

前項までで、回帰問題に対してRVMモデルを適用する方法とその性質を学んだ。RVMはSVMと比べて予測時間が短く、汎化性能も高い一方、学習時間が長いという性質があった。

今回の輪講では、RVMモデルの各パラメータにおける解を逐次的に求めていくことで学習時間を削減する手法を扱い（7.2.2続き）、次に分類問題に対するRVMの適用を扱う（7.2.3）。

7.2.2（続き）

7.2.1ではRVMの事前分布のハイパーパラメータをエビデンス近似（第二種の最尤推定）によって求める方法を学んだ。このときの更新式は陽な形で現されず、パラメータ自身を含む形であった。

$$\alpha_i^{\text{new}} = \frac{r_i}{m_i^2} \quad \leftarrow \alpha_i \text{ を含む} \quad (7.87)$$

$$(\beta^{\text{new}})^{-1} = \frac{\|t - \Phi m\|^2}{N - \sum_i r_i} \quad \leftarrow \beta \text{ を含む} \quad (7.88)$$

一方で、7.2.2 で示した α_i の更新式は α_i 以外の変数を固定して求めたことで陽な形で現された。

$$\alpha_i = \frac{S_i^2}{g_i^2 - S_i} \quad (7.101)$$

$$\begin{pmatrix} S_i = \phi_i^T C_{-i}^{-1} \phi_i \\ g_i = \phi_i^T C_{-i}^{-1} t \end{pmatrix}$$

α_i を含まない。

この性質を用いた学習アルゴリズムは**逐次的疎ベジアン学習アルゴリズム**と呼ばれ、エビデンス近似によるパラメータ更新よりも効率的に学習できることを説明する。

＜逐次的疎ベジアン学習アルゴリズム＞

1. 回帰問題を解く場合には、 β を適当に初期化する。
2. 基底関数 φ_i によって $i = 1$ の S_i, g_i, α_i を初期化する。 $i = 1$ 以外の α_i は無限大とする。
3. \mathcal{E} と M を (7.82, 7.83) から計算すると C_i が求まるので、全ての i について S_i, g_i, α_i を計算する。
4. 次の条件でモデルに追加する候補の基底関数 φ_i を選ぶ。
 $\langle g_i^2 \rangle S_i$ のとき (7.101) で α_i を求め対応する基底関数 φ_i をモデルに追加する
 $\langle g_i^2 \rangle \leq S_i$ のとき $\alpha_i = \infty$ として、対応する基底関数 φ_i をモデルから除外する
5. β を (7.88) で更新する
6. 収束条件を満たした場合は終了し、そうでない場合3から繰り返す

逐次的疎ベジアン学習アルゴリズムがエビデンス近似による解法よりも学習効率が良い理由は、各イテレーションごとに疎な解が得られているためである。エビデンス近似では毎回全ての基底関数から行列 Φ, Σ を計算する必要があるが、逐次的な方法では少数の基底関数のみからなる行列 Φ', Σ' を計算するだけで済むため計算効率が大幅に削減される。

7.2.3 分類問題に対するRVM

回帰問題におけるRVMの議論は、分類問題にも同様に適用することができる。すなわち、各重みパラメータごとにハイパーパラメータを設定し、エビデンスを最大にするようなハイパーパラメータを繰り返し処理で導出することができる。また、「7.2.2 疎性の解析」で扱った、疎性を活かした逐次的なアルゴリズムも適用可能である。

ただし、下記の点については回帰問題と異なるので注意が必要である。

	回帰問題	分類問題
予測モデル	$w^T \phi(x)$	$\sigma(w^T \phi(x))$
尤度関数	正規分布	二項分布
w の事後分布 $p(w \mathcal{D}, \alpha)$	解析的に 正規分布となる。	解析的に求まらないので ラプラス近似する。

これらの違いに着目しながら学習における手続きを解説する。

まずはモデルの形式を次のように定義する。

$$\text{モデル: } y(x, w) = \sigma(w^T \phi(x))$$

$$\text{事前分布: } p(w | \alpha) = \prod_{i=1}^M \mathcal{N}(w_i | 0, \alpha_i^{-1})$$

ラプラス近似により w の事後分布を求める。ラプラス近似は過去資料を参考に、一般に次のステップで行われる。

(ref. 4.4節, 210306 雑談.pdf)


- ① ・ 真の分布のモードを見つける。
- ② $\int f(z)$ について2階微分をかます。
- ③ $\left\{ \begin{array}{l} \cdot f(z) \propto \text{ガウス分布 となるので、全積分が1になるように規格化する。} \\ \cdot \text{規格化によって求めた確率密度分布を } q(z) \text{ とする。これが求めるべき近似分布となる。} \end{array} \right.$

ここで、 w の事後分布は $p(w | t, \alpha)$ であるが、 f を次のように定義すると教科書の表記とつじつまが合う。

$$f(w) = p(w | t, \alpha) p(t | \alpha)$$

$$\left(\Leftrightarrow p(w | t, \alpha) = \frac{1}{p(t | \alpha)} f(w) \right) \quad \begin{array}{l} \text{上巻 p213} \\ \text{(ref. 4.125)} \end{array}$$

求めた分布
↓
f の正規化係数であり、
モデルのエビデンスでもある。

 真の分布のモードを見つける。

事後分布のモードを求めるために、二次微分まで考慮した繰り返し処理を行う。

$$\begin{aligned} \ln p(w | t, \alpha) &= \ln \left\{ \underbrace{p(t | w)}_{\text{二項分布}} \underbrace{p(w | \alpha)}_{\text{正規分布}} \right\} - \underbrace{p(t | \alpha)}_{\text{const.}} \\ &= \sum_{n=1}^N \left\{ t_n \ln y_n + (1 - t_n) \ln (1 - y_n) \right\} - \frac{1}{2} w^T A w + \text{const.} \end{aligned}$$

$$\begin{aligned} \nabla \ln p(w | t, \alpha) &= \sum_{n=1}^N \left\{ \frac{t_n}{y_n} y_n' + \frac{1 - t_n}{(1 - y_n)} (-y_n)' \right\} - A w \\ &= \sum_{n=1}^N \left\{ \frac{t_n}{y_n} y_n (1 - y_n) - \frac{1 - t_n}{1 - y_n} y_n (1 - y_n) \right\} - A w \\ &= \sum_{n=1}^N \left\{ \underbrace{t_n (1 - y_n)}_{\text{正の項}} \underbrace{\phi(x_n)}_{\text{正の項}} - \underbrace{(1 - t_n) y_n}_{\text{正の項}} \underbrace{\phi(x_n)}_{\text{正の項}} \right\} - A w \\ &= \sum_{n=1}^N \left\{ t_n \phi(x_n) - y_n \phi(x_n) \right\} - A w \\ &= \sum_{n=1}^N \phi(x_n) (t_n - y_n) - A w \\ &= \Phi^T (t - y) - A w \end{aligned}$$

$$\begin{aligned}
\nabla \nabla \ell_{np}(w | t, y) &= \sum_{n=1}^N \phi(x_n) (-y_n)' - A \\
&= - \sum_{n=1}^N \phi(x_n) y_n (1 - y_n) \phi(x_n)' - A \\
&= - \Phi^T \underbrace{\begin{pmatrix} y_1(1-y_1) & & 0 \\ & \ddots & \\ 0 & & y_N(1-y_N) \end{pmatrix}}_{\equiv B \text{ とおく}} \Phi - A \\
&= - (\Phi^T B \Phi + A)
\end{aligned}$$

一次微分=0とすることで、 $w = A^{-1} \Phi^T (t - y)$ よりモードが求まるように思えるが、 y の中に w が含まれているため陽に表すことが出来ない。そこで、逐次的再重み付け最小2乗法でモードを見つける。逐次的再重み付け最小2乗法は、ニュートン法の更新式 $w^{new} = w^{old} - H^{-1} \nabla E(w^{old})$ において、ヘッセ行列が重みの更新ごとに変化するケースの更新式である。(ref. 210306輪講.pdf)

これにより得られたモードを w^* とすると、次の式が成り立つ。

$$w^* = A^{-1} \Phi^T (t - y)$$

■ $\ln f(z)$ について二次微分をかます

$$\begin{aligned}\frac{d^2}{dw^2} \ln f(w) &= \nabla \nabla \ln p(w|\xi, \alpha) + 0 \\ &= -(\Phi^T B \Phi + A) \\ &= -\Sigma^{-1} \quad (\text{と } \hat{\alpha})\end{aligned}$$

■ $f(z) \propto$ ガウス分布 となるので、全積分が1になるよう規格化する。

$f(w)$ は w^* 周りで分配が0になり、この点の周りで $f(w)$ をテイラー展開すると、

$$\ln f(w) \simeq \ln f(w^*) - \frac{1}{2} (w - w^*)^T \Sigma^{-1} (w - w^*)$$

となり、

$$f(w) \simeq f(w^*) \exp \left\{ -\frac{1}{2} (w - w^*)^T \Sigma^{-1} (w - w^*) \right\} \quad (*)$$

で近似できると、

$p(w|\xi, \alpha)$ はこれを規格化した次の式で与えられる。

$$\begin{aligned}p(w|\xi, \alpha) &= \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (w - w^*)^T \Sigma^{-1} (w - w^*) \right\} \\ &= \mathcal{N}(w|w^*, \Sigma)\end{aligned}$$

ラプラス近似により w の事後分布が求まったので、これを用いてエビデンスが最大となるハイパーパラメータ α_i を推定する。モデルエビデンス（周辺尤度関数）は次のようになる。

$$\begin{aligned}
 p(t|\alpha) &= \int p(t|w) p(w|\alpha) dw \\
 &\quad \rightarrow \text{事後分布の形がないのでムリやり作る。} \\
 &= \int p(w, t|\alpha) dw \\
 &= \int \underbrace{p(w|t, \alpha)}_{\text{ここが } f(w) \text{ に近づけるに、数値的に設定して、}} p(t|\alpha) dw \\
 &= \int f(w) dw \\
 &\simeq f(w^*) \int \exp\left\{-\frac{1}{2}(w-w^*)^T \Sigma^{-1}(w-w^*)\right\} dw \\
 &= f(w^*) \cdot (2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}} \\
 &= p(t|w^*) \cdot p(w^*|\alpha) \cdot (2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}
 \end{aligned}$$

これを最大にする α を求めるために、対数周辺尤度を α_i について微分して停留点を求める。

$$\begin{aligned}
 \frac{\partial}{\partial \alpha_i} \ln p(t|\alpha) &= \frac{\partial}{\partial \alpha_i} \ln \left\{ p(t|w^*) p(w^*|\alpha) \cdot (2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}} \right\} \\
 &= \frac{\partial}{\partial \alpha_i} \left\{ \underbrace{\sum_{n=1}^N (t_n \ln y_n^* + (1-t_n) \ln (1-y_n^*))}_{4 \text{ 行の } \Sigma \text{ 行列と } w^* \text{ の形}} \right. \\
 &\quad \left. + \sum_{n=1}^m \ln N(w_n^* | 0, \alpha_n^{-1}) + \frac{1}{2} \ln |\Sigma| \right\} \\
 &\quad (\text{ここで、 } y_n^* = \sigma(w^* \Phi(x_n)) \text{ とする})
 \end{aligned}$$

$$= \sum_{n=1}^N \left\{ t_n (1 - y_n^*) \frac{\partial}{\partial \alpha_i} (w^{*T} \phi(x_n)) - (1 - t_n) y_n^* \frac{\partial}{\partial \alpha_i} (w^{*T} \phi(x_n)) \right\} \\ + \frac{\partial}{\partial \alpha_i} \left\{ \ln N(w_i^* | 0, \alpha_i^{-1}) + \frac{1}{2} \ln |\Sigma| \right\}$$

$$= \sum_{n=1}^N (t_n - y_n^*) \frac{\partial}{\partial \alpha_i} \left\{ (t - y)^T \Phi A^{-1} \cdot \phi(x_n) \right\} \\ + \frac{\partial}{\partial \alpha_i} \left\{ -\frac{1}{2} \alpha_i w_i^{*2} + \frac{1}{2} \ln |\Sigma| \right\} \quad \hookrightarrow \text{ここが5行5列}$$

$$= \text{???} - \frac{1}{2} w_i^{*2} + \frac{1}{2} \text{Tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \alpha_i} \right) \quad \left(\begin{array}{c} \text{付録 C} \\ \text{C.2.3} \end{array} \right)$$

$$= \text{???} - \frac{1}{2} w_i^{*2} + \frac{1}{2} \text{Tr} \left\{ \Sigma^{-1} \cdot \left(-\Sigma \frac{\partial \Sigma^{-1}}{\partial \alpha_i} \Sigma \right) \right\} \quad (\text{C.2.1})$$

$$= \text{???} - \frac{1}{2} w_i^{*2} - \frac{1}{2} \text{Tr} \left\{ \frac{\partial}{\partial \alpha_i} (\Phi^T B \Phi + A) \cdot \Sigma \right\}$$

$$= \text{???} - \frac{1}{2} w_i^{*2} - \frac{1}{2} \text{Tr} \left\{ \left(0 + \begin{pmatrix} 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \end{pmatrix} \right) \cdot \Sigma \right\}$$

こことあやしい。

$$= \text{???} - \frac{1}{2} w_i^{*2} - \frac{1}{2} \text{Tr} \left\{ \begin{pmatrix} 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \end{pmatrix} \cdot \Sigma \right\}$$

$$= \text{???} - \frac{1}{2} w_i^{*2} - \frac{1}{2} \sum i i$$

$$\left(\begin{array}{l} \text{教科書と見比べると} \\ = \frac{1}{2 \alpha_i} - \frac{1}{2} w_i^{*2} - \frac{1}{2} \sum i i \quad (7.115) \end{array} \right)$$

この一次微分が 0 となる等式を立てると、

$$\alpha_i^{(n+1)} = \frac{\gamma_i}{(w_i^*)^2} \quad (\text{ただし } \gamma_i = (-d_i \sum ii))$$

が得られる。

これによりRVMのハイパーパラメータを推定することができる。ここでは詳しく解説していないが、前節同様に多くのパラメータ α_i は無限大となり疎な解が得られる。また、
について陽な形で更新式を構築することで高速なアルゴリズムも同様に構築できる。

陽な形の更新式の導出手順をおさらいする。回帰問題の時には周辺尤度を行列 C で表現し、周辺尤度を α_i に関する部分とそうでない部分を切り分けることで導出していた。

$$\ln p(\mathbf{t} | \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{1}{2} \left\{ N \ln(2\pi) + \ln |C| + \mathbf{t}^T C^{-1} \mathbf{t} \right\} \quad (7.85)$$

$$|C| = |C_{-i}| (1 + \alpha_i^{-1} \boldsymbol{\varphi}_i^T C_{-i}^{-1} \boldsymbol{\varphi}_i) \quad (7.94)$$

$$C^{-1} = C_{-i}^{-1} - \frac{C_{-i}^{-1} \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T C_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T C_{-i}^{-1} \boldsymbol{\varphi}_i} \quad (7.95)$$

$$\text{よって、} L(\boldsymbol{\alpha}) = L(\boldsymbol{\alpha}_{-i}) + \lambda(\alpha_i) \quad \text{に分離し、}$$

$\lambda(\alpha_i)$ を最大化する α_i に陽な解を得る。

分類問題でも同様の手続きを行えることを確かめるために、対数周辺尤度を行列 C で表現できるところまでを確認する。そのために目標ラベルを次のように変換する。

$$\hat{\mathbf{t}} = \Phi \mathbf{w}^* + \mathbf{B}^{-1}(\mathbf{t} - \mathbf{y})$$

すると対数周辺尤度は次のように行列 C で現される。

$$\ln p(\mathbf{t}|\alpha) = -\frac{1}{2} \left\{ N \ln(2\pi) + \ln|C| + (\hat{\mathbf{t}})^T C^{-1} \hat{\mathbf{t}} \right\}$$

$$C = B + \Phi A \Phi^T \quad (\Phi \text{ は } n \times m \text{ 行列})$$

これを用いて α_i に関する陽な更新式を求めることで、より高速な逐次的疎ベジアン学習アルゴリズムを構築できる。

次の図7.12は人口データにRVMを適用した結果である。通常のSVMとは異なり、分類面から離れたデータについても関連ベクトルとなりうる性質がある。また、右側の図のように領域ごとのラベルの確率を出力することができるのはSVMと比べて有利な点である。

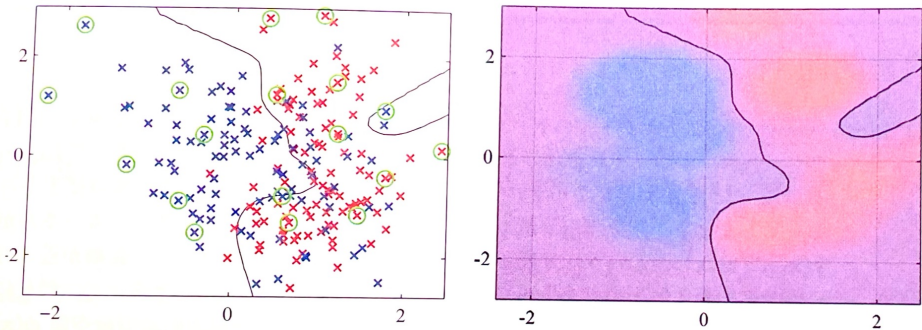


図 7.12

また、RVMではSVMと異なり多クラス分類に拡張しやすいというメリットがある。

多クラスに拡張する場合、シグモイド関数の代わりにソフトマックス関数を代用すればよい。

	2クラス分類	多クラス分類
予測確率	$y(x) = \sigma(w^T \phi(x))$	$y_k(x) = \frac{\exp(a_k)}{\sum \exp(a_k)}$ $i=1 \sim K, \quad a_k = (w_k^T \phi(x))$
損失関数	$l = \sum_{n=1}^N \{t_n \ln y_n + (1-t_n) \ln (1-y_n)\}$	$l = \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}$

RVMについてまとめると、SVMにベイズ的な観点を取り入れ、機能マシマシにしたモデルである。

特に、(i). 出力ラベルの確率を出力することができる点と、(ii). 多クラス分類に応用できる点は特筆すべき点である。代わりに計算量が増加してしまうので、逐次的疎ベジアン学習アルゴリズムのような計算量の工夫が必要になってくる。