

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

- **ベイズ推定法**: モデルをパラメータの事後確率に関して平均することによって推定する方法

$$\hat{p}(x) = \int_{\Theta} q(x | \theta) p(\theta | X) d\theta \quad X = \{x_1, x_2, \dots, x_n\}$$

- **最大事後確率推定法 (MAP推定法)**: パラメータに関して積分するのをやめて, 事後確率を最大にするパラメータ1点を用いる

$$\hat{p}(x) = q(x | \hat{\theta}_{MAP})$$

$$\begin{aligned} \hat{\theta}_{MAP} &= \arg \max_{\theta} p(\theta | X) \\ &= \arg \max_{\theta} \left[\sum_{i=1}^n \log q(x_i | \theta) + \log p(\theta) \right] \end{aligned}$$

- ベイズ推定法やMAP推定法において事前確率に関する知識がない場合、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 で事前分布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 を超パラメータ(hyper-parameter)とよぶ.
- 事前分布の設定の仕方:
 - データから事前分布を定める(経験ベイズ法)
 - 事前分布の事前分布(超事前分布)を考えて, 平均を取る

- 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 $X = \{x_i\}_{i=1}^n$ が生起する確率:

$$p(X; \beta) = \int_{\Theta} p(X | \theta) p(\theta; \beta) d\theta$$

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

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

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

- 周辺尤度を最大にするように β を決定.

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

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

- 対数をとったほうが計算が簡単なことが多い.

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

$$\log ML(\beta) = \int_{\Theta} \left[\log \sum_{i=1}^n q(x_i | \theta) + \log p(\theta; \beta) \right] d\theta$$

- この方法は第Ⅱ種最尤推定法(type II maximum likelihood estimation)ともよばれる.
- 周辺尤度は,
 - 証拠(evidence)
 - 確率的複雑さ(stochastic complexity)
 - 自由エネルギー(free energy)などともよばれる.
- 周辺尤度は, 様々な分野で現れる重要な量 !

最適なハイパーパラメータの探索⁵²

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

- 周辺尤度を最大にする 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

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

- 周辺尤度は積分を含んでいるため、計算が大変.
- そこで、積分を近似計算することにする.
- 以下では、簡単のため $f(\theta)$ の積分

$$\int_{\Theta} f(\theta) d\theta$$

を近似する問題を考える.

■ テイラー展開(Taylor expansion):

$$g(\theta) = g(\theta_0) + (\theta - \theta_0)g'(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2 g''(\theta_0) + \dots$$

(θ が1次元のとき)

■ テイラー展開を適当な次数で打ち切れば,
関数 $g(\theta)$ を多項式で近似できる.

■ θ が多次元のとき,

$$g(\theta) = g(\theta_0) + (\theta - \theta_0)^T b + \frac{1}{2}(\theta - \theta_0)^T B(\theta - \theta_0) + \dots$$

$$b_i = \left. \frac{\partial}{\partial \theta_i} g(\theta) \right|_{\theta=\theta_0}$$

$$B_{i,j} = \left. \frac{\partial^2}{\partial \theta_i \partial \theta_j} g(\theta) \right|_{\theta=\theta_0}$$

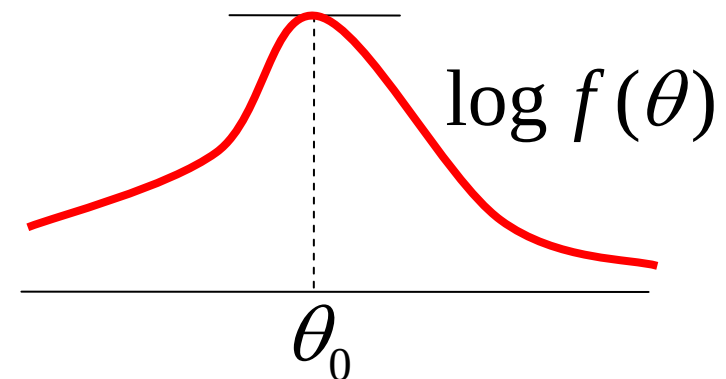
: ヘシアン行列
(Hessian matrix)

- $\theta_0 = \arg \max_{\theta \in \Theta} f(\theta)$ は最大値なので次式を満たす

$$\left. \frac{\partial}{\partial \theta_i} f(\theta) \right|_{\theta=\theta_0} = 0$$

- $\log f(\theta)$ に対しても同様に

$$\left. \frac{\partial}{\partial \theta_i} \log f(\theta) \right|_{\theta=\theta_0} = 0$$



- $\log f(\theta)$ を θ_0 の周りで2次のテーラー展開

$$\log f(\theta) \approx \log f(\theta_0) + 0 + \frac{1}{2}(\theta - \theta_0)^T H(\theta - \theta_0)$$

$$\log \tilde{f}(\theta)$$

$$H_{i,j} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\theta) \Big|_{\theta=\theta_0}$$

- 指数をとれば,

$$f(\theta) \approx \tilde{f}(\theta) = f(\theta_0) \exp\left(\frac{1}{2}(\theta - \theta_0)^T H(\theta - \theta_0)\right)$$

- 正規分布の確率密度関数の積分は1:

$$\int_D \phi(x) dx = 1$$

$$\phi(x) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

- これより,

$$\det(-H^{-1}) = \det(-H)^{-1}$$

$$\int_{\Theta} \exp\left(-\frac{1}{2}(\theta - \theta_0)^T (-H)(\theta - \theta_0)\right) d\theta = \sqrt{\frac{(2\pi)^{\dim \theta}}{\det(-H)}}$$

- 積分のラプラス近似(Laplace approximation):

$$\int_{\Theta} f(\theta) d\theta \approx \int_{\Theta} \tilde{f}(\theta) d\theta = f(\theta_0) \sqrt{\frac{(2\pi)^{\dim \theta}}{\det(-H)}}$$

- $\log f(\theta)$ を2次関数で近似することは,
 $f(\theta)$ を(正規化されていない)ガウス関数
で近似することに対応.
- そのため, ラプラス近似は**ガウス近似**
(Gaussian approximation) と呼ばれる.
- $f(\theta)$ の形状がガウス分布の形状に近いと
き, ラプラス近似は精度がよい.

対数周辺尤度のラプラス近似 259

$$\log ML(\beta) \approx \log PL(\hat{\theta}_{MAP}; \beta) + \frac{\dim \theta}{2} \log 2\pi - \frac{1}{2} \log(\det(-H))$$

$$\hat{\theta}_{MAP} = \arg \max_{\theta} [\log PL(\theta; \beta)]$$

$$H_{i,j} = \left. \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log PL(\theta; \beta) \right|_{\theta = \hat{\theta}_{MAP}}$$

$$\log PL(\theta; \beta) = \sum_{k=1}^n \log q(x_k; \theta) + \log p(\theta; \beta)$$

- 訓練標本数が大きい時, この近似は精度がよい.

周辺尤度を用いた ベイズ流のモデル選択

- 最尤推定するときと同様に, MAP推定でもモデルを適切に選ぶ必要がある.
- 超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

261

- 訓練標本が十分多いことを仮定すると,

$$\sum_{i=1}^n \log q(x_i | \hat{\theta}_{MAP}) = O(n)$$

$$\frac{\dim \theta}{2} \log 2\pi = O(1)$$

$$\log p(\hat{\theta}_{MAP}; \beta) = O(1)$$

$$\hat{\theta}_{MAP} \approx \hat{\theta}_{ML} = O(1)$$

$$\frac{1}{2} \log(\det(-H)) \approx \frac{\dim \theta}{2} \log n = O(\log n)$$

- $O(1)$ の項を無視することになれば

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

- ベイズ情報量規準(Bayesian information criterion):

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

- BICはもはや超パラメータ β に依らない.
- 従って, 超パラメータの決定にBICを用いることは出来ない.
- パラメトリックモデルの選択には使える.

- AICとBICは似た形:

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

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

- BICの方が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がある.
- どちらがよいかは場合と立場による.



まとめ

264

- ベイズの枠組みでは, 周辺尤度を最大にするように超パラメータやモデルを決定するのが自然.
- これは, 今手持ちのデータが最も生起されやすいように超パラメータやモデルを決定することに対応する.
- 周辺尤度は, ラプラス近似やBIC近似により効率よく計算できる.

1. $f(\theta) = p(\theta; 0, 1^2) + p(\theta; 0, 2^2)$

$$p(\theta; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(\theta - \mu)^2}{2\sigma^2}\right)$$

に対して,

$$Z = \int_{-\infty}^{\infty} f(\theta) d\theta$$

の値をラプラス近似で求めよ(真の値は2である).

■ モデル: $q(x | \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2}\right)$

■ 標本: $\{x_i\}_{i=1}^n, x_i \in R$

■ 事前確率: $p(\mu; \beta) = \frac{1}{\sqrt{2\pi\beta^2}} \exp\left(-\frac{\mu^2}{2\beta^2}\right)$ $\beta > 0$

2. 上記の設定(先週の演習と同じ)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.

$$\log ML(\beta) \approx -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^n (x_i - \hat{\mu}_{MAP})^2 - \frac{\hat{\mu}_{MAP}^2}{2\beta^2} - \frac{1}{2} \log(n\beta^2 + 1)$$

$$\hat{\mu}_{MAP} = \frac{1}{n + \beta^{-2}} \sum_{i=1}^n x_i$$

3. Octaveなどを用いた実験:

平均0.5分散1の正規分布から n 個標本を生成し, 2. のモデルに対して経験ベイズ法で β を決定せよ. 具体的には, β の候補を幾つか用意し, それぞれに対して対数周辺尤度をラプラス近似で求め, 対数周辺尤度の最も大きいものを選び.

そうして得られた β を用いてMAP推定法で μ を推定せよ.

上記の実験を真の平均や標本数を変えて実験を行い,

- 経験ベイズ法+MAP推定法
- 最尤推定法

の精度を比較せよ.

Octaveのサンプルプログラム 268

ex9.m

```
clear all
n=20; mu=0.5; sigma=1;
xx=sigma*randn(n,1)+mu;
betas=[0.01:0.01:3]; b=length(betas);

mu_MLE=mean(xx);
sigma_MLE=std(xx,1);
for i=1:b
    beta=betas(i);
    mu_MAP(i)=sum(xx)/(n+beta.^(-2));
    logML(i)=-n/2*log(2*pi) ...
              -sum((xx-mu_MAP(i)).^2)/2 ...
              -mu_MAP(i)^2/(2*beta^2) ...
              -log(n*beta^2+1);
end
[dummy,c]=max(logML);a
```

(右上に続く)

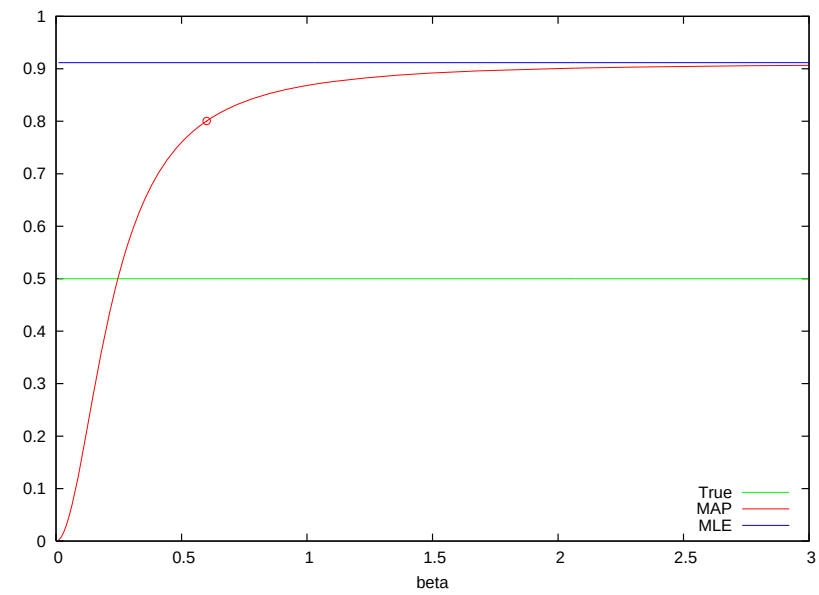
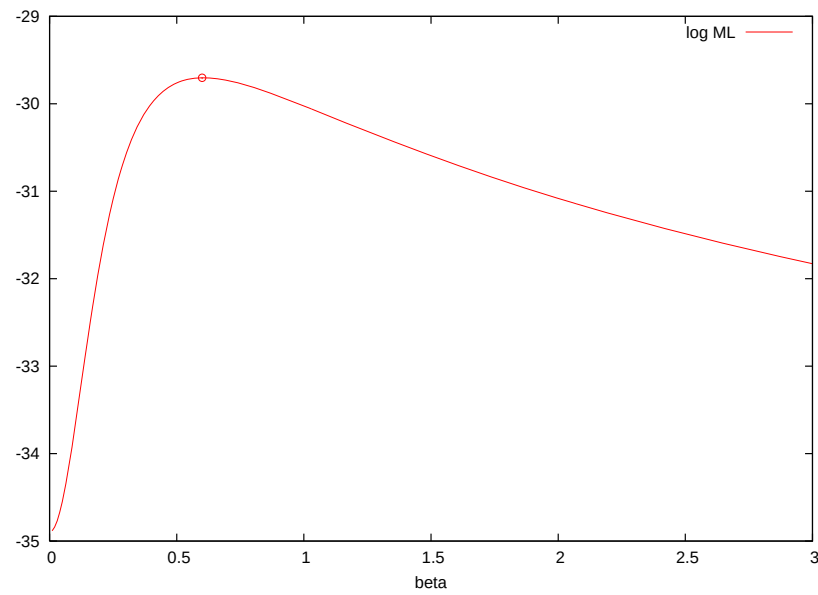
(左下から続き)

```
figure(1); clf; hold on;
plot(betas,logML,'r-')
plot(betas(c),logML(c),'@16')
xlabel('beta')
legend('log ML')
print -deps logML.eps

figure(2); clf; hold on;
plot(betas,mu*ones(1,b),'g-')
plot(betas,mu_MAP,'r-')
plot(betas,mu_MLE*ones(1, b),'b-')
plot(betas(c),mu_MAP(c),'@16')
xlabel('beta')
legend('True','MAP','MLE',4)
print -deps MAP.eps
```


実行例

269



レポート提出に関する注意 270

- 前回(1月14日)と今回(1月18日)レポートの締切はともに次回(1月25日)の講義開始時だが、二つのレポートは別々に作成し、提出すること。
- これまで遅刻レポートを受け取ってきたが、遅刻レポート受け取りの最終期限は講義最終回(1月25日)の講義終了時とする。それ以降は受け取らないので注意すること。

試験について

271

- 2月1日(月)10時40分～12時10分
- S222講義室
- 試験内容:
 - 専門用語の英単語(日本語→英語)
 - 自由記述問題(以下より2問選択)
 - 識別関数のよさを測る規準について
 - 最尤推定法について
 - ノンパラメトリック法について
 - モデル選択について
 - 自分が考えたパターン認識の例について
- 教科書, ノートの持ち込みは不可.