

早稲田大学「パターン認識」レポート 解答例と 解説

2006-11-27 出題

提出期限: 2006-12-11 授業後

赤穂昭太郎

1 問題

設定 spam メールかどうかを判定するために，5 個の単語をピックアップしてそれぞれの単語が含まれているかどうかの 2 値ベクトル $X = (X_1, \dots, X_5)$ を特徴ベクトルとして用意する．

データ	spam メール (クラス A)	no-spam メール (クラス B)
メール 1:	(1,0,1,0,1)	メール 6: (0,0,1,1,0)
メール 2:	(0,1,0,0,0)	メール 7: (1,0,0,0,1)
メール 3:	(0,0,1,1,1)	メール 8: (0,0,0,1,0)
メール 4:	(1,0,0,1,1)	
メール 5:	(0,0,1,1,1)	
および未分類の メール 9:	(1,0,1,1,0)	

確率モデル

$$p(\mathbf{X} = \mathbf{x}, C = c) = p(\mathbf{X} | C = c)p(C = c), \quad c = A \text{ or } B$$

$$p(\mathbf{X} = \mathbf{x} | C = c) = \prod_{k=1}^5 \mu_{ck}^{x_k} (1 - \mu_{ck})^{1-x_k} \Leftarrow \text{これが間違ってました}$$

$$p(C = A) = q, \quad \text{注) } p(C = B) = 1 - q$$

(1) パラメータ $(\mu_{A1}, \dots, \mu_{A5}, \mu_{B1}, \dots, \mu_{B5}, q)$ の最尤推定量を求めよ．これに基づいてメール 1~9 を spam 判定せよ．

(2) 事前分布として $p(\mu_{ck}) \propto \mu_{ck}(1 - \mu_{ck})$, $p(q) \propto q(1 - q)$ をとる．このとき，

(2-1) パラメータの MAP 推定量を求めよ．これに基づいてメール 1~9 を spam 判定せよ．

(2-2) メール 9 が spam である事後確率を求めよ．

解答例

便宜上以下のように記号を定義する .

- θ : パラメータ集合 $\theta = \{\mu_{A1}, \dots, \mu_{A5}, \mu_{B1}, \dots, \mu_{B5}, q\}$
- K : 特徴ベクトルのサイズ ($K = 5$)
- N_c : クラス c (= A or B) のサンプル数 ($N_A = 5, N_B = 3$)
- $r_c(\mathbf{X}; \theta) = p(\mathbf{X} | c)p(c)$: パラメータ θ を固定したときのクラス c (=A or B) の確率 . 具体的にはクラス A について

$$r_A(\mathbf{X}; \theta) = \prod_{k=1}^K \mu_{kA}^{X_k} (1 - \mu_{kA})^{1-X_k} q \quad (1)$$

であり , クラス B については

$$r_B(\mathbf{X}; \theta) = \prod_{k=1}^K \mu_{kB}^{X_k} (1 - \mu_{kB})^{1-X_k} (1 - q) \quad (2)$$

となる .

- \mathcal{D} : サンプル集合 $\{(\mathbf{X}^{(i)}, c^{(i)})\}_{i=1}^{N_A+N_B}$ つまり , i 番目のメールの特徴ベクトル $\mathbf{X}^{(i)}$ とそれが spam か no-spam であるかのクラスラベル $c^{(i)}$ (=A or B) の集合 .

(1) 最尤推定

パラメータが与えられたときのデータの確率分布 (尤度) は

$$p(\mathcal{D}; \theta) = \prod_{i=1}^{N_A+N_B} p(\mathbf{X}^{(i)} | c^{(i)}) p(c^{(i)}) = \prod_{i=1}^{N_A} r_A(\mathbf{X}^{(Ai)}) \prod_{i=1}^{N_B} r_B(\mathbf{X}^{(Bi)}) \quad (3)$$

である . ただし $\mathbf{X}^{(ci)}$ はクラス c (=A or B) の i 番目のメールの特徴ベクトル¹ .

最尤推定はこの値を最大にするパラメータ値として与えられる . 簡単のため , その対数 (対数尤度) をとって考えると , (1), (2) 式より ,

$$\begin{aligned} L^{\text{ML}} &= \log p(\mathcal{D}; \theta) \\ &= N_A \log q + N_B \log(1 - q) \end{aligned} \quad (4)$$

$$+ \sum_{c \in \{A, B\}} \sum_{k=1}^K [m_{ck} \log \mu_{ck} + (N_{ck} - m_{ck}) \log(1 - \mu_{ck})] \quad (5)$$

¹ 上付き添え字とべき乗が多少紛らわしいので注意 . ここでは上付き添え字を使う場合は括弧をつけた .

となる。ただし、ここでクラス c (=A or B) のサンプルの k 番目の特徴ベクトルのうち 1 のものの総数を m_{ck} と置いた。例えば、クラス A (spam) の特徴ベクトルの第一成分は、メール 1~5 の順に 1, 0, 0, 1, 0 で、1 の数は 2 個なので $m_{A1} = 2$ となる。

最大値を求めるために、 L_{ML} を各パラメータで微分し 0 と置く。まず q については、

$$\frac{\partial L^{ML}}{\partial q} = \frac{N_A}{q} - \frac{N_B}{1-q} = 0 \quad (6)$$

より

$$q^{ML} = \frac{N_A}{N_A + N_B} = 5/8 = 0.625$$

を得る。

次に、 μ_{ck} については、

$$\frac{\partial L^{ML}}{\partial \mu_{ck}} = \frac{m_{ck}}{\mu_{ck}} - \frac{N_{ck} - m_{ck}}{1 - \mu_{ck}} = 0 \quad (7)$$

より、

$$\mu_{ck}^{ML} = \frac{m_{ck}}{N_{ck}} \quad (8)$$

を得る。

m_{ck} は問題のデータから、

c	k=1	k=2	k=3	k=4	k=5
A	2	1	3	3	4
B	1	0	1	2	1

となるから、 μ_{ck}^{ML} は

c	k=1	k=2	k=3	k=4	k=5
A	2/5	1/5	3/5	3/5	4/5
B	1/3	0	1/3	2/3	1/3

となる。

これをもとにして、各メールが spam であるかどうかを判定する。ここでは、誤り率最小になるように、 $p(C = A | \mathbf{X}; \theta) \geq 1/2$ で spam, それ以外で no-spam と判定することにする。

$$p(C = A | \mathbf{X}; \theta) = \frac{p(C = A, \mathbf{X})}{p(\mathbf{X})} = \frac{r_A(\mathbf{X}; \theta)}{r_A(\mathbf{X}; \theta) + r_B(\mathbf{X}; \theta)} = \frac{1}{1 + r_B(\mathbf{X}; \theta)/r_A(\mathbf{X}; \theta)} \quad (9)$$

だから、 $r_B(\mathbf{X}; \theta)/r_A(\mathbf{X}; \theta)$ を計算すればよい。

$$\frac{r_B(\mathbf{X}; \theta^{\text{ML}})}{r_A(\mathbf{X}; \theta^{\text{ML}})} = \frac{1-q}{q} \prod_{k=1}^K \left\{ \left(\frac{\mu_{kB}^{\text{ML}}}{\mu_{kA}^{\text{ML}}} \right)^{X_k} \left(\frac{1-\mu_{kB}^{\text{ML}}}{1-\mu_{kA}^{\text{ML}}} \right)^{1-X_k} \right\} \quad (10)$$

各メールについてこの値を計算して，(9) 式に代入すると以下のようにになる（数値は小数第 2 位未満四捨五入）.

メール	$p(C = A \mathbf{X}; \theta^{\text{ML}})$	判定
1	0.89	spam
2	1.00	spam
3	0.82	spam
4	0.67	spam
5	0.82	spam
6	0.37	no-spam
7	0.73	spam
8	0.16	no-spam
9	0.44	no-spam

(2) ベイズ推定

ベイズ推定ではパラメータも確率変数なので， $p(X; \theta)$ は $p(X | \theta)$ などと書き，前節で定義したものもこれに従うこととする．

(2-1) MAP 推定

MAP 推定では尤度の代わりにパラメータの事後分布 $p(\theta | \mathcal{D})$ を最大化する θ を求め る． 実際には，これに θ に依存しない $p(\mathcal{D})$ をかけて， $p(\theta, \mathcal{D}) = p(\mathcal{D} | \theta)p(\theta)$ を最大 化する． 対数をとって考えれば，これは対数尤度に事前分布の尤度を足した，

$$L^{\text{MAP}} = L^{\text{ML}} + \log p(\theta) \quad (11)$$

を最大化すればよい．

$$\begin{aligned} L^{\text{MAP}} &= (N_A + 1) \log q + (N_B + 1) \log(1 - q) \\ &+ \sum_{c \in \{A, B\}} \sum_{k=1}^K [(m_{ck} + 1) \log \mu_{ck} + (N_{ck} - m_{ck} + 1) \log(1 - \mu_{ck})] \end{aligned} \quad (12)$$

だから，最尤推定の場合と同様にして，

$$q^{\text{MAP}} = \frac{N_A + 1}{N_A + N_B + 2} = 0.6$$

これは最尤推定より少しだけ 0.5 に偏っている .

μ_{ck} についても ,

$$\mu_{ck}^{\text{MAP}} = \frac{m_{ck} + 1}{N_{ck} + 2} \quad (13)$$

であり ,

c	k=1	k=2	k=3	k=4	k=5
A	3/7	2/7	4/7	4/7	5/7
B	2/5	1/5	2/5	3/5	2/5

となる .

また , 各メールが spam であるかどうかの判定は , 最尤推定と同様にして , $r_A(\mathbf{X}; \theta^{\text{MAP}})/r_B(\mathbf{X}; \theta^{\text{MAP}})$ の値に基づいて判定する .

各メールについて計算すると , 以下のようになる .

メール	$p(C = A \mathbf{X}, \theta^{\text{MAP}})$	判定
1	0.80	spam
2	0.43	no-spam
3	0.76	spam
4	0.64	spam
5	0.76	spam
6	0.45	no-spam
7	0.66	spam
8	0.29	no-spam
9	0.48	no-spam

メール 9 についての値を見ると , 最尤推定の時に比べてかなり際どい判定である .

(2-2) 事後確率

サンプル集合 \mathcal{D} が得られているとき , 新たなメール \mathbf{X} のクラス C の事後確率

$$p(C = c | \mathbf{X}, \mathcal{D}) \quad (14)$$

を計算したい . これは

$$p(C = c | \mathbf{X}, \mathcal{D}) = \frac{p(C = c, \mathbf{X} | \mathcal{D})}{p(C = A, \mathbf{X} | \mathcal{D}) + p(C = B, \mathbf{X} | \mathcal{D})} \quad (15)$$

だから右辺の分子が計算できればよい .

これは既知の関数を使って

$$p(C = c, \mathbf{X}, \mathcal{D}) = \int p(C = c, \mathbf{X}, \mathcal{D}, \theta) d\theta = \int p(C = c, \mathbf{X} | \theta) p(\theta | \mathcal{D}) d\theta \quad (16)$$

によって計算できる。この積分は、既知の関数が θ に依存しているので、その θ を消すために必要なものである。また、最右辺の分解は新たなメールが過去のサンプル \mathcal{D} とは θ に関して条件付きで独立に発生することを仮定することから得られる。

$$\begin{aligned} p(C = c, \mathbf{X} \mid \mathcal{D}) &= \int p(C = c, \mathbf{X} \mid \theta) p(\theta \mid \mathcal{D}) d\theta \\ &= \int \prod_{k=1}^K p(\mathbf{X} \mid C = c, \theta) p(C = c) \exp(L^{\text{MAP}}) d\theta \end{aligned} \quad (17)$$

L^{MAP} は (12) 式で与えられるので、まず $c = A$ のときは

$$\begin{aligned} p(C = A, \mathbf{X} \mid \mathcal{D}) &= \int \prod_{k=1}^K [\mu_{Ak}^{X_k} (1 - \mu_{Ak})^{1-X_k}] q \\ &\quad \times q^{N_A+1} (1 - q)^{N_B+1} \prod_{c' \in \{A, B\}} \prod_{k=1}^K \mu_{ck}^{m_{c'k}+1} (1 - \mu_{c'k})^{N_{c'k}-m_{c'k}+1} d\theta \end{aligned} \quad (18)$$

$$\begin{aligned} &= \int_0^1 q^{N_A+2} (1 - q)^{N_B+1} dq \\ &\quad \times \prod_{k=1}^K \left[\int_0^1 \mu_{Ak}^{X_k+m_{Ak}+1} (1 - \mu_{Ak})^{N_A-m_{Ak}+2-X_k} d\mu_{Ak} \right. \\ &\quad \left. \times \int_0^1 \mu_{Bk}^{m_{Bk}+1} (1 - \mu_{Bk})^{N_B-m_{Bk}+1} d\mu_{Bk} \right] \end{aligned} \quad (19)$$

となり、 $c = B$ のときも同様に、

$$\begin{aligned} p(C = B, \mathbf{X} \mid \mathcal{D}) &= \int \prod_{k=1}^K [\mu_{Bk}^{X_k} (1 - \mu_{Bk})^{1-X_k}] (1 - q) \\ &\quad \times q^{N_A+1} (1 - q)^{N_B+1} \prod_{c' \in \{A, B\}} \prod_{k=1}^K \mu_{ck}^{m_{c'k}+1} (1 - \mu_{c'k})^{N_{c'k}-m_{c'k}+1} d\theta \end{aligned} \quad (20)$$

$$\begin{aligned} &= \int_0^1 q^{N_A+1} (1 - q)^{N_B+2} dq \\ &\quad \times \prod_{k=1}^K \left[\int_0^1 \mu_{Ak}^{m_{Ak}+1} (1 - \mu_{Ak})^{N_A-m_{Ak}+1} d\mu_{Ak} \right. \\ &\quad \left. \times \int_0^1 \mu_{Bk}^{X_k+m_{Bk}+1} (1 - \mu_{Bk})^{N_B-m_{Bk}+2-X_k} d\mu_{Bk} \right] \end{aligned} \quad (21)$$

となる。ここで、

$$\int_0^1 x^k (1 - x)^l dx = \frac{k!l!}{(k + l + 1)!} \quad (22)$$

より、

$$p(C = A, \mathbf{X} \mid \mathcal{D}) = \frac{(N_A + 2)!(N_B + 1)!}{(N_A + N_B + 4)!} \times \prod_{k=1}^K \left[\frac{(X_k + m_{Ak} + 1)!(N_A - m_{Ak} + 2 - X_k)!}{(N_A + 4)!} \frac{(m_{Bk} + 1)!(N_B - m_{Bk} + 1)!}{(N_B + 3)!} \right] \quad (23)$$

$$p(C = B, \mathbf{X} \mid \mathcal{D}) = \frac{(N_A + 1)!(N_B + 2)!}{(N_A + N_B + 4)!} \times \prod_{k=1}^K \left[\frac{(m_{Ak} + 1)!(N_A - m_{Ak} + 1)!}{(N_A + 3)!} \frac{(X_k + m_{Bk} + 1)!(N_B - m_{Bk} + 2 - X_k)!}{(N_B + 4)!} \right] \quad (24)$$

クラス A である事後確率は (15) 式から計算できるが、結局比だけが問題となるので、(23) と (24) の比を計算すると、

$$\frac{p(C = B, \mathbf{X} \mid \mathcal{D})}{p(C = A, \mathbf{X} \mid \mathcal{D})} = \frac{N_B + 2}{N_A + 2} \left(\frac{N_A + 4}{N_B + 4} \right)^K \prod_{k=1}^K f(X_k, m_{Ak}, m_{Bk}) \quad (25)$$

となる。ただし、 $f(X_k, m_{Ak}, m_{Bk})$ は

$$f(X_k, m_{Ak}, m_{Bk}) = \frac{(m_{Ak} + 1)!(N_A - m_{Ak} + 1)!(X_k + m_{Bk} + 1)!(N_B - m_{Bk} + 2 - X_k)!}{(X_k + m_{Ak} + 1)!(N_A - m_{Ak} + 2 - X_k)!(m_{Bk} + 1)!(N_B - m_{Bk} + 1)!} \quad (26)$$

で、 $X_k = 1$ のとき

$$f(1, m_{Ak}, m_{Bk}) = \frac{m_{Bk} + 2}{m_{Ak} + 2}, \quad (27)$$

$X_k = 0$ のとき

$$f(1, m_{Ak}, m_{Bk}) = \frac{N_B - m_{Bk} + 2}{N_A - m_{Ak} + 2} \quad (28)$$

となる。メール 9 の特徴ベクトルの値 $\mathbf{X} = (1, 0, 1, 1, 0)$ や各定数の値を代入すると、

$$\frac{p(C = B, \mathbf{X} \mid \mathcal{D})}{p(C = A, \mathbf{X} \mid \mathcal{D})} = \frac{5}{7} \times \left(\frac{9}{7} \right)^5 \times \frac{3}{4} \times \frac{5}{6} \times \frac{3}{5} \times \frac{4}{5} \times \frac{4}{3} \quad (29)$$

これを使って事後確率は

$$p(C = A \mid \mathbf{X}, \mathcal{D}) = \frac{1}{1 + p(C = B, \mathbf{X} \mid \mathcal{D})/p(C = A, \mathbf{X} \mid \mathcal{D})} \simeq [0.499] \quad (30)$$

となる。これは、ほとんど 0.5 に一致している。

解説

- 基本的には 2005 年度課題と類似の問題なので , 共通部分については以下の pdf の解説を参照のこと .<http://www.neurosci.aist.go.jp/~akaho/waseda/report-1.pdf>
- 付け加えるとすれば , 今回の spam メールの例は , 「ナイーブベイズ法」になっていることに注意 .
- この解答例では , 誤り率最小基準で spam か no-spam を判定したが , 実際には no-spam を spam と判定されては困るので , 「コスト関数」を導入し , その期待値ができるだけ小さくなるような判定も考えられる . これは簡単にできるので , 各自試して頂きたい .
- 完全に正解していた人が今年度は 2 名いた .