情報認識
「経験ベイズ法」

- 担当教員： 杉山　将（計算工学専攻）
- 居室： W8E-505
- 電子メール: sugi@cs.titech.ac.jp

これまでの復習

- **最大事後確率則：**

$$x \in \omega_k : k = \arg\max_i p(\omega_i \mid x)$$

- **ベイズの定理より**

$$p(\omega_i \mid x) \propto \underbrace{p(x \mid \omega_i)}_{\text{条件付き確率}} \underbrace{p(\omega_i)}_{\text{事前確率}}$$

- **事前確率はそのカテゴリに含まれる標本の割合で推定**

$$\hat{p}(\omega_i) = \frac{n_i}{n}$$

これまでの復習

■条件付き確率の推定1

- **最尤推定法**による確率密度関数推定

$$\hat{p}_j(x \mid \omega_i) = q_j(x; \hat{\theta}_{ML}^{(i,j)}) \qquad \hat{\theta}_{ML}^{(i,j)} = \arg\max_{\theta \in \Theta_j} L_{i,j}(\theta)$$

$$L_{i,j}(\theta) = \prod_{x \in \omega_i} q_j(x; \theta)$$

- **赤池の情報量規準**によるモデルの選択

$$\hat{p}(x \mid \omega_i) = \hat{p}_{a_i}(x \mid \omega_i)$$

$$a_i = \arg\max_j AIC_i(j)$$

$$AIC_i(j) = -\sum_{x \in \omega_i} \log q_j(x; \hat{\theta}_{ML}^{(i,j)}) + \dim \Theta_j$$

これまでの復習

■条件付き確率の推定2

- ベイズ推定法による確率密度関数推定

$$\hat{p}_j(x \mid \omega_i) = \int_{\Theta} q_j(x;\theta) \frac{\prod_{x \in \omega_i} q_j(x;\theta) p_j(\theta)}{\int_{\Theta} \prod_{x \in \omega_i} q_j(x;\theta') p_j(\theta') d\theta'} d\theta$$

■条件付き確率の推定3

- 最大事後確率推定法による確率密度関数推定

$$\hat{p}(x \mid \omega_i) = q(x; \hat{\theta}_{MAP}^{(i,j)})$$

$$\hat{\theta}_{MAP}^{(i,j)} = \arg\max_{\theta \in \Theta_j} \left[L_{i,j}(\theta) p_j(\theta) \right]$$

■モデルの選択，事前確率の設定は？

事前確率の設定

- 事前確率に関する知識がない場合，事前確率を自分で設定する必要がある．
- 事前確率によって推定結果が変わるため，客観的な方法で事前確率を設定しないと意味のある推定結果が得られない．
- あるパラメータ β で事前確率が制御できる場合を考える： $p(\theta;\beta)$
- 普通のパラメータ θ と区別するため，β を**ハイパーパラメータ**(hyper-parameter)とよぶ．
- 訓練標本を用いて β を客観的に設定したい．

経験ベイズ法

- **経験ベイズ法**(empirical Bayes method)：手元にある訓練標本が最も生起されやすいようにハイパーパラメータ β を設定.
- 訓練標本 $\{x_i\}_{i=1}^n$ が生起する確率：

$$p(x_1, x_2, \ldots, x_n; \beta) = \int_\Theta p(x_1, x_2, \ldots, x_n \mid \theta) p(\theta; \beta) d\theta$$

$$= \int_\Theta \prod_{i=1}^n q(x_i; \theta) p(\theta; \beta) d\theta$$

- **周辺尤度**(marginal likelihood)：これを β の関数と見たもの.

$$ML(\beta) = \int_\Theta \prod_{i=1}^n q(x_i; \theta) p(\theta; \beta) d\theta$$

経験ベイズ法 (続き)

- 周辺尤度を最大にするように β を決定.

$$\beta_{EB} = \arg\max_{\beta} ML(\beta)$$

$$ML(\beta) = \int_{\Theta} \prod_{i=1}^{n} q(x_i;\theta) p(\theta;\beta) d\theta$$

- この方法は**第 II 種最尤推定法**(type II maximum likelihood estimation method)ともよばれる.
- 周辺尤度は, **証拠**(evidence), **確率的複雑さ**(stochastic complexity), あるいは**自由エネルギー**(free energy)ともよばれる.

周辺尤度の計算

$$ML(\beta) = \int_{\Theta} \prod_{i=1}^{n} q(x_i;\theta) p(\theta;\beta) d\theta$$

- 周辺尤度は積分を含んでいるため，計算が大変．
- ラプラス近似：

$$ML(\beta) \approx \prod_{i=1}^{n} q(x_i;\hat{\theta}_{MAP}) p(\hat{\theta}_{MAP};\beta) \sqrt{\frac{(2\pi)^{\dim\theta}}{|-H|}}$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid x_1, x_2, \ldots, x_n)$$

$$H_{i,j} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \left[\sum_{k=1}^{n} \log q(x_k;\theta) + \log p(\theta;\beta) \right]_{\theta=\hat{\theta}_{MAP}}$$

周辺尤度の計算 (続き)

- 実際には，対数周辺尤度の方が計算しやすいことが多い．

$$\log ML(\theta) \approx \sum_{i=1}^{n} \log q(x_i;\hat{\theta}_{MAP}) + \log p(\hat{\theta}_{MAP};\beta) + \frac{\dim\theta}{2}\log 2\pi - \frac{1}{2}\log|-H|$$

- 訓練標本が十分に多いとき，この近似は精度がよい．

最適なハイパーパラメータの探索

$$\beta_{EB} = \arg\max_{\beta} \int_{\Theta} p(x_1, x_2, \ldots, x_n \mid \theta)\, p(\theta; \beta)\, d\theta$$

- 周辺尤度の積分はラプラス近似で簡単に計算できるが，それを最大にする β_{EB} を解析的に求めるのは困難．
- 以下のように探索する．
  1. β の候補をいくつか用意
  2. それらに対して周辺尤度を計算
  3. その中で周辺尤度が最大のものを選ぶ

周辺尤度を用いたモデル選択

- 最尤推定のときと同様に，MAP推定でもモデルを適切に選ぶ必要がある．
- ハイパーパラメータの設定と同様に，全てのモデルの候補に対して周辺尤度を計算し，周辺尤度最大のモデルを選べばよい．

ラプラス近似の更なる近似

- 訓練標本が十分多いことを仮定すると，

$$\sum_{i=1}^{n} \log q(x_i;\hat{\theta}_{MAP}) = O(n) \qquad \frac{\dim\theta}{2}\log 2\pi = O(1)$$

$$\log p(\hat{\theta}_{MAP};\beta) = O(1) \qquad \frac{1}{2}\log|H| = O(\log n)$$

$$\hat{\theta}_{MAP} \approx \hat{\theta}_{ML} = O(1) \qquad \frac{1}{2}\log|H| \approx \frac{\dim\theta}{2}\log n$$

- $O(1)$ の項を無視することにすれば

$$\log\int_{\Theta}\prod_{i=1}^{n} q(x_i;\theta)p(\theta;\beta)d\theta$$

$$\approx \sum_{i=1}^{n}\log q(x_i;\hat{\theta}_{ML}) - \frac{\dim\theta}{2}\log n$$

ベイズ情報量規準

- **ベイズ情報量規準**(Bayesian information criterion)：**これを－2倍したもの**

$$BIC = -2\sum_{i=1}^{n} \log q(x_i;\hat{\theta}_{ML}) + \dim\theta \log n$$

- **AICとBICは似た形**

$$AIC = -2\sum_{i=1}^{n} \log q(x_i;\hat{\theta}_{ML}) + 2\dim\theta$$

- **BICの方は複雑なモデルに対する罰則が強いため，単純なモデルを好む傾向．**
- **どちらがよいかは場合と立場による．**

小レポート

- **モデル：** $$q(x;\sigma^2)=\frac{1}{(2\pi\sigma^2)^{d/2}}\exp\left(-\frac{x^T x}{2\sigma^2}\right)$$
- **標本：** $\{x_i\}_{i=1}^n,\ x_i\in R^d$
- **事前確率：** $p(\sigma^2;\beta)=\beta\exp(-\beta\sigma^2)\qquad \beta>0$
- **以上の設定のもと，対数周辺尤度のラプラス近似が次式で与えられることを示せ**

$$\log ML(\theta)\approx-\frac{nd\log 2\pi}{2}-\frac{nd\log\hat{\sigma}_{MAP}^2}{2}-\frac{\sum_{i=1}^n x_i^T x_i}{2\hat{\sigma}_{MAP}^2}+\log\beta-\beta\hat{\sigma}_{MAP}^2$$
$$+\frac{1}{2}\log 2\pi-\frac{1}{2}\log\left(-\frac{nd}{2\hat{\sigma}_{MAP}^4}+\frac{\sum_{i=1}^n x_i^T x_i}{\hat{\sigma}_{MAP}^6}\right)$$

$$\hat{\sigma}_{MAP}^2=\frac{2}{nd+\sqrt{c}}\sum_{i=1}^n x_i^T x_i\qquad c=n^2d^2+8\beta\sum_{i=1}^n x_i^T x_i$$