, \cdots, x_n \mid c) \end{aligned} \tag{2.2}$$

このとき，$p(x_1, \cdots, x_n \mid c)$ はモデルのデータ $\langle x_1, \cdots, x_n \rangle$ に対する尤度に対応し，式 (2.1) を識別関数と呼ぶ．ベイズ分類器の尤度 $p(x_1, \cdots, x_n \mid c)$ の計算法は仮定するモデルによってさまざまに変化する．モデルの制約が強い場合（単純なモデルの場合）は計算が容易であるが，モデルが複雑になるに従い計算も複雑になる．以下、このモデルを単純なものから徐々に一般化して学んでいくことにしよう．

2.1　Naive Bayes

まず最初に最も単純な構造をもつベイズ分類器である Naive Bayes を学ぼう．Naive Bayes では，図 2.1 のように，目的変数が与えられた際，説明変数間の条件付き独立を仮定している．これにより, 同時確率分布を以下のように，単純な確率の積で表すことができる．

$$p(X_0, X_1, ..., X_n) = p(X_0) \prod_{i=1}^{n} p(X_i \mid X_0)$$

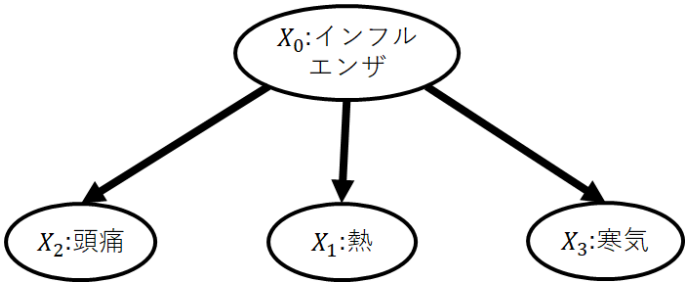

図 2.1: インフルエンザ判別に用いられる Naive Bayes の例

ここで，$p(X_0 = c)$ $(c = 1, \cdots, r_0)$，$p(X_i = k \mid X_0 = c)$ $(i = 1, \cdots, n; c = 1, \cdots, r_0; k = 1, \cdots, r_i)$ を示すパラメータをそれぞれ $\theta_{X_0=c}, \theta_{X_i=k|X_0=c}$ で表すと，説明変数のデータ（インスタンス）$x_1, \cdots, x_n$ に対する Naive Bayes の識別関数は以下で表される．

$$\widehat{x}_0 = \underset{c \in \{1,\cdots,r_0\}}{\operatorname{argmax}} \theta_{X_0=c} \prod_{i=1}^{n} \theta_{X_i=x_i|X_0=c} \tag{2.3}$$

識別関数は、非常に少ないパラメータを元に計算することが可能であり，モデル全体のパラメータ数は変数数に対して線形関数的に増加するに留まる．

今，全変数がインスタンス化されたサンプルが N 個あり，t 番目のサンプルを $\mathbf{d}^t = \langle x_0^t, x_1^t, \cdots, x_n^t \rangle$ と表し，訓練データを $D = \langle \mathbf{d}^1, \cdots, \mathbf{d}^t, \cdots, \mathbf{d}^N \rangle$ と表すと，D を所与とした時の Naive Bayes の対数尤度は以下で表される．

$$L = \sum_{c=1}^{r_0} N_{X_0=c} \log \theta_{X_0=c} + \sum_{i=1}^{n} \sum_{c=1}^{r_0} \sum_{k=1}^{r_i} N_{X_i=k, X_0=c} \log \theta_{X_i=k|X_0=c} \tag{2.4}$$

ここで，$N_{X_0=c}$ は D において $X_0 = c$ となる頻度を表し，$N_{X_i=k,X_0=c}$ は D において $X_i = k$ かつ $X_0 = c$ となる頻度を表す．さらに，Naive Bayes の最尤推定量は以下のように表される．

$$\hat{\theta}_{X_0=c} = \frac{N_{X_0=c}}{N}, \quad \hat{\theta}_{X_i=k|X_0=c} = \frac{N_{X_i=k,X_0=c}}{N_{X_0=c}} \tag{2.5}$$

課題 11

Naive Bayes のパラメータの最尤推定量 $\hat{\theta}_{X_0=c}, \hat{\theta}_{X_i=k|X_0=c}$ が (2.5) 式になることを確かめよ.
ヒント 1：束縛条件 $\sum_{c'=1}^{r_0} \theta_{X_0=c'} - 1 = 0$ を含んだラグランジュの未定乗数法を用いると，ラグランジェ関数 F_{X_0} は

$$F_{X_0} = L + \lambda \left(\sum_{c'=1}^{r_0} \theta_{X_0=c'} - 1 \right)$$

である．ここで，L は Naive Bayes の対数尤度（(2.4) 式）である．次式を満たす $\theta_{X_0=c}$ が求める最尤推定量 $\hat{\theta}_{X_0=c}$ である．

$$\frac{\partial F_{X_0}}{\partial \theta_{X_0=c}} = 0$$

また，束縛条件 $\sum_{k'=1}^{r_i} \theta_{X_i=k'|X_0=c} - 1 = 0$ を含んだラグランジュの未定乗数法を用いると，ラグランジェ関数 $F_{X_i|X_0}$ は

$$F_{X_i|X_0} = L + \lambda \left(\sum_{k'=1}^{r_i} \theta_{X_i=k'|X_0=c} - 1 \right)$$

である．次式を満たす $\theta_{X_i=k|X_0=c}$ が求める最尤推定量 $\hat{\theta}_{X_i=k|X_0=c}$ である．

$$\frac{\partial F_{X_i|X_0}}{\partial \theta_{X_i=k|X_0=c}} = 0$$

ヒント 2：確率の公理 $\sum_{c=1}^{r_0} \theta_{X_0=c} = 1$，$\sum_{k=1}^{r_i} \theta_{X_i=k|X_0=c} = 1$ を用いる．

2.1.1 ゼロ頻度問題

テストデータの中で，訓練データに含まれないデータを 1 つでも含んでいると，識別関数の計算に用いるパラメータ $\theta_{X_0=c}$ や $\theta_{X_i=k|X_0=c}$ の最尤推定値が 0 となってしまうことがある．この場合，識別関数の値も 0 となり（対数のときは $\log 0$ となり計算できない），そのカテゴリの確率は 0 になってしまう．

たとえばスパムメール分類を行うとき，メール文に “無料”，“儲かる” などの単語が含まれていると，スパムメールである確率が高まっていく．しかし，訓練時には含まれなかった新単語 “稼ぐ”が出現すると，そのパラメータは 0 と推定されるため，このメールがスパムである確率は 0 になってしまう．この問題をゼロ頻度問題と呼ぶ．ゼロ頻度問題を緩和する方法として，最尤推定量の計算に用いるデータの頻度が 0 となる時は最尤推定値を 1 とする方

法がある．

課題 12

http://www.ai.lab.uec.ac.jp/実験/にある MICS2019_NB_TAN プロジェクトをインポートし，”NB.java”内の関数”getParameters”，”classification”，”setFrequencyTable”を実装し，Naive Bayes による分類プログラムを完成させよ．eclipse プロジェクトのインポート方法は http://www.ai.lab.uec.ac.jp/wp-content/uploads/2018/11/cabebc8b347bc1e77dcdbf08de59ff4c.pdf を参照してください．
ただし，NB.java がわかりにくい人は Naive Bayes の分類プログラムを一から作成してもよい．また，プロジェクト内に含まれるデータセット”spam”，”sentiment”に対して Naive Bayes による分類精度を求めよ．データセット spam と sentiment は，ある単語がメールに含まれているか否かを説明変数とし，そのメールがスパムメールか否かを目的変数としたデータセットである．ゼロ頻度問題を解消するため，最尤推定量の計算に用いるデータの頻度が 0 となる時は，そのパラメータの最尤推定値を 1 とせよ．

※本実験で公開しているデータセットでは，最も右側の列を目的変数の列としており，NB.java でもデータの最も右側の列を目的変数として扱っている．つまり，テキストでは X_0 を目的変数としているが，NB.java 内では X_n を目的変数として扱っているため注意．

課題 12 の目標は spam と sentiment のテストデータ TD.csv の各行 (各メール) がスパムか (1 をとるか) 否か (0 をとるか) を (2.3) 式で分類し，全行 (全メール) に対する正答率を測定することである．そのためには，パラメータ $\theta_{X_0=c}$ と $\theta_{X_i=k|X_0=c}$ を spam の学習データ LD.csv から推定しなければならない．ここでは，パラメータを最尤推定量の式 (2.5) により推定する．式 (2.5) を計算するには N，$N_{X_0=k}$，$N_{X_i=k,X_0=c}$ をそれぞれ求めなければならない．N は学習データのサイズ，すなわち LD.csv の行数である．以上をまとめると，課題 12 の NB.java では，関数”setFrequencyTable”で $N_{X_0=k}$，$N_{X_i=k,X_0=c}$ を求め，それを用いて関数”getParameters”でパラメータ $\theta_{X_0=c}$ と $\theta_{X_i=k|X_0=c}$ の最尤推定値を計算し，それを用いて関数”classification”で TD.csv の正答率を測定する．正しく実装すると，spam の分類精度は 0.81 程度となる．

2.2 Tree Augmented Naive Bayes

前節で紹介した Naive Bayes は，各説明変数が目的変数を所与として条件付き独立であることを仮定している．しかし，一般にこの仮定は成り立たない．例えば図 2.2 に示されるネットワークのように，熱がある場合はインフルエ

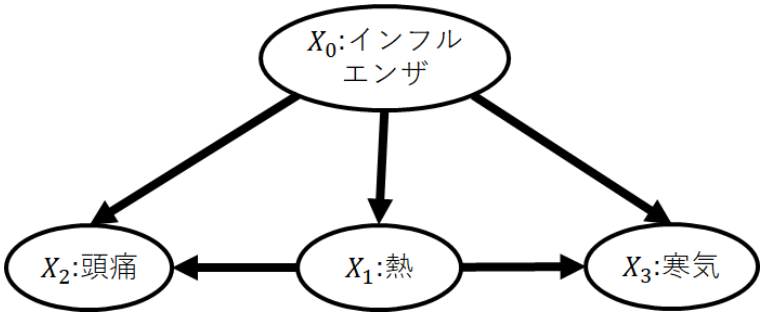

図 2.2: インフルエンザ判別に用いられる TAN の例

ンザの感染に関係なく悪寒がする確率が高まるため，明らかに二つの変数 X_1 と X_3 は X_0 を所与としても従属関係である．しかし，Naive Bayes では X_0 を所与として X_1 と X_3 は条件付き独立と仮定しているため，誤った確率を推定してしまう．この問題を解決するには，従属関係にある説明変数間に適切にエッジを引く必要がある．しかし，全ての説明変数同士の従属関係をチェックするには，膨大な計算時間がかかってしまう．そこで，説明変数間の従属関係を考慮し，かつ学習の計算量が少ないモデルとして，Tree Augmented Naive Bayes (TAN) が提案されている．TAN は，Naive Bayes のように目的変数が各説明変数の親となっており，説明変数間で木構造をとるモデルである（図 2.2）.

TAN の説明変数間の木構造はルート変数を必ず持つ．本テキストでは簡単のためルート変数を X_1 と仮定する．すなわち，X_1 は X_0 以外に親変数を持たない．TAN は同時確率分布を次のように表す.

$$p(X_0, X_1, \cdots, X_n) = p(X_0)p(X_1 \mid X_0) \prod_{i=2}^{n} p(X_i \mid X_{\pi(i)}, X_0)$$

ここで，$X_{\pi(i)}$ は TAN 構造における X_i の親となる説明変数である.

今，$X_{\pi(i)} = j$ かつ $X_0 = c$ のときに $X_i = k$ となる条件付き確率 $p(X_i = k \mid X_{\pi(i)} = j, X_0 = c)$ を示すパラメータを $\theta_{X_i=k|X_{\pi(i)}=j,X_0=c}$ と表すと，説明変数のインスタンス $x_1, \cdots, x_n$ に対する TAN の識別関数は次式で表される.

$$\widehat{x}_0 = \underset{c \in \{1,\cdots,r_0\}}{\operatorname{argmax}} \theta_{X_0=c}\theta_{X_1=x_1|X_0=c} \prod_{i=2}^{n} \theta_{X_i=x_i|X_{\pi(i)}=x_{\pi(i)},X_0=c} \tag{2.6}$$

また，D を所与とした TAN の対数尤度は以下で表される.

$$\begin{aligned} L = &\sum_{c=1}^{r_0} N_{X_0=c} \log \theta_{X_0=c} + \sum_{c=1}^{r_0}\sum_{k=1}^{r_1} N_{X_1=k,X_0=c} \log \theta_{X_1=k|X_0=c} \\ &+ \sum_{i=2}^{n}\sum_{c=1}^{r_0}\sum_{j=1}^{r_{\pi(i)}}\sum_{k=1}^{r_i} N_{X_i=k,X_{\pi(i)}=j,X_0=c} \log \theta_{X_i=k|X_{\pi(i)}=j,X_0=c} \end{aligned} \tag{2.7}$$

パラメータ $\theta_{X_0=c}$，$\theta_{X_1=k|X_0=c}$，$\theta_{X_i=k|X_{\pi(i)}=j,X_0=c}$ の最尤推定量は以下で表される．

- $\widehat{\theta}_{X_0=c} = \dfrac{N_{X_0=c}}{N}$
- $\widehat{\theta}_{X_1=k|X_0=c} = \dfrac{N_{X_1=k,X_0=c}}{N_{X_0=c}}$
- $\widehat{\theta}_{x_i=k|X_{\pi(i)}=j,X_0=c} = \dfrac{N_{X_i=k,X_{\pi(i)}=j,X_0=c}}{N_{X_{\pi(i)}=j,X_0=c}}$

ここで，$N_{X_i=k,X_{\pi(i)}=j,X_0=c}$ は $X_i = k$ かつ $X_{\pi(i)} = j$ かつ $X_0 = c$ となる頻度である．

TAN の説明変数が成す木構造は未知であるから，データから学習する必要がある．今，TAN の全パラメータ集合を Θ とし，TAN のとりうる全グラフ集合を $\mathcal{G}_{TAN}$ とする．TAN の構造学習では，次の尤度を最大にする構造 G^* を探索する．

$$G^* = \underset{G\in\mathcal{G}_{TAN}}{\operatorname{argmax}}\, p(D \mid G, \widehat{\Theta})$$

ここで，$\widehat{\Theta}$ は Θ の最尤推定量である．G^* を得るには，次の 5 つのステップを行えばよい．

1. 次の条件付き相互情報量 $I(X_i; X_j \mid X_0)$ を異なる二つの説明変数の組み $(X_i, X_j), i < j$ に対して計算する．

$$I(X_i; X_j \mid X_0) = \sum_{\substack{k\in\{1,\cdots,r_i\}\\ m\in\{1,\cdots,r_j\}\\ c\in\{1,\cdots,r_0\}}} p(X_i = k, X_j = m, X_0 = c) \log \frac{p(X_i = k, X_j = m \mid X_0 = c)}{p(X_i = k \mid X_0 = c)p(X_j = m \mid X_0 = c)} \tag{2.8}$$

   条件付き相互情報量の計算に必要な確率 $p(X_i = k, X_j = m, X_0 = c)$，$p(X_i = k, X_j = m \mid X_0 = c)$，$p(X_i = k \mid X_0 = c)$，$p(X_j = m \mid X_0 = c)$ はそれぞれ次のように推定する．

   - $\hat{p}(X_i = k, X_j = m, X_0 = c) = \dfrac{N_{X_i=k,X_j=m,X_0=c}}{N}$
   - $\hat{p}(X_i = k, X_j = m \mid X_0 = c) = \dfrac{N_{X_i=k,X_j=m,X_0=c}}{N_{X_0=c}}$
   - $\hat{p}(X_i = k \mid X_0 = c) = \dfrac{N_{X_i=k,X_0=c}}{N_{X_0=c}}$
   - $\hat{p}(X_j = m \mid X_0 = c) = \dfrac{N_{X_j=m,X_0=c}}{N_{X_0=c}}$

2. 各説明変数をノードとした完全無向グラフを生成し，各無向エッジ $(X_i, X_j), 1 \leq i < j \leq n$ に重み $I(X_i; X_j \mid X_0)$ を割り当てる．

3. 生成した重み付き完全グラフから，最大全域木を生成する．

4. 木のルートノードを一つ選び，そのルートノードから外側にエッジの方向をつけていく．

5. 目的変数から，構築された木構造の各説明変数に向けてエッジを加える．

注意：$N_{X_i=k,X_j=m,X_0=c}=0$ の時は $p(X_i=k,X_j=m,X_0=c)=0$ と考える．したがって，$N_{X_i=k,X_j=m,X_0=c}=0$ となるパターン k,m,c に対して式 (2.8) の被総和部は 0 となる．よって，$N_{X_i=k,X_j=m,X_0=c}>0$ となる変数値のパターンのみを考えれば良いため，木構造の学習においてゼロ頻度問題は生じない．

課題 13

課題 12 でインポートしたプロジェクトに含まれている”TAN.java”内の関数”getParameters”，”classification”，”setFrequencyTable”，”getConditionalMutualInformation”，”getMaximumSpanningTree”を実装し，TAN による分類プログラムを完成させよ．ただし，TAN.java がわかりにくい人は，TAN の分類プログラムを一から作成しても良い．また，データセット”spam”,”sentiment”に対して TAN の分類精度を求め，Naive Bayes と比較・考察せよ．ただし，ゼロ頻度問題を解消するため，最尤推定量の計算に用いるデータの頻度が 0 となる時は，そのパラメータの最尤推定値を 1 とせよ．

※課題 12 の NB.java と同様に，TAN.java でも目的変数がデータの一番右側にあることを前提としていることに注意.

課題 13 では，TAN を用いて二つのデータセット spam と sentiment の分類精度 (正答率) を測定する．課題 12 で用いた Naive Bayes は学習データに依らずグラフが確定していたが，課題 13 では TAN のグラフを学習データから推定しなければならない．具体的には，テキストの図 2.2 に示される $X2 \leftarrow X1 \rightarrow X3$ のように，説明変数間で構成する木構造をデータから推定する．木構造の推定方法として，テキスト 19〜20 ページに記載されているように，各説明変数間の条件付き相互情報量を重みとした最大全域木を求める．最大全域木を求めるアルゴリズムとして，プリム法やクラスカル法がある．なお，TAN.java では LD.csv, TD.scv で一番左側の列の変数を木構造のルートノードと定めていることに注意する．条件付き相互情報量の計算は関数 getConditionalMutualInformation に実装し，それを用いた説明変数間の木構造推定は関数 getMaximumSpanningTree に実装する．TAN.java では木構造の格納先を int 型配列 str_tan としており，求めた木構造において X_i の親変数（目的変数以外）が X_j の時，str_tan[i] = j として木構造を表せる．TAN のグラフを推定したら，そのグラフにしたがって各変数のパラメータを

最尤法で推定し，さらにその後分類精度を測定する．木構造のルートノード以外の説明変数は目的変数以外にも親変数を持つため，それに対応するように関数 setFrequencyTable, getParameters, classification を実装する必要がある．正しく実装すると，データセット sentiment の分類精度は約 0.58 程度となる．

これまで紹介した Naive Bayes と TAN では，パラメータを最尤法で推定した．しかし，よく知られているように，ベイズ推定はより強力である．以下で紹介しよう．

第3章　ディリクレモデル

ベイズアプローチに従い，パラメータの事前分布を考えることにしよう．事前分布を考える場合，様々な考え方があるが，最も合理的であると考えられるのは，事前分布と事後分布の分布形が同一になるような事前分布，すなわち，自然共益事前分布の導入であろう．尤度は多項分布に従うので，その自然共益事前分布であるディリクレ分布が事前分布としてよく用いられる．ここで，G を Naive Bayes または TAN 構造，G_{TAN} を TAN 構造とし，$\Theta_{X_0} = \bigcup_{c=1}^{r_0}\{\theta_{X_0=c}\}$，$\Theta_{X_i|X_0=c} = \bigcup_{k=1}^{r_i}\{\theta_{X_i=k|X_0=c}\}$，$\Theta_{X_i|X_{\pi(i)}=m,X_0=c} = \bigcup_{k=1}^{r_i}\{\theta_{X_i=k|X_{\pi(i)}=m,X_0=c}\}$ とおく．ディリクレ分布 $p(\Theta_{X_0} \mid G)$，$p(\Theta_{X_i|X_0=c} \mid G)$，$p(\Theta_{X_i|X_{\pi(i)}=m,X_0=c} \mid G_{TAN})$ はそれぞれ次のように表せる．

$$p(\Theta_{X_0} \mid G) = \frac{\Gamma\left(\sum_{c=1}^{r_0}\alpha_{0:c}\right)}{\prod_{c=1}^{r_0}\Gamma(\alpha_{0:c})}\prod_{c=1}^{r_0}\theta_{X_0=c}^{\alpha_{0:c}-1},$$

$$p(\Theta_{X_i|X_0=c} \mid G) = \frac{\Gamma\left(\sum_{k=1}^{r_i}\alpha_{i:kc}\right)}{\prod_{k=1}^{r_i}\Gamma(\alpha_{i:kc})}\prod_{k=1}^{r_i}\theta_{X_i=k|X_0=c}^{\alpha_{i:kc}-1}$$

$$p(\Theta_{X_i|X_{\pi(i)}=m,X_0=c} \mid G_{TAN}) = \frac{\Gamma\left(\sum_{k=1}^{r_i}\alpha_{i:kmc}\right)}{\prod_{k=1}^{r_i}\Gamma(\alpha_{i:kmc})}\prod_{k=1}^{r_i}\theta_{X_i=k|X_{\pi(i)}=m,X_0=c}^{\alpha_{i:kmc}-1}$$

ここで，$\alpha_{0:c}$ は $N_{X_0=c}$ に，$\alpha_{i:kc}$ は $N_{X_i=k,X_0=c}$ に，$\alpha_{i:kmc}$ は $N_{X_i=k,X_j=m,X_0=c}$ 対応する事前の知識を表現する擬似サンプルとしてのハイパーパラメータを示す．

事後分布は，事前分布を尤度に掛け合わせることにより得ることができる．先に求められた尤度とディリクレ分布を掛け合わせるとそれぞれ以下のような事後分布を得ることができる．

$$p(D,\Theta_{X_0} \mid G) = \frac{\Gamma\left(\sum_{c=1}^{r_0}\alpha_{0:c}\right)}{\prod_{c=1}^{r_0}\Gamma(\alpha_{0:c})}\prod_{c=1}^{r_0}\theta_{X_0=c}^{N_{X_0=c}+\alpha_{0:c}-1} \tag{3.1}$$

$$p(D,\Theta_{X_i|X_0=c} \mid G) = \frac{\Gamma\left(\sum_{k=1}^{r_i}\alpha_{i:kc}\right)}{\prod_{k=1}^{r_i}\Gamma(\alpha_{i:kc})}\prod_{k=1}^{r_i}\theta_{X_i=k|X_0=c}^{N_{X_i=k,X_0=c}+\alpha_{i:kc}-1} \tag{3.2}$$

$$\begin{aligned} &p(D,\Theta_{X_i|X_{\pi(i)}=m,X_0=c} \mid G_{TAN}) \\ &\quad= \frac{\Gamma\left(\sum_{k=1}^{r_i}\alpha_{i:kmc}\right)}{\prod_{k=1}^{r_i}\Gamma(\alpha_{i:kmc})}\prod_{k=1}^{r_i}\theta_{X_i=k|X_{\pi(i)}=m,X_0=c}^{N_{X_i=k,X_{\pi(i)}=m,X_0=c}+\alpha_{i:kmc}-1} \end{aligned} \tag{3.3}$$

これらの事後分布を最大にする MAP(Maximum A Posteriori) 推定値は，以下のとおりである．

$$\widehat{\theta}_{X_0=c} = \frac{N_{X_0=c} + \alpha_{0:c} - 1}{\sum_{c=1}^{r_0}(N_{X_0=c} + \alpha_{0:c} - 1)}$$

$$\widehat{\theta}_{X_i=k|X_0=c} = \frac{N_{X_i=k,X_0=c} + \alpha_{i:kc} - 1}{\sum_{k=1}^{r_i}(N_{X_i=k,X_0=c} + \alpha_{i:kc} - 1)}$$

$$\widehat{\theta}_{X_i=k|X_{\pi(i)}=m,X_0=c} = \frac{N_{X_i=k,X_{\pi(i)}=m,X_0=c} + \alpha_{i:kmc} - 1}{\sum_{k=1}^{r_i}(N_{X_i=k,X_{\pi(i)}=m,X_0=c} + \alpha_{i:kmc} - 1)}$$

MAP 推定値では，全てのハイパーパラメータを $\alpha_{x_0} = 1$，$\alpha_{ijk} = 1$ として一様分布に設定すると最尤推定値 (Maximum Likelihood estimator) に一致する．

しかし，ベイズ統計学では，MAP 推定値よりも事後分布の期待値である EAP(Expected A Posteriori) 推定値のほうが頑健で予測効率がよいことが知られている．ディリクレ分布，式 (3.1)，(3.2)，(3.3) の期待値はそれぞれ

$$\widehat{\theta}_{X_0=c} = \frac{N_{X_0=c} + \alpha_{0:c}}{\sum_{c=1}^{r_0}(N_{X_0=c} + \alpha_{0:c})} \tag{3.4}$$

$$\widehat{\theta}_{X_i=k|X_0=c} = \frac{N_{X_i=k,X_0=c} + \alpha_{i:kc}}{\sum_{k=1}^{r_i}(N_{X_i=k,X_0=c} + \alpha_{i:kc})} \tag{3.5}$$

$$\widehat{\theta}_{X_i=k|X_{\pi(i)}=m,X_0=c} = \frac{N_{X_i=k,X_{\pi(i)}=m,X_0=c} + \alpha_{i:kmc}}{\sum_{k=1}^{r_i}(N_{X_i=k,X_{\pi(i)}=m,X_0=c} + \alpha_{i:kmc})} \tag{3.6}$$

となる．ベイズ学習では，EAP 推定値が最も一般的に用いられる．ベイズ推定では，式 (3.4)，(3.5)，(3.6) のように事前分布のハイパーパラメータによってゼロ頻度問題を解消できる．一般に，ハイパーパラメータはベイジアンネットワークの尤度等価 (Heckerman et al., 1995) を満たすように次式で計算する．

$$\alpha_{0:c} = \frac{\alpha}{r_0} \tag{3.7}$$

$$\alpha_{i:kc} = \frac{\alpha}{r_i r_0} \tag{3.8}$$

$$\alpha_{i:kmc} = \frac{\alpha}{r_i r_{\pi(i)} r_0} \tag{3.9}$$

ここで，α は全ハイパーパラメータに共通するハイパーパラメータであり，α が大きくなるほど EAP 推定量は一様分布に近く．

課題 14

NB と TAN のパラメータを EAP 推定量（式 (3.4)，(3.5)，(3.6)）で推定した時のデータセット”spam”,”sentiment”の分類精度を求めよ．この時，ハイパーパラメータは式 (3.7)，(3.8)，(3.9) で計算し，α を動かして分類精度がどのように変化するか観察・考察せよ．また，最尤推定と EAP 推定の違いを比較・考察せよ．

関連図書

[1] Akaike,H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6),716-723

[2] Box,G.E.P. and Tiao,G.C. (1992) Bayesian Inference in Statistical Analysis. New York, N.Y., John Wiley and Sons.

[3] Bouckaert,R. (1994) Probabilistic network construction using the minimum description length principle. *Technical Report ruu-cs-94-27*,Utrecht University.

[4] Buntine,W.L. (1991) Theory refinement on Bayesian networks. In B. D'Ambrosio, P. Smets and P. Bonissone (eds.),*Proc. 7th Conf. Uncertainty in Artificial Intelligence*,52-60.LA,California,Morgan Kaufmann.

[5] Castillo,E., Hadi,A.S., and Solares,C. (1997) Learning and updating of uncertainty in Dirichlet models. *Machine Learning*, **26**,43-63

[6] Chickering,D.M. (1996) Learning Bayesian networks is NP-complete. In D.Fisher and H.Lenz (eds.),*Proc. International Workshop on Artificial Intelligence and Statistics*,121-130

[7] Chickering,D.M. and Heckerman,D. (2000) A comparison of scientific and engineering criteria for Bayesian model selection. *Statistics and Computing*, **10**, 55-62

[8] Cooper,G.F. and Herskovits,E. (1992) A Bayesian Methods for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309-347

[9] Darwiche, A. (2009) Modeling and reasoning with Bayesian networks, Cambridge University Press

[10] Dawid, A.P. (1979) Conditional Independence in Statistical Theory, Journal of the Royal Statistical Society, Series B, 41, 1-33

[11] Dawid, A.P. (1980) Conditional Independence for Statistical Operations, Annals of Statistics, 8, 598-617

[12] Heckerman,D. , Geiger,2012/10/24, and Chickering,D. (1995) Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* **20**,197-243

[13] Rissanen,J. (1978) Modeling by shortest data description, *Automatica* **14**, 465-471

[14] Koivisto,M. and Sood,K. (2004) Exact Bayesian structure discovery in Bayesian networks, *Journal of Machine Learning Research*, **5**, 549-573.

[15] Malone,B.M., Yuan,C., Hansen,E.A. , Bridges,S. (2011) Improving the Scalability of Optimal Bayesian Network Learning with External-Memory Frontier Breadth-First Branch and Bound Search. In A. Preffer and F.G. Cozman (eds.) *Proc. the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 479-488

[16] Schwarz,G.E. (1978) Estimating the dimension of a model, *Annals of Statistics* **6**(2), 461-464.

[17] Silander,T. and Myllymaki,P. (2006) A simple approach for finding the globally optimal Bayesian network structure. In R. Dechter and T. Richardson(eds.), *Proc. 22nd Conf. Uncertainty in Artificial Intelligence*, 445-452

[18] Silander,T., Kontkanen,P., and Myllymaki,P. (2007) On sensitivity of the MAP Bayesian network structure to the equipment sample size parameter. In K.B. Laskey, S.M. Mahoney, J.Goldsmith (eds.), *Proc. 23d Conf. Uncertainty in Artificial Intelligence*, 360-367

[19] Steck,H. and Jaakkola,T.S. (2002). On the Dirichlet Prior and Bayesian Regularization. In S.Becker, S.Thrun, K.Obermayer(eds.),*Advances in Neural Information Processing Systems (NIPS)*,697-704

[20] Steck,H. (2008) Learning the Bayesian network structure: Dirichlet Prior versus Data. In D.A. McAllester and P.Myllymaki (eds.),*Proc. 24th Conf. Uncertainty in Artificial Intelligence*, 511-518

[21] Suzuki,J. (1993) A Construction of Bayesian networks from Databases on an MDL Principle. In Heckerman, E.H.Mamdani (eds.),*Proc. 9th Conf. Uncertainty in Artificial Intelligence*, 266-273

[22] Suzuki, J. (2006) On strong consistency of model selection in classification. *IEEE Transactions on Information Theory* **52**(11), 4726-4774

[23] Ueno,M. (2010) Learning networks determined by the ratio of prior and data. In P. Grunwald and P. Spirtes (eds.), *Proc. the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, 598-605

[24] Ueno,M. (2011) Robust learning Bayesian networks for prior belief. In A. Preffer and F.G. Cozman (eds.) *Proc. the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 698-707

[25] Jensen, F.V. and Nielsen, T.D. (2007) Bayesian networks and decision graphs, Springer

[26] Lauritzen, S.L. (1974) Sufficiency, Prediction and Extreme Models, Scandinavian Journal of Statistics, 1, 128-134