情報認識
「経験ベイズ法（第11章）」

- 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

連絡事項

- １月２０日は，今週の宿題をするための計算機実習とする（出席は取らない）．

- レポートは１月２７日の授業開始時に提出すること

ベイズ推定法

- 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

$$\hat{p}(x) = \int q(x \mid \theta) p(\theta \mid X) d\theta$$

$$X = \{x_1, x_2, \ldots, x_n\}$$

$$p(\theta \mid X) = \frac{p(X \mid \theta) p(\theta)}{\int p(X \mid \theta') p(\theta') d\theta'}$$

- 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

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

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

事前確率の設定

- 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.

事前確率の設定(続き)

- 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

- 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超パラメータ β の値を設定.
- 訓練標本 $X=\{x_i\}_{i=1}^{n}$ が生起する確率:

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

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

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

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

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.

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

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

- 対数をとったほうが計算が簡単なことが多い.

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

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

経験ベイズ法(続き)

- この方法は第Ⅱ種最尤推定法(typeⅡ 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!

最適な超パラメータの探索

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

- 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

周辺尤度の計算

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

- 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

■テイラー展開(Taylor expansion):

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

（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，

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

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

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

：ヘッセ行列 (Hessian matrix)

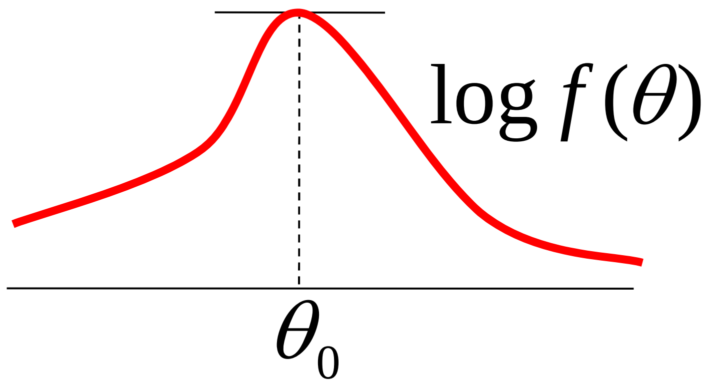
積分の近似

- $\theta_0 = \arg\max_{\theta \in \Theta} f(\theta)$ 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

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



積分の近似(続き)

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

$$\log f(\theta) \approx \underbrace{\log f(\theta_0) + 0 + \tfrac{1}{2}(\theta-\theta_0)^T H(\theta-\theta_0)}_{\log \tilde{f}(\theta)}$$

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

- 指数をとれば,

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

ラプラス近似

- 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，

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

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

- 積分のラプラス近似(Laplace approximation)：

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

ラプラス近似(続き)

- $\log f(\theta)$ 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) とも呼ばれる.
- $f(\theta)$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

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

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

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

$$\log PL(\theta;\beta) = \sum_{i=1}^{n}\log q(x_i \mid \theta) + \log p(\theta;\beta)$$

- 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.
- 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

- 訓練標本が十分多いことを仮定すると,

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

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

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

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

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

- $O(1)$ の項を無視することにすれば

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

ベイズ情報量規準

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

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

- 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.

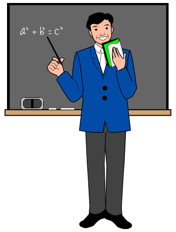
AIC vs. BIC

- AICとBICは似た形：

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

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

- 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

- 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．

小レポート

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）．

小レポート（続き）

- モデル：
$$q(x\mid\mu)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2}\right)$$
- 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（前回の演習と同じ）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．

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

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

小レポート（続き）

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べ．そして，そのようにして
得られた μ を用いて，MAP推定法でβ 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

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);
```

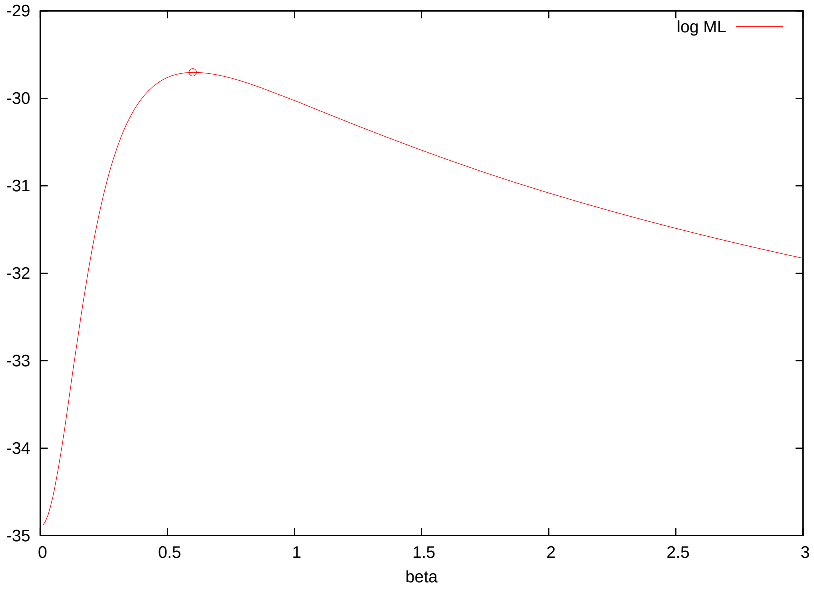
（右上に続く）

（左下から続き）

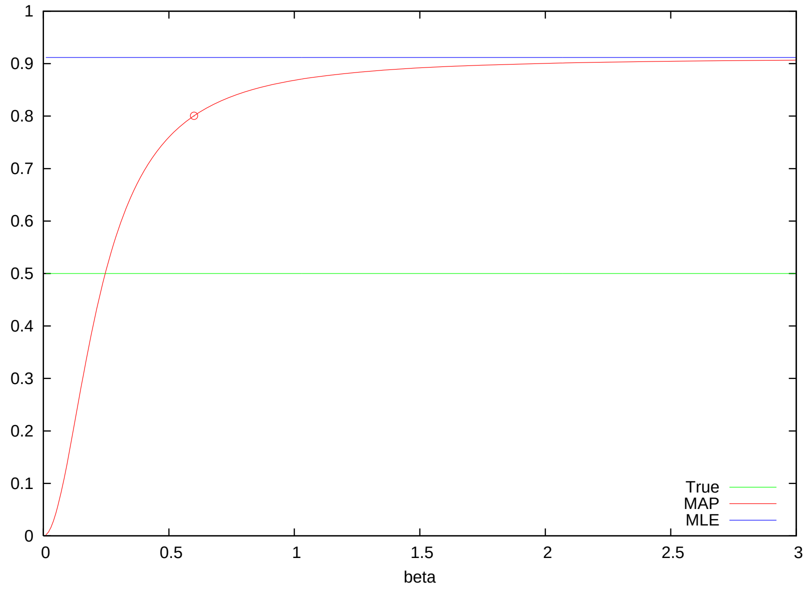
```
figure(1); clf; hold on;
plot(betas,logML,'r-')
plot(betas(c),logML(c),'@16')
xlabel('beta')
legend('log ML')
print -deps logML.eps

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



log ML
-29
-30
-31
-32
-33
-34
-35
0
0.5
1
1.5
2
2.5
3
beta



True
MAP
MLE
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
1.5
2
2.5
3
beta

連絡事項

- １月２０日は，今週の宿題をするための計算機実習とする（出席は取らない）．
- レポートは１月２７日の授業開始時に提出すること

試験について

- 2月3日（月）10時45分～12時15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．

レポート提出に関する注意

- 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, 1月27日の講義終了時とする.
- それ以降は一切受け付けないので注意すること.