Ch.3 Bayes推論による学習と予測.

§3.1 学習と予測.

* 学習

"独立同分布で"ということ.

i.i.d.に生成 → $\mathcal{D}$: データ

どんなデータが生成される？（予測分布）

生成 → x_* : 新規データ

モデル $p(x|\theta)$

パラメータ(r.v.): $\theta \sim \underbrace{p(\theta)}_{\text{事前分布}}$

頻度論と異なるところ.
（頻度論ではパラメータは定数）

パラメータの分布を更新

どうやって分布を更新する？（学習）

モデル $p(x|\theta)$

更新された分布.

パラメータ(r.v.): $\theta \sim \underbrace{p(\theta|\mathcal{D})}_{\text{事後分布}}$

- データを表現するモデル: $p(\mathcal{D},\theta) = \underbrace{p(\mathcal{D}|\theta)}_{\text{尤度}}p(\theta)$.

  - $p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(x_n|\theta)$. （データは i.i.d.）

- データ$\mathcal{D}$を観測した後の $p(\theta|\mathcal{D})$ の更新（学習）:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \quad (\because \text{Bayes' Th.})$$

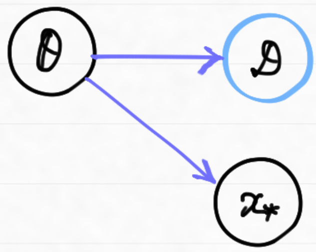
$\mathcal{D}$についての特徴をとらえているはず.

- 予測分布: データ$\mathcal{D}$を観測したうえでの新規データx_*の分布:

$$p(x_*|\mathcal{D}) = \mathbb{E}_{p(\theta|\mathcal{D})}\left[p(x_*|\theta)\right]$$

$$= \int p(x_*|\theta)p(\theta|\mathcal{D})d\theta.$$

graphical model.



$$p(\mathcal{D}, x_*, \theta) = p(\mathcal{D}|\theta)p(x_*|\theta)p(\theta).$$ (データが i.i.d.)

$\mathcal{D}$: given のとき.

$$p(x_*, \theta|\mathcal{D}) = \frac{p(\mathcal{D}, x_*, \theta)}{p(\mathcal{D})}$$

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

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

$$= p(x_*|\theta)p(\theta|\mathcal{D}).$$

$\downarrow$ θを積分除去

$p(x_*|\mathcal{D})$. 予測分布

- データ$\mathcal{D}$の情報をすべて事後分布 $p(\theta|\mathcal{D})$ に押し込められる.

→ 計算が便利.

- データ量に対して, モデルの表現能力は変化しない.

cf) Gauss過程 などの Baysian ノンパラメトリクス.

* 共役事前分布

事前分布 $p(\theta)$ と事後分布 $p(\theta|\mathcal{D})$ 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.

- 尤度 $p(\mathcal{D}|\theta)$ のかたちに依存.
- 学習させるパラメータにも依存.

✓ 事後分布 $p(\theta|\mathcal{D})$ や予測分布 $p(x_*|\mathcal{D})$ の計算が楽!

- 共役でない事前分布をつかう.

→ 事後分布 $p(\theta|\mathcal{D})$ は事前分布 $p(\theta)$ と違う種類の分布になる.

→ MCMC, 変分推論の利用.

◦ 変分推論

$p(\theta|\mathcal{D}) \approx q(\theta;\underline{\eta})$ と仮定.

└ 変分パラメータ

η の最適化:

$$\eta_{opt} := \underset{\eta}{\operatorname{argmin}} \, KL[q(\theta;\eta) \| p(\theta|\mathcal{D})].$$

ふつう, これは解析的に解けない. → 勾配法.

§3.2 離散確率分布の学習と予測.

* Bernoulli分布.

$p(x|\mu) = \mathrm{Bern}(x|\mu)$.　　μの分布を推論

$\mu \in (0,1)$を生成する分布? → ベータ分布.

$p(\mu) = \mathrm{Beta}(\mu|a,b)$,　a,b : ハイパーパラメータ

・ Bernoulli分布からのN個のデータ $\mathcal{X} = \{x_1, \cdots, x_N\}$: given

$$p(\mu|\mathcal{X}) = \frac{p(\mathcal{X}|\mu)p(\mu)}{p(\mathcal{X})} \quad (\because \text{Bayes' Th.})$$

$$\propto p(\mathcal{X}|\mu)p(\mu) \quad (p(\mathcal{X}) = \text{const.})$$

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

μの分布に興味があるから、
μが関わるところだけ見れば十分!

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

$$= \sum_{i=1}^{N} x_i \log\mu + \sum_{i=1}^{N}(1-x_i)\log(1-\mu) + (a-1)\log\mu + (b-1)\log(1-\mu) + \text{const.}$$

$$= \Big(\underbrace{\sum_{i=1}^{N} x_i + a - 1}_{=:\hat{a}}\Big)\log\mu + \Big(\underbrace{N - \sum_{i=1}^{N} x_i + b - 1}_{=:\hat{b}}\Big)\log(1-\mu) + \text{const.}$$

$\log \mathrm{Beta}(\mu|\hat{a},\hat{b})$ の形をしている

$x_i = 1$となる回数

$x_i = 0$となる回数

$$\therefore p(\mu|\mathcal{X}) = \mathrm{Beta}(\mu|\hat{a},\hat{b}).$$

データを観測した後の分布も事前分布と同じベータ分布なので
共役事前分布であることが分かる.

・未観測値 $x_* \in \{0,1\}$ の予測分布.

$$
\begin{aligned}
p(x_*) &= \int p(x_* \mid \mu)\, p(\mu)\, d\mu \\
&= \int \mathrm{Bern}(x_* \mid \mu)\, \mathrm{Beta}(\mu \mid a, b)\, d\mu \\
&= C_B(a,b) \int \mu^{x_*} (1-\mu)^{1-x_*} \mu^{a-1} (1-\mu)^{b-1}\, d\mu \\
&= C_B(a,b) \frac{1}{C_B(x_*+a,\, 1-x_*-b)} \\
&= \frac{\Gamma(a+b)\Gamma(x_*+a)\Gamma(1-x_*+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+1)} \\
&= \frac{1}{a+b} \frac{\Gamma(x_*+a)\Gamma(1-x_*+b)}{\Gamma(a)\Gamma(b)}.
\end{aligned}
$$

$x_* \in \{0,1\}$ なので場合分け.

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

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

$$
\therefore\ p(x_*) = \left(\frac{a}{a+b}\right)^{x_*} \left(\frac{b}{a+b}\right)^{1-x_*} = \mathrm{Bern}\left(x_* \,\middle|\, \frac{a}{a+b}\right).
$$

→ データがあるときも同様に

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

＊ カテゴリ分布.（K値）

$$p(s|\pi) = \mathrm{Cat}(s|\pi), \qquad \underline{\sum_{k=1}^{K} \pi_k = 1, \ \pi_k \in (0,1)}$$

―(＊)

(＊)をみたす π を生成する分布? → Dirichlet分布.

$$p(\pi) = \mathrm{Dir}(\pi|\alpha) = C_D(\alpha) \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}.$$

・ カテゴリ分布からのN個のデータ $\mathcal{S} = \{s_1, \cdots, s_N\}$: given.

$$p(\pi|\mathcal{S}) \propto p(\mathcal{S}|\pi)p(\pi).$$

$$= \prod_{i=1}^{N} \mathrm{Cat}(s_i|\pi)\mathrm{Dir}(\pi|\alpha).$$

$$\log p(\pi|\mathcal{S}) = \sum_{i=1}^{N} \log \mathrm{Cat}(s_i|\pi) + \log \mathrm{Dir}(\pi|\alpha) + \mathrm{const}.$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} s_{n,k} \log \pi_k + \sum_{k=1}^{K} (\alpha_k - 1)\log \pi_k + \underbrace{\log C_D(\alpha)}_{\mathrm{const.}} + \mathrm{const}.$$

$$= \sum_{k=1}^{K} \Big(\underbrace{\sum_{i=1}^{N} s_{n,k} + \alpha_k}_{=: \hat{\alpha}_k} - 1 \Big) \log \pi_k + \mathrm{const}.$$

事後分布もDirichlet分布になる.

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

・ 未観測データ s_* の予測分布.

$$p(s_*) = \int p(s_*|\pi)p(\pi)d\pi.$$

$$= \int \mathrm{Cat}(s_*|\pi)\mathrm{Dir}(\pi|\alpha)d\pi$$

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

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

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

$$= \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\prod_{k=1}^{K}\Gamma(s_{*,k}+\alpha_k)}{\left(\prod_{k=1}^{K}\Gamma(\alpha_k)\right)\Gamma\left(\sum_{k=1}^{K}(s_{*,k}+\alpha_k)\right)}$$

$$p(s_{*,k'}=1) = \frac{\cancel{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}\ \alpha_{k'}\ \cancel{\prod_{k=1}^{K}\Gamma(\alpha_k)}}{\cancel{\left(\prod_{k=1}^{K}\Gamma(\alpha_k)\right)}\left(\sum_{k=1}^{K}\alpha_k\right)\cancel{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}} = \frac{\alpha_{k'}}{\sum_{k=1}^{K}\alpha_k}$$

$s_{*,k'}=1$ のとき

$k \neq k'$ では

$s_{*,k}=0$ に注意.

ガンマ関数の性質 $\Gamma(x+1)=x\Gamma(x)$ をつかう.

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

データ $\mathcal{S}$ があるときも同様に,

$$p(s_*|\mathcal{S}) = \mathrm{Cat}\left(s_* \,\middle|\, \left(\frac{\hat{\alpha}_k}{\sum_{k'=1}^{K}\hat{\alpha}_{k'}}\right)_{k=1}^{K}\right)$$

* Poisson分布.

$$p(x|\lambda) = \mathrm{Poi}(x|\lambda), \quad \lambda \in \mathbb{R}_{>0}.$$

$\lambda \in \mathbb{R}_{>0}$ を生成する dist.? → ガンマ分布.

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

・Poisson分布からの N個のデータ $\mathcal{X} = \{x_1, \cdots, x_N\}$.

$$p(\lambda|\mathcal{X}) \propto p(\mathcal{X}|\lambda)p(\lambda)$$

$$= \prod_{i=1}^{N} \mathrm{Poi}(x_i|\lambda)\,\mathrm{Gam}(\lambda|a,b)$$

$$\log p(\lambda|\mathcal{X}) = \sum_{i=1}^{N} \log \mathrm{Poi}(x_i|\lambda) + \log \mathrm{Gam}(\lambda|a,b) + \mathrm{const.}$$

$$= \sum_{i=1}^{N}\Big(x_i \log\lambda - \underbrace{\log x_i!}_{\mathrm{const.}} - \lambda\Big) + (a-1)\log\lambda - b\lambda + \underbrace{\log C_G(a,b)}_{\mathrm{const.}} + \mathrm{const.}$$

$$= \Big(\underbrace{\sum_{i=1}^{N} x_i + a - 1}_{=:\hat{a}}\Big)\log\lambda - \underbrace{(N+b)}_{=:\hat{b}}\lambda + \mathrm{const.}$$

$$\therefore p(\lambda|\mathcal{X}) = \mathrm{Gam}(\lambda|\hat{a},\hat{b}).$$

・x_* の予測分布

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

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

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

$$= \frac{1}{x_*!}\frac{b^a}{\Gamma(a)}\frac{\Gamma(x_*+a)}{(1+b)^{x_*+a}}$$

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

$$= \frac{\Gamma(x_*+a)}{x_*!\,\Gamma(\underbrace{a}_{=:r})}\left(1-\frac{1}{1+b}\right)^a\left(\underbrace{\frac{1}{1+b}}_{=:p}\right)^{x_*}$$

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

$$=: \mathrm{NB}(x_* \mid r, p) :$$ 負の2項分布

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

Bernoulli試行を行うとき, r回の成功がおこるまでに失敗した試行回数の分布
(を $r \in \mathbb{Z}_{>0}$ から $r \in \mathbb{R}$ に拡張したもの)

ちなみに r=1 のときは 幾何分布.

・負の2項分布の平均.

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \frac{\Gamma(x+r)}{x!\,\Gamma(r)}(1-p)^r p^x$$

$$= \sum_{x=1}^{\infty} x \frac{\Gamma(x+r)}{x!\,\Gamma(r)}(1-p)^r p^x$$

$$= \sum_{x=1}^{\infty} \frac{\Gamma(x+r)}{(x-1)!\,\Gamma(r)}(1-p)^r p^x$$

$$= p\sum_{x=1}^{\infty} \frac{\Gamma(x+r)}{(x-1)!\,\Gamma(r)}(1-p)^r p^{x-1}$$

x-1 を改めて x とおく.

$$= p\sum_{x=0}^{\infty} \frac{\Gamma(x+1+r)}{x!\,\Gamma(r)}(1-p)^r p^x$$

$$= p\sum_{x=0}^{\infty} (x+r)\frac{\Gamma(x+r)}{x!\,\Gamma(r)}(1-p)^r p^x$$

$$= p\underbrace{\sum_{x=0}^{\infty} x \frac{\Gamma(x+r)}{x!\,\Gamma(r)}(1-p)^r p^x}_{=\mathbb{E}[X]} + pr\underbrace{\sum_{x=0}^{\infty} \frac{\Gamma(x+r)}{x!\,\Gamma(r)}(1-p)^r p^x}_{=1}$$

$$= p\mathbb{E}[X] + pr$$

$$\therefore \mathbb{E}[X] = \frac{pr}{1-p}.$$

・負の2項分布の分散.

$$\mathbb{E}[X^2] = \sum_{x=0}^{\infty} x^2 \frac{\Gamma(x+r)}{x!\,\Gamma(r)}(1-p)^r p^x$$

$$= \sum_{x=1}^{\infty} x \frac{\Gamma(x+r)}{(x-1)!\,\Gamma(r)}(1-p)^r p^x$$

$$= p\sum_{x=1}^{\infty} x \frac{\Gamma(x+r)}{(x-1)!\,\Gamma(r)}(1-p)^r p^{x-1}$$

$x-1$を改めてxとおく.

$$= p\sum_{x=0}^{\infty} (x+1) \frac{\Gamma(x+r+1)}{x!\,\Gamma(r)}(1-p)^r p^x$$

ガンマ関数の性質

$$= p\sum_{x=0}^{\infty} (x+1)(x+r) \frac{\Gamma(x+r)}{x!\,\Gamma(r)}(1-p)^r p^x$$

$$= p\sum_{x=0}^{\infty} (x^2+(r+1)x+r) \frac{\Gamma(x+r)}{x!\,\Gamma(r)}(1-p)^r p^x$$

$$= p\mathbb{E}[X^2] + p(r+1)\mathbb{E}[X] + pr$$

$$= p\mathbb{E}[X^2] + p(r+1)\frac{pr}{1-p} + pr.$$

$$\therefore \mathbb{E}[X^2] = p(r+1)\frac{pr}{(1-p)^2} + \frac{pr}{1-p} = \frac{pr(pr+1)}{(1-p)^2}.$$

$$\therefore \mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{pr}{(1-p)^2}$$