ベイズ推論

Chapter 3 (3.1, 3.2)

3. ベイズ推論による学習と予測
・パラメータの事後分布および未観測値の予測分布の解析的な計算.

3.1. 学習と予測
① 確率推論を用いたパラメータの学習と未観測の値の予測.

学習 (training, learning)
モデルのもつパラメータの値をデータから決定すること.
… ベイズ推論では、パラメータを確率変数として扱う.
⇝ 確率計算によってデータを観測した後の事後分布を求めることが学習にあたる.

予測も行う.

3.1.1. パラメータの事後分布
$\mathcal{D}$: 訓練データ集合.

$$p(\mathcal{D}, \theta) = p(\mathcal{D} \mid \theta) p(\theta) \tag{3.1}$$

θ : モデルに含まれる未知パラメータ.
$p(\theta)$: パラメータに関する事前分布.
$p(\mathcal{D} \mid \theta)$: 尤度関数
… 特定のパラメータ θ からどのようにしてデータ $\mathcal{D}$ が発生するのかを記述.

② ベイズの定理を用いたパラメータの不確実性の更新.

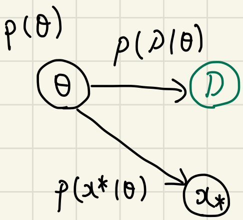
$$p(\theta \mid \mathcal{D}) = \frac{\boxed{p(\mathcal{D} \mid \theta)} p(\theta)}{p(\mathcal{D})} \tag{3.2}$$

⇝ 条件付分布 $p(\theta \mid \mathcal{D})$ を計算することが, ベイズ学習における「学習」にあたる.

尤度関数を通すことでより観測データ $\mathcal{D}$ に関する特徴を捉えていることが期待.

3.1.2 予測分布

$$p(x_*|D) = \underbrace{\int p(x_*|\theta)\,\underbrace{p(\theta|D)}_{\text{重み}}\,d\theta}_{\text{平均をとってるイメージ}} \quad (3.3)$$



- D も x_* も パラメータ θ によって発生過程が支配
- D と x_* の間の直接的な依存関係は仮定しない。

⇝ パラメータが与えられたもとで条件付独立。

対応する同時分布は

$$p(D, x_*, \theta) = p(D|\theta)\,p(x^*|\theta)\,p(\theta) \quad (3.4)$$

データ D だけが手元にあるとする。
このとき、残りの変数の事後分布は

$$p(x_*, \theta|D) = \frac{p(D, x_*, \theta)}{p(D)} \quad (\because \text{Bayes})$$

$$= \frac{\boxed{p(D|\theta)}\,p(x^*|\theta)\,\boxed{p(\theta)}}{\boxed{p(D)}} \quad (\because (3.4))$$

$$= p(x_*|\theta)\,p(\theta|D) \quad (\because (3.2)) \quad (3.5)$$

⇝ (3.5) から θ を積分消去

$$(\text{左辺}) = \int_\theta p(x_*, \theta|D)\,d\theta = p(x_*|D)$$

$$(\text{右辺}) = \int_\theta p(x_*|\theta)\,p(\theta|D)\,d\theta$$

より、(3.3) が得られる。

☆ 予測分布
事後分布から関心のない変数を積分消去した周辺分布。
⇒ 事前知識 $p(\theta)$ だけを頼り。

まとめ

(3.4)で表されるモデル

$$p(D, x_*, \theta) = p(D|\theta)\, p(x_*|\theta)\, p(\theta)$$

を使ってx_*を予測する。

1. (3.2)

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

を使ってパラメータの事後分布を学習。

2. データ自体は捨て、得られた事後分布から (3.3)

$$p(x_*|D) = \int p(x_*|\theta)\, p(\theta|D)\, d\theta$$

を用いて積分計算を行うことで、x_*に関する事後分布を求める。

◎ データDの情報をすべて事後分布 $p(\theta|D)$に押し込められる。

△ データ量に関してモデルの表現能力が変化しない。

(参考)

データDに合わせて予測モデルの表現能力を変えていく確率モデル

- ガウス過程
- ベイジアンノンパラメトリクス

3.1.3 共役事前分布

共役事前分布 (conjugate prior)

- (3.2)

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

における事前分布 $p(\theta)$ と事後分布 $p(\theta \mid D)$ が同じ種類の確率分布をもつように設定された事前分布。

- どのような事前分布が共役になりうるかは、尤度関数 $p(D \mid \theta)$ の設計の仕方にも依存。

- 参考文献 Chapter 3, p.p. 75, 表1 を参照。

- パラメータを2つ持つ分布 (e.g. ガウス分布) では、どのパラメータを学習させたいかによって共役事前分布が異なる。

⇝ 各確率分布と、それに対応する共役分布を使い、パラメータの事後分布や予測分布の解析的計算を確認していくことがこの本の中心テーマ。

② 共役事前分布を使うことのメリット。

- 事後分布や予測分布の計算が簡単かつ効率的に実行できる。
  … 逐次学習のフレームワークの構築。

  データセット D_1 を観測したあとの事後分布。

$$p(\theta \mid D_1) \propto p(D_1 \mid \theta)\, p(\theta) \quad (3.7)$$

  新規のデータセット D_2 を観測後、

$$p(\theta \mid D_1, D_2) \propto p(D_2 \mid \theta)\, p(\theta \mid D_1) \quad (3.8)$$

⇝ 共役事前分布を使えば、$p(\theta)$, $p(\theta \mid D_1)$, $p(\theta \mid D_1, D_2)$ はすべて同じ形式。

○ 解析的な事後分布の計算ができない複雑な拡張モデル
… 近似手法を用いた事後分布の計算
〜> 共役分布を組み合わせて全体のモデルを構築していけば
計算効率の高い近似アルゴリズムを導ける。

3.1.4. 共役でない事前分布の利用。
(発展)
データに関する興味深い構造を捉えるために、共役でない事前分布を採用することも。

(例) ガウス分布の平均パラメータ
事前分布：ガンマ分布 (共役でない)

方法：MCMC、変分推論

事後分布 $p(\theta|D)$ を変分パラメータ η を使った何らかの分布 $q(\theta;\eta)$ で近似的に表現できると仮定。
→ KLダイバージェンスの最小化。

$$\eta_{opt} = \underset{\eta}{\mathrm{argmin}}\, KL[q(\theta;\eta)\|p(\theta|D)] \tag{3.9}$$

勾配法

3.2. 離散確率分布の学習と予測

3.2.1. ベルヌーイ分布の学習と予測

$x \in \{0, 1\}$

$$p(x|\mu) = \mathrm{Bern}(x|\mu) \quad \left(= \mu^{x}(1-\mu)^{1-x} \right) \tag{3.10}$$

⇝ μ の分布を訓練データから推論.

③ ベイズ学習の考え方.
不確実性を持つ値はすべて確率変数として取扱い,
確率分布を事前に設定する必要がある.

⇒ $\mu \in (0, 1)$ を生成する確率分布(ベータ分布)を採用.

$$p(\mu) = \mathrm{Beta}(\mu|a, b) \quad \left(= \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{C_B(a,b)} \mu^{a-1}(1-\mu)^{b-1} \right) \tag{3.11}$$

a, b : ハイパーパラメータ
… given

i.i.d. assumption ?

データ $\mathbf{X} = \{x_1, \ldots, x_N\}$ を観測した後の事後分布を求める.

$$p(\mu|\mathbf{X}) \propto p(\mathbf{X}|\mu)p(\mu)$$

$$\propto \prod_{i=1}^{N} p(x_i|\mu)\, p(\mu) \tag{3.12}$$

(3.12) に対数をとる (指数分布12番目)

$$\ln p(\mu|\mathbf{x}) = \ln\left[\prod_{i=1}^{N} p(x_i|\mu)\, p(\mu)\right] + \text{const.}$$

$$= \sum_{i=1}^{N} \ln p(x_i|\mu) + \ln p(\mu) + \text{const.}$$

$$= \sum_{i=1}^{N} \ln\left[\mu^{x_i}(1-\mu)^{1-x_i}\right]$$

$$+ \ln\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\right] + \text{const.}$$

$$= \sum_{i=1}^{N} x_i \ln\mu + \sum_{i=1}^{N}(1-x_i)\ln(1-\mu)$$

$$+ (a-1)\ln\mu + (b-1)\ln(1-\mu) + \text{const.}$$

$$= \left(\sum_{i=1}^{N} x_i + a - 1\right)\ln\mu$$

$$+ \left(N - \sum_{i=1}^{N} x_i + b - 1\right)\ln(1-\mu) + \text{const.} \tag{3.13}$$

これと、(2.43)

$$\ln \mathrm{Beta}(\mu|\hat{a},\hat{b}) = (\hat{a}-1)\ln\mu + (\hat{b}-1)\ln(1-\mu) + \ln C_B(\hat{a},\hat{b})$$

を比較すると、事後分布は次のベータ分布で表されることが分かる。

$$p(\mu|\mathbf{x}) = \mathrm{Beta}(\mu|\hat{a},\hat{b}) \tag{3.14}$$

where

$$\hat{a} = \underbrace{\boxed{\sum_{i=1}^{N} x_i}}_{x=1\text{となる回数}} + a$$

$$\hat{b} = \underbrace{N - \sum_{i=1}^{N} x_i}_{x=0\text{となる回数}} + b \tag{3.15}$$

→) ベータ分布はいままでに何回コインの表と裏が出たか記憶する役割。

(参考) 経験ベイズ法

・ベイズ学習におけるパラメータの学習は、μに関する事後分布を求めるだけ。

⇝ 事後分布がベータ分布である ＋ パラメータの値を明らかにした

… 事前分布のパラメータ a、b を直接更新しているわけではない。

c.f. 経験ベイズ法

ハイパーパラメータを観測データに合わせて直接更新

… 確率推論によって導かれる手法ではない。

① ベイズ学習の枠組みでは、事前分布のハイパーパラメータは、"問題に関するドメイン知識を反映した上で、固定値として設定される"

⇝ もしハイパーパラメータもデータから学習させたい場合は、ハイパーパラメータに対してさらに事前分布を用意する。

○ どんな $\underbrace{\text{事前分布の設定}}_{\text{ハイパーパラメータの設定 ?}}$ をしたとしても、データ数 N が増えると対応する事前分布は次第に一致する。

○ 証拠となるデータが少ないときは、それぞれの予測者は自身の持つ信念 (事前分布) に大きく影響される。

⇒ データ数が増えると、それぞれの意見は次第に一致。

未観測の値 $x_* \in \{0,1\}$ に対する予測分布の計算.

簡単のため $p(\mu)$ を使った予測を行う.

$$p(x_*) = \int \overbrace{p(x_* \mid \mu)\, p(\mu)}^{p(x_*,\mu)}\, d\mu$$

$$= \int \mathrm{Bern}(x_* \mid \mu)\, \mathrm{Beta}(\mu \mid a, b)\, d\mu$$

$$= \int \mu^{x_*} (1-\mu)^{1-x_*}\, C_B(a,b)\, \mu^{a-1} (1-\mu)^{b-1}\, d\mu$$

$$= C_B(a,b) \int \mu^{x_*+a-1} (1-\mu)^{1-x_*+b-1}\, d\mu \quad (3.16)$$

ここで.

$$\mathrm{Beta}(\mu \mid \bar{a}, \bar{b}) = C_B(\bar{a}, \bar{b})\, \mu^{\bar{a}-1} (1-\mu)^{\bar{b}-1}$$

について. $\int \mathrm{Beta}(\mu \mid \bar{a}, \bar{b})\, d\mu = 1$

なので, (3.16) について.

$$\begin{cases} \bar{a} = x_* + a \\ \bar{b} = 1 - x_* + b \end{cases}$$

とおくと,

$$p(x_*) = \frac{C_B(a,b)}{C_B(\bar{a},\bar{b})} \int C_B(\bar{a},\bar{b})\, \mu^{\bar{a}-1} (1-\mu)^{\bar{b}-1}\, d\mu$$

$$= \frac{C_B(a,b)}{C_B(x_*+a,\, 1-x_*+b)}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(x_*+a)\Gamma(1-x_*+b)}{\Gamma(a+b+1)} \quad (3.18)$$

☆ $\Gamma(x+1) = x\Gamma(x)$

よって,

$$p(x_*=1) = \frac{\Gamma(a+b)}{\Gamma(a)\cancel{\Gamma(b)}} \cdot \frac{\Gamma(a+1)\cancel{\Gamma(b)}}{\Gamma(a+b+1)}$$

$$= \frac{\Gamma(a+b) \cdot a\Gamma(a)}{\Gamma(a)\,(a+b)\,\Gamma(a+b)} = \frac{a}{a+b} \quad (3.19)$$

$$p(x_* = 0) = \frac{\Gamma(a+b)}{\cancel{\Gamma(a)}\Gamma(b)} \cdot \frac{\cancel{\Gamma(a)}\Gamma(1+b)}{\Gamma(a+b+1)}$$

$$= \frac{\Gamma(a+b)\, b\, \Gamma(b)}{\Gamma(b)(a+b)\Gamma(a+b)} = \frac{b}{a+b} \qquad (3.20)$$

なので、

$$p(x_*) = \left(\frac{a}{a+b}\right)^{x_*} \left(\frac{b}{a+b}\right)^{1-x_*}$$

$$= \left(\frac{a}{a+b}\right)^{x_*} \left(1 - \frac{a}{a+b}\right)^{1-x_*}$$

$$= \mathrm{Bern}\left(x_* \,\middle|\, \frac{a}{a+b}\right) \qquad (3.21)$$

◎これと、$p(\mu|\mathbf{X})$ がベータ分布であることを用いる。

$$p(x_*|\mathbf{X}) = \mathrm{Bern}\left(x_* \,\middle|\, \frac{\hat{a}}{\hat{a}+\hat{b}}\right)$$

$$= \mathrm{Bern}\left(x_* \,\middle|\, \frac{\sum_{i=1}^{N} x_i + a}{N+a+b}\right) \qquad (3.22)$$

3.2.2. カテゴリ分布の学習と予測

$$p(\mathbf{s} \mid \boldsymbol{\pi}) = \mathrm{Cat}(\mathbf{s} \mid \boldsymbol{\pi}) \left(= \prod_{k=1}^{K} \pi_k^{s_k} \right) \quad (3.23)$$

where $\sum_{k=1}^{K} \pi_k = 1, \ \pi_k \in (0, 1)$

⇝ $\boldsymbol{\pi}$を生成する分布はディリクレ分布.

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \underbrace{C_D(\boldsymbol{\alpha})}_{= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)}} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \quad (3.24)$$

⇝ ディリクレ分布がカテゴリ分布に対する共役事前分布であることを確認する.

データ: $\mathbf{S} = \{ \mathbf{s}_1, \ldots, \mathbf{s}_N \}$ …離散値

$$p(\boldsymbol{\pi} \mid \mathbf{S}) \propto p(\mathbf{S} \mid \boldsymbol{\pi}) \, p(\boldsymbol{\pi}) \quad (\because \text{Bayes})$$

$$= \left(\prod_{i=1}^{N} \mathrm{Cat}(\mathbf{s}_i \mid \boldsymbol{\pi}) \right) \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \quad (3.25)$$

対数をとる.

$$\begin{aligned}
\ln p(\boldsymbol{\pi} \mid \mathbf{S}) &= \sum_{i=1}^{N} \ln \mathrm{Cat}(\mathbf{s}_i \mid \boldsymbol{\pi}) + \ln \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) + \text{const.} \\
&= \sum_{i=1}^{N} \ln \left(\prod_{k=1}^{K} \pi_k^{s_{i,k}} \right) + \ln \left(C_D(\boldsymbol{\alpha}) \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} s_{i,k} \ln \pi_k + \sum_{k=1}^{K} (\alpha_k - 1) \ln \pi_k + \text{const.} \\
&= \sum_{k=1}^{K} \ln \pi_k \left(\boxed{\sum_{i=1}^{N} s_{i,k}} + \alpha_k - 1 \right) + \text{const.} \quad (3.26)
\end{aligned}$$

事前分布のハイパーパラメータにこれを加えたもの

where $s_{i,k}$: $\mathbf{s}_i$のk番目の要素.

ここで、ディリクレ分布 $\mathrm{Dir}(\pi|\hat{\alpha})$ の p.d.f. の対数は

$$\ln\left(C_D(\hat{\alpha})\prod_{k=1}^{K}\pi_k^{\hat{\alpha}_k-1}\right) = \sum_{k=1}^{K}(\hat{\alpha}_k-1)\ln\pi_k + \mathrm{const}.$$

より、

$$p(\pi|\mathcal{S}) = \mathrm{Dir}(\pi|\hat{\alpha}) \qquad (3.27)$$

where $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_K)$,

$$\hat{\alpha}_k = \sum_{i=1}^{N} s_{i,k} + \alpha_k, \quad \text{for } k=1,\dots,K \qquad (3.28)$$

… 事後分布もディリクレ分布になる。

次に、予測分布の計算を行う

… 未観測の $\mathcal{S}_*$ を ハイパーパラメータ α を持つ事前分布 $p(\pi)$ に基づいて予測。

⇝ π を周辺化。

$$p(\mathcal{S}_*) = \int \overbrace{p(\mathcal{S}_*|\pi)\,p(\pi)}^{p(\mathcal{S}_*,\pi)}\,d\pi$$

$$= \int \mathrm{Cat}(\mathcal{S}_*|\pi)\,\mathrm{Dir}(\pi|\alpha)\,d\pi$$

$$= \int \prod_{k=1}^{K}\pi_k^{s_{*k}}\,C_D(\alpha)\prod_{k=1}^{K}\pi_k^{\alpha_k-1}\,d\pi$$

$$= C_D(\alpha)\int\prod_{k=1}^{K}\pi_k^{s_{*,k}+\alpha_k-1}\,d\pi$$

これと、ディリクレ分布 $\mathrm{Dir}(\pi|\bar{\alpha})$ の p.d.f. の積分

$$\int \mathrm{Dir}(\pi|\bar{\alpha})\,d\pi = C_D(\bar{\alpha})\int\prod_{k=1}^{K}\pi_k^{\bar{\alpha}_k-1}\,d\pi = 1$$

より、

$$p(\mathbf{s}_* \mid \pi) = \frac{C_D(\alpha)}{C_D(\bar{\alpha})} C_D(\bar{\alpha}) \int \prod_{k=1}^{K} \pi^{\bar{\alpha}_k - 1} d\pi$$

$$(\bar{\alpha}_k = s_{*,k} + \alpha_k)$$

$$= \frac{C_D(\alpha)}{C_D((s_{*,k} + \alpha_k)_{k=1}^{K})} \qquad (3.29)$$

where $(s_{*,k} + \alpha_k)_{k=1}^{K}$: K dim vector.

これと、

$$\begin{cases} C_D(\alpha) = \dfrac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \\ C_D((s_{*,k} + \alpha_k)_{k=1}^{K}) = \dfrac{\Gamma(\sum_{k=1}^{K} s_{*,k} + \alpha_k)}{\prod_{k=1}^{K} \Gamma(s_{*,k} + \alpha_k)} \end{cases}$$

より、

$$p(\mathbf{s}_* \mid \pi) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \cdot \frac{\prod_{k=1}^{K} \Gamma(s_{*,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} s_{*,k} + \alpha_k)} \qquad (3.30)$$

を得る。

ここで、あるk'に対して、$s_{*,k'} = 1$ となる場合を考えると、

$$p(\mathbf{s}_{*,k'} = 1) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \cdot \frac{\prod_{k \neq k'} \Gamma(\alpha_k) \Gamma(1 + \alpha_{k'})}{\Gamma(1 + \sum_{k=1}^{K} \alpha_k)}$$

$$= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(\alpha_{k'})} \cdot \frac{\alpha_{k'} \Gamma(\alpha_{k'})}{\sum_{k=1}^{K} \alpha_k \cdot \Gamma(\sum_{k=1}^{K} \alpha_k)}$$

$$= \frac{\alpha_{k'}}{\sum_{k=1}^{K} \alpha_k} \qquad (3.31)$$

よって、カテゴリ分布としてまとめなおすと、

$$p(\mathbf{s}_*) = \mathrm{Cat}\left(\mathbf{s}_* \,\middle|\, \left(\frac{\alpha_{k'}}{\sum_{k=1}^{K} \alpha_k}\right)_{k=1}^{K}\right) \qquad (3.32)$$

-- $p(\mathbf{s}_* \mid \mathbf{S})$ は $\alpha \leftarrow \hat{\alpha}$ とすれば得られる.

3.2.3. ポアソン分布の学習と予測

◎ ポアソン分布を使った事後分布と予測分布の計算.

$$p(x \mid \lambda) = \mathrm{Poi}(x \mid \lambda) \left(= \frac{\lambda^x}{x!} e^{-\lambda} \right) \qquad (3.33)$$

where $\lambda \in \mathbb{R}_+$.

⇝ λの事前分布として.

$$P(\lambda) = \mathrm{Gam}(\lambda \mid a, b) \left(= \underbrace{C_G(a,b)}_{= \frac{b^a}{\Gamma(a)}} \lambda^{a-1} e^{-b\lambda} \right) \qquad (3.34)$$

where $a, b \in \mathbb{R}_+$ are fixed. (Hyper Parameter)

Given: $\mathbb{X} = \{x_1, \ldots, x_N\} \sim$ Poisson distribution

$$\Rightarrow p(\lambda \mid \mathbb{X}) \propto p(\mathbb{X} \mid \lambda)\, p(\lambda)$$

$$= \left(\prod_{i=1}^{N} \mathrm{Poi}(x_i \mid \lambda) \right) \mathrm{Gam}(\lambda \mid a, b) \qquad (3.35)$$

対数をとると.

$$\ln p(\lambda \mid \mathbb{X}) = \sum_{i=1}^{N} \ln \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} + \ln C_G(a,b) \lambda^{a-1} e^{-b\lambda} + \text{const.}$$

$$= \sum_{i=1}^{N} \left(x_i \ln \lambda - \lambda - \ln x_i! \right)$$

$$+ (a-1) \ln \lambda - b\lambda + \ln C_G(a,b) + \text{const.}$$

$$= \left(\sum_{i=1}^{N} x_i + a - 1 \right) \ln \lambda + (N - b) \lambda + \text{Const.} \qquad (3.36)$$

よって、これはガンマ分布の対数と同じ形.

よって、

$$p(\lambda | X) = \mathrm{Gam}(\lambda | \hat{a}, \hat{b}) \quad (3.37)$$

where $\hat{a} = \sum_{i=1}^{N} x_i + a, \quad \hat{b} = N + b \quad (3.38)$

次に、予測分布を求める.

事前分布 : (3.34)

$$p(\lambda) = \mathrm{Gam}(\lambda | a, b)$$

⇝ 未観測データ x_* に関する予測分布は、

$$p(x_*) = \int p(x_* | \lambda) p(\lambda) \, d\lambda$$

$$= \int \mathrm{Poi}(x_* | \lambda) \mathrm{Gam}(\lambda | a, b) \, d\lambda \quad (3.39)$$

$$= \int \frac{\lambda^{x_*}}{x_*!} e^{-\lambda} C_G(a, b) \lambda^{a-1} e^{-b\lambda} \, d\lambda$$

$$= \frac{C_G(a,b)}{x_*!} \int \lambda^{x_* + a - 1} e^{-(1+b)\lambda} \quad (3.40)$$

ここで、ガンマ分布 $\mathrm{Gam}(\lambda | \bar{a}, \bar{b}) = C_G(\bar{a}, \bar{b}) \lambda^{\bar{a}-1} e^{-\bar{b}\lambda}$ に関して、

$$\int C_G(\bar{a}, \bar{b}) \lambda^{\bar{a}-1} e^{-\bar{b}\lambda} \, d\lambda = 1$$

より、

$$p(x_*) = \frac{C_G(a,b)}{x_*! \, C_G(\bar{a}, \bar{b})} \int C_G(\bar{a}, \bar{b}) \lambda^{\bar{a}-1} e^{-\bar{b}\lambda} \, d\lambda \quad (\bar{a} = x_* + a, \bar{b} = 1 + b)$$

$$= \frac{C_G(a, b)}{x_*! \, C_G(x_* + a, 1 + b)} \quad (3.42)$$

⇝ 負の二項分布。 $\quad \overset{\shortparallel}{\frac{b^a}{\Gamma(a)} \cdot \frac{1}{x_*!} \cdot \frac{\Gamma(x_* + a)}{(1+b)^{x_* + a}}}$

Let $r \in \mathbb{R}_+$, $p \in (0,1)$

$$p(x_*) = NB(x_* \mid r, p)$$
$$= \frac{\Gamma(x_* + r)}{x_*! \Gamma(r)} (1-p)^r p^{x_*} \quad (3.43)$$

を p.d.f. として持つ分布を負の二項分布という.
ただし,

$$\begin{cases} r = a \\ p = \dfrac{1}{b+1} \end{cases} \quad (3.44)$$

$$1 - \frac{1}{b+1} = \frac{b+1-1}{b+1} = \frac{b}{b+1}$$

$$\left(\frac{b}{b+1}\right)^a \cdot \left(\frac{1}{b+1}\right)^{x_*}$$
$$= b^a (1+b)^{-(a+x_*)}$$

であり,

$$\langle x_* \rangle = \frac{pr}{1-p} \quad (3.45)$$

$$\langle x_*^2 \rangle = \frac{pr(pr+1)}{(1-p)^2} \quad (3.46)$$

である.