早稲田大学「パターン認識」レポート
2005/11/21 出題
解答例と解説

赤穂昭太郎

最終更新：平成 18 年 1 月 18 日

問題

2 クラス問題 $(k = 0, 1)$ ．クラスの確率分布 $q(x|k)$ は，平均 μ_0, μ_1，分散 1 の正規分布とする．各クラス確率は π_0, π_1 とする (ただし $\pi_0 + \pi_1 = 1$ であるため独立なパラメータは π_0 だけを考えればよい).

クラス 0 のサンプル $x_1, \ldots, x_m$ と，クラス 1 のサンプル $x_{m+1}, \ldots, x_n$ が与えられたとき,

1. パラメータの最尤推定量 $\mu_0^*, \mu_1^*, \pi_0^*$
2. パラメータの MAP 推定量 $\mu_0^{\mathrm{MAP}}, \mu_1^{\mathrm{MAP}}, \pi_0^{\mathrm{MAP}}$
3. 事後予測分布 $p(k \mid x, \mathcal{D})$

を求めよ[1]．ただし，π_0 の事前分布 $r(\pi_0)$ は $[0, 1]$ の一様分布，μ_0, μ_1 の事前分布 $r(\mu_0), r(\mu_1)$ は平均 0, 分散 1 の正規分布とする．

解答例

以下では簡単のためパラメータ μ_0, μ_1, π_0 をまとめて θ であらわすことにする．

§ まず最尤推定について

クラス k とデータ x の同時分布は

$$p(x, k \mid \theta) = \pi_k q(x_i \mid k) = \pi_k \exp\left\{-\frac{(x-\mu_k)^2}{2} - \frac{1}{2}\log(2\pi)\right\} \tag{1}$$

である． 従って，全サンプルから計算される対数尤度は,

$$\begin{aligned} L(\mathcal{D}; \theta) &= \sum_{i=1}^{m} \log p(x_i, k=0 \mid \theta) + \sum_{i=m+1}^{n} \log p(x_i, k=1 \mid \theta) \\ &= -\frac{1}{2}\sum_{i=1}^{m}(x_i - \mu_0)^2 + m\log\pi_0 - \frac{m}{2}\log(2\pi) \end{aligned} \tag{2}$$

[1] パラメータの π_k と円周率の π は若干紛らわしいので間違えないように.

$$-\frac{1}{2}\sum_{i=m+1}^{n}(x_i-\mu_1)^2+(n-m)\log\pi_1-\frac{n-m}{2}\log(2\pi) \tag{3}$$

$$= -\frac{1}{2}\sum_{i=1}^{m}(x_i-\mu_0)^2-\frac{1}{2}\sum_{i=m+1}^{n}(x_i-\mu_1)^2$$
$$+m\log\pi_0+(n-m)\log(1-\pi_0)-\frac{n}{2}\log(2\pi) \tag{4}$$

となる．$L(\mathcal{D};\theta)$ をそれぞれのパラメータについて微分して 0 とおいて解くことにより最尤推定量が求まる．

- π_0 については，

$$\frac{\partial L(\mathcal{D};\theta)}{\partial\pi_0}=\frac{m}{\pi_0}-\frac{n-m}{1-\pi_0}=0 \tag{5}$$

より，

$$\boxed{\pi_0^*=\frac{m}{n}}$$

- μ_0 については，

$$\frac{\partial L(\mathcal{D};\theta)}{\partial\mu_0}=\sum_{i=1}^{m}(x_i-\mu_0)=0 \tag{6}$$

より，

$$\boxed{\mu_0^*=\frac{1}{m}\sum_{i=1}^{m}x_i}$$

- μ_1 については，

$$\frac{\partial L(\mathcal{D};\theta)}{\partial\mu_1}=\sum_{i=m+1}^{n}(x_i-\mu_1)=0 \tag{7}$$

より，

$$\boxed{\mu_1^*=\frac{1}{n-m}\sum_{i=m+1}^{n}x_i}$$

§ 次に MAP 推定について

パラメータの事後分布は，事前分布を $r(\pi_0), r(\mu_0), r(\mu_1)$ とおくと，

$$p(\theta\mid\mathcal{D})=\frac{\exp(L(\mathcal{D};\theta))r(\pi_0)r(\mu_0)r(\mu_1)}{p(\mathcal{D})} \tag{8}$$

である．事前分布は

$$r(\pi_0)=1,\qquad 0\le\pi_0\le 1, \tag{9}$$

$$r(\mu_k)=\exp(-\mu_k^2/2)/\sqrt{2\pi},\qquad k=0,1 \tag{10}$$

だから，対数事後分布は，

$$l(\theta\mid\mathcal{D})=\log p(\theta\mid\mathcal{D})=L(\mathcal{D};\theta)-\frac{\mu_0^2}{2}-\frac{\mu_1^2}{2}-\log(2\pi)-\log p(\mathcal{D}). \tag{11}$$

となる．$l(\theta\mid\mathcal{D})$ をそれぞれのパラメータについて微分して 0 とおいて解くことにより MAP 推定量が求まる．

- π_0 については,

$$\frac{\partial l(\theta \mid \mathcal{D})}{\partial \pi_0} = \frac{m}{\pi_0} - \frac{n-m}{1-\pi_0} = 0 \tag{12}$$

より,

$$\boxed{\pi_0^{\mathrm{MAP}} = \frac{m}{n}}$$

- μ_0 については,

$$\frac{\partial l(\theta \mid \mathcal{D})}{\partial \mu_0} = \sum_{i=1}^{m}(x_i - \mu_0) - \mu_0 = 0 \tag{13}$$

より,

$$\boxed{\mu_0^{\mathrm{MAP}} = \frac{1}{m+1}\sum_{i=1}^{m} x_i}$$

- μ_1 については,

$$\frac{\partial l(\theta \mid \mathcal{D})}{\partial \mu_1} = \sum_{i=m+1}^{n}(x_i - \mu_1) - \mu_1 = 0 \tag{14}$$

より,

$$\boxed{\mu_1^{\mathrm{MAP}} = \frac{1}{n-m+1}\sum_{i=m+1}^{n} x_i}$$

§ 最後に事後予測分布について

事後予測分布は,

$$p(k \mid x, \mathcal{D}) = \frac{p(x, k \mid \mathcal{D})}{p(x, k=0 \mid \mathcal{D}) + p(x, k=1 \mid \mathcal{D})} \tag{15}$$

だから,

$$p(x, k \mid \mathcal{D}) = \int p(x, k \mid \theta) p(\theta \mid \mathcal{D}) d\theta \tag{16}$$

が計算できればよい．このうち $p(x, k \mid \theta)$ は (1) 式であり，$p(\theta \mid \mathcal{D})$ は (11) 式の対数事後尤度を使って $p(\theta \mid \mathcal{D}) = \exp(l(\theta \mid \mathcal{D}))$ となる．

それぞれの式の値を代入すると,

$$\begin{aligned} p(x, k \mid \mathcal{D}) = (2\pi)^{(-n+3)/2} \iiint \pi_k \exp\left\{-\frac{(x-\mu_k)^2}{2}\right\} \\ \cdot \exp\left\{ -\frac{1}{2}\sum_{i=1}^{m}(x_i - \mu_0)^2 - \frac{1}{2}\sum_{i=m+1}^{n}(x_i - \mu_1)^2 + \right. \\ \left. m\log\pi_0 + (n-m)\log\pi_1 - \frac{\mu_0^2}{2} - \frac{\mu_1^2}{2}\right\} d\pi_0 d\mu_0 d\mu_1. \end{aligned} \tag{17}$$

ここで[2],

$$\int_0^1 \pi_0^k \pi_1^l d\pi_0 = \int_0^1 \pi_0^k (1-\pi_0)^l d\pi_0 = B(k+1, l+1) = \frac{k!l!}{(k+l+1)!}, \tag{18}$$

$$\int_{-\infty}^{\infty} \exp(-\frac{a}{2}(\mu_k^2 - b\mu_k)) d\mu_k = \sqrt{\frac{2\pi}{a}} \exp\left(\frac{ab^2}{2}\right), \qquad a > 0 \tag{19}$$

[2]なお，$B(x, y)$ はベータ関数．この二つの関係式は覚えておいて損はない重要な式である．

となることを使うとこの積分は解析的に計算できる．以下式を簡単にするため，

$$\mu_0^{\mathrm{MAP}} = \frac{1}{m+1}\sum_{i=1}^{m} x_i, \quad \mu_1^{\mathrm{MAP}} = \frac{1}{n-m+1}\sum_{i=m+1}^{n} x_i, \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \tag{20}$$

と置く[3]．

- $k=0$ のときは[4]，

$$\begin{aligned} p(x, k=0 \mid \mathcal{D}) &= (2\pi)^{(-n+3)/2} \exp\left(-\frac{n}{2}\sigma^2\right) \int_0^1 \pi_0^{m+1}\pi_1^{n-m} d\pi_0 \\ &\quad \cdot \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-\mu_0)^2}{2} - \frac{m+1}{2}(\mu_0^2 - 2\mu_0\mu_0^{\mathrm{MAP}})\right\} d\mu_0 \\ &\quad \cdot \int_{-\infty}^{\infty} \exp\left\{-\frac{n-m+1}{2}(\mu_1^2 - 2\mu_1\mu_1^{\mathrm{MAP}})\right\} d\mu_1 \\ &= (2\pi)^{(-n+3)/2} \exp\left(-\frac{n}{2}\sigma^2\right) B(m+2, n-m+1) \\ &\quad \cdot\sqrt{\frac{2\pi}{m+2}}\, n \exp\left\{-\frac{m+1}{2(m+2)}(x-\mu_0^{\mathrm{MAP}})^2 + \frac{m+1}{2}\left(\mu_0^{\mathrm{MAP}}\right)^2\right\} \\ &\quad \cdot\sqrt{\frac{2\pi}{n-m+1}} \exp\left\{\frac{n-m+1}{2}\left(\mu_1^{\mathrm{MAP}}\right)^2\right\} \end{aligned} \tag{21}$$

- $k=1$ のときもほぼ同様に，

$$\begin{aligned} p(x, k=1 \mid \mathcal{D}) &= (2\pi)^{(-n+3)/2} \exp\left(-\frac{n}{2}\sigma^2\right) \int_0^1 \pi_0^{m}\pi_1^{n-m+1} d\pi_0 \\ &\quad \cdot \int_{-\infty}^{\infty} \exp\left\{-\frac{m+1}{2}(\mu_0^2 - 2\mu_0\mu_0^{\mathrm{MAP}})\right\} d\mu_0 \\ &\quad \cdot \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-\mu_1)^2}{2} - \frac{n-m+1}{2}(\mu_1^2 - 2\mu_1\mu_1^{\mathrm{MAP}})\right\} d\mu_1 \\ &= (2\pi)^{(-n+3)/2} \exp\left(-\frac{n}{2}\sigma^2\right) B(m+1, n-m+2) \\ &\quad \cdot\sqrt{\frac{2\pi}{m+1}} \exp\left\{\frac{m+1}{2}\left(\mu_0^{\mathrm{MAP}}\right)^2\right\} \\ &\quad \cdot\sqrt{\frac{2\pi}{n-m+2}} \exp\left\{-\frac{n-m+1}{2(n-m+2)}(x-\mu_1^{\mathrm{MAP}})^2 + \frac{n-m+1}{2}\left(\mu_1^{\mathrm{MAP}}\right)^2\right\} \end{aligned} \tag{22}$$

これらの結果を (16) 式に代入すると，$k=0$ についての事後予測分布は，

$$\boxed{p(k=0 \mid x, \mathcal{D}) = \frac{\hat{\pi}_0}{Z(x)} \exp\left\{-\frac{m+1}{2(m+2)}(x-\mu_0^{\mathrm{MAP}})^2\right\}}$$

となり，$k=1$ についての事後予測分布は，

$$\boxed{p(k=1 \mid x, \mathcal{D}) = \frac{\hat{\pi}_1}{Z(x)} \exp\left\{-\frac{n-m+1}{2(n-m+2)}(x-\mu_1^{\mathrm{MAP}})^2\right\}}$$

となる． ただし，

$$\hat{\pi}_0 = \frac{(m+1)^{3/2}}{(m+2)^{1/2}n}, \quad \hat{\pi}_1 = \frac{(n-m+1)^{3/2}}{(n-m+2)^{1/2}n} \tag{23}$$

であり[5]，$Z(x)$ はそれぞれの分子の和

$$Z(x) = \hat{\pi}_0 \exp\left\{-\frac{m+1}{2(m+2)}(x-\mu_0^{\mathrm{MAP}})^2\right\} + \hat{\pi}_1 \exp\left\{-\frac{n-m+1}{2(n-m+2)}(x-\mu_1^{\mathrm{MAP}})^2\right\} \tag{24}$$

[3] $\mu_0^{\mathrm{MAP}}, \mu_1^{\mathrm{MAP}}$ については MAP 推定の値そのものである．

[4] 途中の式変形は大幅に省略してあるので確認のこと．

[5] 分母の n は共通なのでなくてもよいが，π_0^* や π_0^{MAP} と比較させやすくするためにつけてある．詳しくは解説の「推定の比較」の表を参照．

解説

§ 問題における暗黙の仮定

問題にはいくつかの**暗黙の仮定**があり，学生諸君にはわかりにくい部分もあったかも知れない．

まず第一に，**未知なる物は何か**ということである．パラメータ μ_0, μ_1, π_0 は未知としている．レポートの中に,「μ_0, μ_1, π_0 は既知なので，最尤推定も MAP 推定も $\mu_0^* = \mu_0^{\mathrm{MAP}} = \mu_0$」などと答えていた人がいるが，それではそもそも問題が成立しない．

次に，授業でも説明したように，**サンプルはみな独立**と仮定している．独立なサンプルの場合には尤度が積の形で (従って，対数尤度が和の形で) 書くことができ，最尤推定などが簡単に求められる．ただし，現実問題ではサンプルがみなサイコロを振るように独立というわけにはいかないのが普通であるので，その場合には注意が必要なこともある．

また，パラメータの事前分布も独立，すなわち，$r(\pi_0, \mu_0, \mu_1) = r(\pi_0)r(\mu_1)r(\mu_2)$ を仮定している．独立性はパラメータに対しても簡単化に重要な要素となっている．

記法に関して，若干の注意．確率分布を表すのに基本的にはすべて $p()$ という記号を使っている． ただし，それ以外のものを使う場合がある．$q(x \mid k)$ や $r(\pi_0)$ などがそれにあたるが，前者は確率モデルを規定していることを強調するため，後者は事前分布であることを強調するために用いた． なお，$p(x)$ に関して紛らわしくないように $p_x(x)$ などと書く流儀もある． もう一つ，確率の教科書などで「確率変数は大文字，その実現値は小文字」という流儀で書かれているものが多いが，この講義では紛らわしくない場合を除いて原則としてすべて小文字で表している．

§ 最尤推定について

最尤推定を求める基本的なやり方は，対数尤度をパラメータについて微分して 0 と置いた方程式を解けばよい．この問題では微分した式が正規分布などの仮定によりいずれも閉じた形で解くことができる．

§MAP 推定について

MAP (Maximum a posteriori：最大事後確率) 推定は，最尤推定の対数尤度に**事前分布の対数を加えただけ**である．

特に，事前分布が一様分布ならば加わる部分が 0 となるため，最尤推定と MAP 推定は一致する． そのため，最尤推定は MAP 推定で事前分布が一様分布の場合とみなすこともできる．ただし，実数空間のように無限に広がった空間では一様分布を考えることはできないので，この問題では平均値に関しては正規分布を仮定している．

発展的な話題　MAP 推定が簡単になるような事前分布として**共役事前分布**と呼ばれる分布があるので興味があれば調べてみるとよい．本問題では，π_0 の事前分布として一様分布を取ったが，$r(\pi_0) = c\pi_0^{\gamma_0}(1-\pi_0)^{\gamma_1}$ としても同じように簡単に解ける (この分布をベータ分布という．ただし c, γ_0, γ_1 は正定数)． γ_0, γ_1 が 0 でなければ，これは確率が 0 や 1 という極端な値を避けると言う事前知識になっている．

§ 事後予測分布について

授業での説明が不十分だったのと，計算が結構複雑なためレポートで正解まで達した人はいなかった．ということでこの部分に関しては失題と言える．

ここでも正規分布の性質から運よく積分を解析的に求めることができたが，一般には積分の計算は困難である．そのためのいろいろな近似法が研究されている．

発展的な話題　代表的な近似法として Laplace 近似 (事後分布を正規分布で近似する)，MCMC 法 (マルコフ連鎖モンテカルロ法: 事後分布に従う乱数を発生させる)，変分ベイズ法 (事後分布を簡単な分布で近似する) の 3 種類があるので興味があれば調べてみるとよい．

§ 推定の比較

最尤推定や MAP 推定では，求めたパラメータを $p(k \mid x, \theta)$ に代入して，$p(k \mid x, \theta^*)$ や $p(k \mid x, \theta^{\mathrm{MAP}})$ により与えられた x のクラスを推定する．このような推定量を**プラグイン推定量**という．

これはパラメータについて平均を取る事後予測分布 (16) 式と大きく違う点である． ただし，この問題の場合は，すべて

$$p(k \mid x) = \frac{\alpha_k}{Z(x)} \exp\left(-\frac{\beta_k}{2}(x - \gamma_k)^2\right) \tag{25}$$

という形をしている． もともと仮定したモデルでは $\beta_k = 1$ となるはずで，プラグイン推定量では確かにそうなっている．一方，事後予測分布ではそうなっていない． 事後予測分布は一般に最初に想定したモデルからはみだしたモデルになる．

$\alpha_0, \beta_0, \gamma_0$ を推定法ごとに表にまとめてみよう．($\alpha_1, \beta_1, \gamma_1$ もほとんど同様)

推定法	α_0	β_0	γ_0
最尤推定	$\pi_0^* = \dfrac{m}{n}$	1	$\mu_0^* = \dfrac{\sum_{i=1}^m x_i}{m}$
MAP 推定	$\pi_0^{\mathrm{MAP}} = \dfrac{m}{n}$	1	$\mu_0^{\mathrm{MAP}} = \dfrac{\sum_{i=1}^m x_i}{m+1}$
事後予測分布	$\hat{\pi}_0 = \dfrac{(m+1)^{3/2}}{(m+2)^{1/2} n}$	$\dfrac{m+1}{m+2}$	μ_0^{MAP}

表を見ればすぐにわかるように，m **の値が大きくなればどれもみな同じものになる**．逆に m が少ない場合，例えば $m = 0$ という，$k = 0$ のサンプルが一つもないときを考えよう． このとき，最尤推定は (β_0 の分母に m があるので) 破綻し，MAP 推定も $k = 0$ の予測確率が 0 になる．一方，事後予測分布では，一つも観測されなくても，それはたまたま観測されないだけだと判断して 0 ではない値を出す．**サンプル数が少ないうちは MAP 推定や事後予測分布は事前分布を重視し，サンプルが増えるにつれてサンプルを重視するようになっていく．**