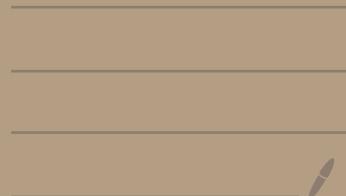


ベイス理論

Chapter 1 (後半)



1.5. グラフ, カルマモデル

グラフ, カルマモデル (graphical model)

- 確率モデル上に存在する複数の変数, 因果性を (ドアや矢印) を
使って表現する記法.

~ 基本的なモデル (因果など) や, エモーション確率モデルで
視覚的に表現できる.

モデル上, 変数間の独立性の判定などに, 手計算を使うよりも
グラフ上で考察の方が便利.

② ここでは, DAG (directed acyclic graph) による表現を説明.

[レポート等でない有向グラフ]

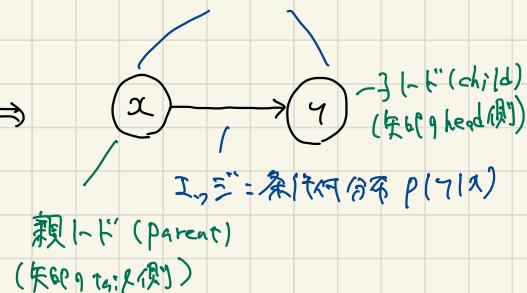
トド: 子変数

1.5.1. 有向グラフ.

赤玉白玉, 例の同時分布:

$$p(x, y) = p(y|x) p(x) \quad (1.55)$$

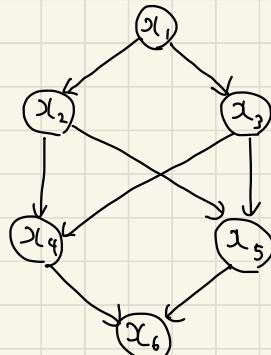
x と y の具体的な
関係性を表している



例: 同時分布

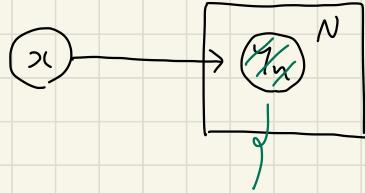
$$p(x_1, x_2, x_3, x_4, x_5, x_6)$$

$$= p(x_1) p(x_2|x_1) p(x_3|x_1) p(x_4|x_2, x_3) p(x_5|x_2, x_3) p(x_6|x_4, x_5) \quad (1.56)$$



(1.56) のグラフモデル

変数がN個存在する場合の表現を使うことによって、
例) (1, 5) が同じ分布で同時に分布。



この変数が条件付けていく

かを記述。

(好んでコードを書く)

…観測データを1つ上で明記。

推論アルゴリズムを導出する際に条件付分布を解釈する手段。

参考

マルチタスク学習 (multi-task learning)

…複数の関連するタスクを同時に解くことを目的、個々のタスクの予測精度を向上させようというアプローチ。

②複数の予測対象や観測データをうまく、モデルに統合して予測精度を上げるには、機械学習アルゴリズム構築の本質(?)。

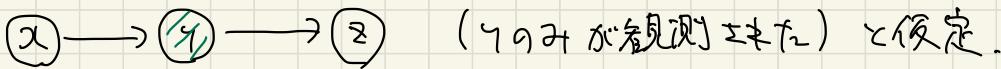
…複数のモデルを使い、関連性を複数のデータを統合(?)。

異なるタスクを同時に分布として扱う場合は、データの相似度を用いてそれを抽出する可能性がある。

1.5.2. ハードの条件付分布

ある確率モデル上で、条件付分布の計算と、グラフ上で、対応関係。
③)

$$p(x, y, z) = p(x) p(y|x) p(z|y) \quad (1.57)$$



→ 式 1.27 のハードの事後分布は、

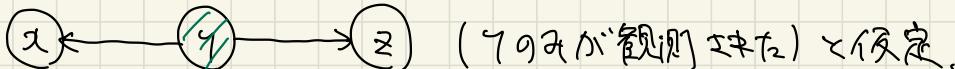
$$\begin{aligned} p(x, z|y) &= \frac{p(x, y, z)}{p(y)} \\ &= \frac{p(x) p(y|x) p(z|y)}{p(y)} \quad (\because (1.57)) \\ &= \underline{p(x|y)} \underline{p(z|y)} \quad (1.58) \end{aligned}$$

条件付独立性 (conditional independence)

モデルの式 (1.57) で y が観測された場合、
式 1.27 のハードの事後分布は独立分布に分解できる。

③) 2

$$p(x, y, z) = p(x|y) p(y|z) p(z) \quad (1.59)$$



→ 式 1.27 のハードの事後分布は、

$$\begin{aligned} p(x, z|y) &= \frac{p(x, y, z)}{p(y)} \\ &= \frac{p(x|y) p(y|z) p(z)}{p(y)} \\ &= \underline{p(x|y)} \underline{p(z|y)} \quad (1.60) \end{aligned}$$

条件付独立性が成り立つ、

13

$$p(x, y, z) = p(y|x, z) p(x) p(z) \quad (1.61)$$



(i) 一端が二つも観測されていない状態.

$$\begin{aligned} p(x, z) &= \sum_y p(x, y, z) \\ &= \sum_y p(y|x, z) p(x) p(z) \\ &= \underline{p(x) p(z)} \end{aligned} \quad (1.62)$$

独立かつ分布の分解.

(ii) y が観測された場合.

3x129 (一端が事後分布),

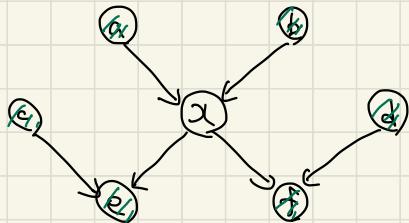
$$\begin{aligned} p(x, z|y) &= \frac{p(x, y, z)}{p(y)} \\ &= \underline{\frac{p(y|x, z)p(x)p(z)}{p(y)}} \end{aligned} \quad (1.63)$$

→ もともと独立 f_1, f_2 x, z が (一端 y が観測) または \rightarrow 依存関係をもつま f_2 .

④ 機械学習で使われるモデルには二つ以上の (一端向) の関係性を持つもので、 $x-y$ を観測した後 y の事後分布は直線の逆数が複数の組み合、 f_2 分布 (2 つ) が f_2 が f_2 で取る。

1.9.3. マルコフゲーティング

マルコフゲーティング (Markov blanket)



- も、と大きくは確率モデルから話を中心にした部分だけ取り扱う。

≈ グラフモデル上、他の変数の条件付独立性を考へる。

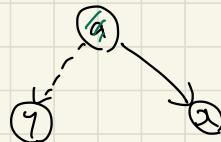
aとxの関係性

- 2つのヘッドが矢印で直結つながれてる場合、依存関係を持つ。
- aはxの父であり子ヘッドや親ヘッドが存在しても、aに対するtail-to-tail型かhead-to-tail型
- ≈ aが親(元)ではなくて子としての独立性



head-to-tail-tail

or

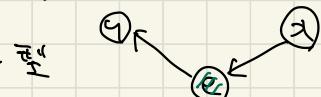


tail-to-tail-tail

eとxの関係性

- 2つのヘッドに直接箭がいるので、依存関係がある。

eは子ヘッドがいた場合、
xに対するhead-to-tail型



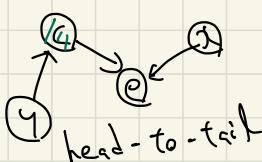
≈ xとは独立

eはcのような親ヘッドがいた場合。
≈ head-to-tail-head型

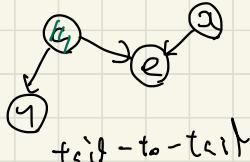


≈ xはcによって依存関係を持つ。

cは親ヘッドや子ヘッドがいた場合



head-to-tail



tail-to-tail

子は親は

≈ eとは独立
≈ xも独立

浅いが b, f, d (2 種) の 2 因子の議論が成立。

2 次上まである。

① ジーラー カルモデル上で入力と 2 变数が全て観測された場合、
入力候補を肉眼で見て、1 つには

- 直接親 (a, b)
- 間接親 (e, f)
- 共同親 (co-parent) (c, d)
(ef. 親)

② マルコフ フィールド、トポトピカル [他] 観測 [1 つが本, 1 つも。
また、2 つとも条件付分布の影響を受ける。

… サンプリングアンドザンギングに有用。

1.6. ベイズ学習, アポロ-7.

ベイズ学習 (Bayesian machine learning)

… 確率モデル = P と確率推論を利用して本機械学習, アポロ-7.

1.6.1. モデル構築と推論.

— ベイズ学習によるモデル構築と推論.

1. モデル構築

観測 D → P と観測されていない未知の変数 X について、
同時分布 $P(P, X)$ を構築する。

2. 推論, 対応.

$$\text{事後分布 } P(X|D) = \frac{P(D, X)}{P(D)} \quad \text{を解釈的または統計的につなぐ}.$$

ステップ1：モデル構築

・確率分布と組み合わせることで、観測データと未観測の変数の関係性を記述する。

① $p(\gamma | \theta)$

$$p(\gamma | \theta) = p(\gamma, \theta) / p(\theta) = \left(\prod_{n=1}^N p(\gamma_n | \theta) \right) p(\theta)$$

$\begin{cases} \gamma : \text{観測} \\ \theta : \text{未知, 変数} \end{cases}$

ステップ2：推論導出

・ステップ1で構築したモデルに基づいて未観測の変数の条件付き分布を求める。

$p(D) : \text{モデルエビデンス (model evidence)}$ /
周辺尤度 (marginal likelihood)

・モデルからデータ D が出現するまでのエビデンス。

$p(X | D)$

・離散分布には分子とモード分布によるものとある。

・ $\theta < 9$ 実用的な確率モデルにおける $p(X | D)$ は、形式が明確な確率分布を帰着できない。

= 対後分布 $p(X | D)$ を求めた中には必要な周辺尤度

$$p(D) = \int p(D, X) dX$$

が解析的に計算できない。

→ $\theta = \gamma_0 = \gamma'$ / 边り推論

$p(X | D)$ を簡易的な表現で用いて解釈。

1.6.2. 各タスクにおけるベイズ推論

ここでは、決定する機械学習タスクに適合する概略を述べる。

…具体的的な各確率分布の設定、仕方や事後分布、計算方法には
触れない。

着目

- 観測値と未観測値、関係
- 事後分布の推論がどう計算形式の(?)形で来るか。

④ 線形回帰、分類。

観測データ

$$X = \{x_1, \dots, x_n\} : \text{入力値}.$$

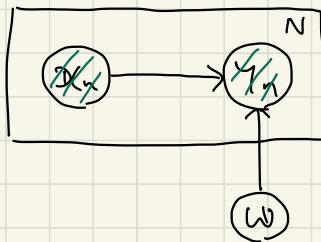
$$Y = \{y_1, \dots, y_n\} : \text{出力値}.$$

ω : 未知ベクトル (各点 x_n が y_n を予測する確率 ω)
… ω から推論。

このモデル、同時確率分布は、

$$p(Y, X, \omega) = p(\omega) \prod_{n=1}^N p(y_n | x_n, \omega) p(x_n) \quad (1.64)$$

$\underbrace{\qquad\qquad}_{\text{事前分布}}$



データ X および Y が観測された後、
 ω の事後分布

$$p(\omega | Y, X) \propto p(\omega) \prod_{n=1}^N p(y_n | x_n, \omega) \quad (1.65)$$

④ 一般の回帰や分類では、 ω の事後分布を計算するニーズが「学習」
に対するもの。

実際、 γ を計算

(1.65) の右边を計算。結果を正规化。

学習された w の分布を用いて新しい入力値 x_* に対する未知の出力値 γ_* (内側) 予測分布 (predictive distribution) を求めよとするが、どうぞ。

$$p(\gamma_* | \alpha_*, Y, X) = \int p(\gamma_* | \alpha_*, w) p(w | Y, X) dw \quad (1.66)$$

→ γ を観測した後、
 w が不確実。

~) α_* が γ_* の予測モデル $p(\gamma_* | \alpha_*, w)$ の重み付けを手配。

④ 因果的および分類モデルは、条件付きモデル (conditional model)
と呼ぶ。

… 入力 x_n が与えられた条件付き分布 $p(\gamma_n | \alpha_n, w)$ を直接モデル化
 \leftrightarrow joint model

④ γ の定義 = γ

Setting γ の長さが事前に分かっている。

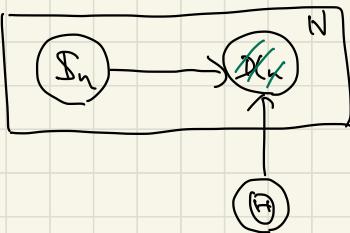
$X = \{\alpha_1, \dots, \alpha_N\}$: 観測データ

$S = \{S_1, \dots, S_N\}$: γ の各要素の当該。

$\Theta = \{\theta_1, \dots, \theta_K\}$: 各 γ の中心位置などを表すパラメータ。

~) γ を生成過程

$$p(X, S, \Theta) = p(\Theta) \prod_{n=1}^N p(\alpha_n | S_n, \Theta) p(S_n) \quad (1.67)$$



- 表示事前分布は従、 $\text{I}(\theta) + \text{I}(x)$ が決定。
- データ x の事後分布 $p(x|N)$ を算出する。グラスバウムの式で N が決まる。
- ここで、 D_n は既定である θ と $p(D_n|\theta)$ の分布を従う生成。
- D_n は θ と x を決定する。
... 隠れ変数 (hidden variable) / 潜在変数 (latent variable)

④ 二つめのモデル化
 ~) θ - x に対するグラスバウムの式で、中心変数に内蔵する確率的構造を表現を得る。
 (グラスバウムの内蔵を確実にするため推論によらず、 θ と x が独立)

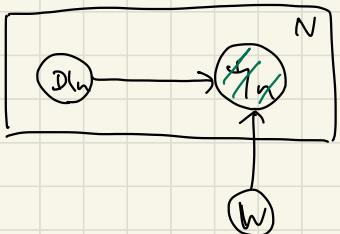
④ 総形の元前成

$$\begin{aligned} p(Y, X, W) &= p(Y|X, W) p(X) p(W) \\ &= p(W) \prod_{n=1}^N p(Y_n | x_n, W) p(x_n) \quad (1.68) \end{aligned}$$

W : 行列

本質的には (1.64) の総形固有モデルと同じ式。

⑤ 総形の元前成の場合、潜在変数 X が観測されることは前段



~) $\text{I}(\theta) + \text{I}(y)$ 事後分布 $p(w|y)$
 および事後分布 $p(x|y)$ を計算する。

特に θ - y の確率可視化を用いて、 x の事後分布 $p(x|y)$ の平均値を算出。

「生成モデル」や「隠形馬元前」成るトドーリ子「生成モデル」(generative model)
…すべて、観測データには必ず潜在変数が存在し、生成過程を記述する。 //

② 聯合モデル (joint model)

すべてのデータや未知の変数は同一の事後分布を直接観測できる。

まとめ ベイズ推論は、各機械学習の代表的な手法と並んで統一されたアプローチを示す。

1.6.3. 複雑な事後分布に対する近似入

④ 隠形变量モデルはおほく推論式 (1.65) を用いていた。
解釈的には事後分布が得られる。

分類モデルは「隠形モデル」と「ラグランジカルモデル」を含むが、
二つともに内訳 (1.4) のように非隠形変数が内に入っている。
解釈的には事後分布が得られる。

(3): 確率モデル $p(D, \mathbf{x})$ は必ず事後分布 $p(\mathbf{x}|D)$ を持つ。

• 比較的簡単なモデルでは、事後分布は单純な確率分布
(e.g. ガウス分布、ベーネーイ分布) に帰着。

… (1.64), 隠形变量
観測データ、 $\mathbb{P}(\mathbf{x}|D)$ 生成にガウス (高さ) 分布
 $\sim \mathcal{N}(\mathbf{x}|D)$ 事後分布、潜規データに対する確率分布も
ガウス分布。

④ 「ラグランジカル」と「ベーネーイ」の推論

固近代 $\left(\int p(D, \mathbf{x}) d\mathbf{x}, \sum_{\mathbf{x}} p(D, \mathbf{x}) \right)$ の計算が不可能

④ 解析的計算で主な未知確率分布 $p(\theta | D)$ を「知る」手段

1. $\pi = \gamma^{\ast} \pi^{\ast}$ (Sampling)

- 計算木幾何法, $\pi = \gamma$ 分布から $\pi = \gamma^{\ast}$ に $\sim p(\theta | D)$ で大量に得る.

→ 事後分布 $p(\theta | D)$ の平均値や分散を調査する.
分布の性質を調査する.

$\pi = \gamma^{\ast} \pi^{\ast} = \gamma$ 手法

MCMC (Markov Chain Monte Carlo)

ギブスサンプリング (Gibbs Sampling)

ハミルトンモンテカルロ (Hamiltonian Monte Carlo)

順次モンテカルロ (sequential Monte Carlo)

= 一本筋道, ギブス $\pi = \gamma^{\ast} \pi^{\ast} = \gamma$ を中心に普及,

2. 局部化された計算

e.g.: 事後分布 $p(\theta | D)$ の計算 (2 例) 困難な部分と容易な部分で局部化計算可能で簡単に実現.

・ 事後分布自体を直接 $p(\theta | D) \approx g(\theta)$ により直接分布表現.

手筋

ラプラス近似 (Laplace approximation)

変分推論 (Variational inference)

期待伝播 (expectation propagation)

= 一本筋道 变分推論を普及,

1.6.4. 不確実性に基づく意思決定

確率的理論

…対象となる現象が不確実性 (uncertainty) を定量的に表す。

手段

～推論。結果自体が何かしら意思決定 (decision making) を行なうわけではなく。

例) 明日、天気を予測するような推論アルゴリズムを構築

～ $p(y)$: 推論結果

予測

$$p(y = \text{晴}) = 0.8$$

(1.69)

$$p(y = \text{雨}) = 0.2$$

⇒ 外出する際は傘を持ちいくべき?

① (1.69) で表された推論結果は、明日の天気に関する不確実性を表したものだ。

～傘を持ちいくかどうかは意思決定する人の価値観や状況によって変わること。

⇒ 損失関数 (loss function) を考えることにより、より容易に定量化。

例): 雨でも濡れると嫌う A イン

$$L_A(y = \text{晴}, x = \text{傘なし}) = 0$$

$$L_A(y = \text{雨}, x = \text{傘なし}) = 100$$

$$L_A(y = \text{晴}, x = \text{傘あり}) = 10$$

$$L_A(y = \text{雨}, x = \text{傘あり}) = 15$$

～期待値

$$\sum_y L_A(y, x = \text{傘なし}) p(y) = 0 \times 0.8 + 100 \times 0.2 = 20$$

$$\sum_y L_A(y, x = \text{傘あり}) p(y) = 10 \times 0.8 + 15 \times 0.2 = 11$$

⇒ 雨が降る確率が低いが、傘を持ち出かけて方より期待損失が少ない。

例：荷物は傘を待つ歩くが嫌う B エン.

$$L_B (\gamma = \text{晴}, x = \text{傘なし}) = 0$$

$$L_B (\gamma = \text{雨}, x = \text{傘なし}) = 50$$

$$L_B (\gamma = \text{晴}, x = \text{傘あり}) = 20$$

$$L_B (\gamma = \text{雨}, x = \text{傘あり}) = 25$$

~) 期待値

$$\sum_{\gamma} L_B (\gamma, x = \text{傘なし}) p(\gamma) = 10$$

$$\sum_{\gamma} L_B (\gamma, x = \text{傘あり}) p(\gamma) = 21$$

⇒ 傘を持たない方が期待損失は少なくてよし.

① ベイズ学者は不確実的知識論とそれに伴う意思決定を明確に分けたことを基本とす.

了義化學習でも、不確実性の表現は重要.

ベイズ最適化

が「入力過程」での確実モデルを使うことにより、解析の難しさ、システムや実験の最適化パラメータを不確実性に基づいて探索.

1.6.5. ベイズ學習の利点と欠点

利点

1. さまざまな知識が一貫性を持て解決する.
2. 対象の不確実性を定量的に取り扱うことができる.
3. 利用可能な知識を自然に取り入れることができる.
4. 適切な適合度がある.

欠点

1. 理論的な知識を要する.
2. 計算コストがかかる.