情報数理科学 VII　　教師あり学習

松島　慎

2018 年 10 月 9 日

$(\mathbf{x}_i, y_i)$ の組がいくつか与えられている時 $(i = 1, 2, \ldots, n)$、未知のデータを予測したい。どのように選べばよいか？

1　最も簡単な例（線形回帰の最小二乗法）

- 回帰問題
  $\mathbf{x} \in \mathbb{R}^d$,$y_i \in \mathbb{R}$ のとき、回帰という。線形モデルを用いた予測
  $\mathbb{R}^d$ から $\mathbf{w}$ を一つ選び（＝学習し）、未知の $\mathbf{x}$ に対し、$\hat{y}$ を以下の式で予測する

$$\hat{y} = \sum_{j=1}^{d} w_j x_j = \langle \mathbf{w}, \mathbf{x} \rangle \tag{1}$$

  ここで $\langle \mathbf{a}_1, \mathbf{a}_2 \rangle$ は $\mathbf{a}_1$ と $\mathbf{a}_2$ の内積。$\mathbf{x}$ の第 j 要素を x_j とする。$\mathbf{w}$ の第 j 要素を w_j とする。

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

- 線形モデル

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j$$
$$F = \left\{ f : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle \,|\, \mathbf{w} \in \mathbb{R}^d \right\}$$

  バイアス項も次元をあげることで考慮できる。

  （教師ありだけでなく、一般に）学習とは、与えられたデータをもとに、モデル（関数の集合）F の中から予測器（関数）を一つ選ぶこと
- 最小二乗法
  以下の目的関数 $J(\mathbf{w})$ を最小化するような $\mathbf{w}$ を一つ選び、それを予測器とする。

$$\begin{aligned} J(\mathbf{w}) &= \sum_i \left(y_i - \sum_j w_j x_{ij} \right)^2 \\ &= \sum_i \left(y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle \right)^2 \\ &= \| \mathbf{y} - X\mathbf{w} \|_2^2 . \end{aligned}$$

目的関数は経験誤差。本当の目的は汎化誤差 $(= \mathbb{E}_{\mathbf{x},y} \left(y - \langle \mathbf{w}, \mathbf{x} \rangle \right)^2)$
ここで

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}, X = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1d} \\ \vdots & \ddots & & \\ x_{1n} & & & x_{dn} \end{bmatrix}}_{\text{計画行列（データ行列）}} .$$

最小解は以下のようになる。

$$\begin{aligned} &\mathbf{w}^* = \operatorname{argmin} J(\mathbf{w}) \\ &\quad \overbrace{\text{グラム行列 (GramianMatrix)}} \\ &\to \qquad \overbrace{X^\top X} \qquad\qquad \mathbf{w}^* = X^\top \mathbf{y} \\ &\to \mathbf{w}^* = \left(X^\top X \right)^{-1} X^\top \mathbf{y} \end{aligned}$$

※ $X^\top X$ に逆行列が存在する場合

$$J(\mathbf{w}) = \sum_i \left(y_i - \sum_j w_j x_{ij} \right)^2$$
$$\nabla J(\mathbf{w}) = \begin{bmatrix} \frac{\partial J(\mathbf{w})}{\partial w_1} \\ \vdots \\ \frac{\partial J(\mathbf{w})}{\partial w_d} \end{bmatrix} .$$

$$\begin{aligned} \frac{\partial J(\mathbf{w})}{\partial w_{j'}} &= \frac{\partial}{\partial w_{j'}} \sum_i \left(y_i - \sum_j w_j x_{ij} \right)^2 \\ &= \sum_i -2 x_{ij'} \left(y_i - \sum_j w_j x_{ij} \right) \\ &= -2 \underbrace{\sum_i x_{ij'} y_i}_{\left\langle X^\top y, \mathbf{e}_{j'} \right\rangle} + 2 \underbrace{\sum_i \sum_j x_{ij'} x_{ij} w_j}_{\left\langle X^\top X \mathbf{w}, \mathbf{e}_{j'} \right\rangle} \\ &\Rightarrow \nabla J(\mathbf{w}) = -2X^\top y + 2X^\top X \mathbf{w} \end{aligned}$$

$$
\begin{aligned}
&J(\mathbf{w}) = \sum_i \left(y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle\right)^2 \\
&\nabla J(\mathbf{w}) = \sum_i \left(2(y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)\right) \mathbf{x}_i. \\
&(\nabla_{\mathbf{x}} F(\langle \mathbf{a}, \mathbf{x} \rangle) = F'(\langle \mathbf{a}, \mathbf{x} \rangle)\mathbf{a}) \\
&= 2\sum_i \mathbf{x}_i y_i - 2\left(\sum_i \mathbf{x}_i \mathbf{x}_i^\top\right)\mathbf{w}. \\
&= 2\begin{bmatrix}\mathbf{x}_1 \dots \mathbf{x}_n\end{bmatrix}\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - 2\begin{bmatrix}\mathbf{x}_1 \dots \mathbf{x}_n\end{bmatrix}\begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix}\mathbf{w}.
\end{aligned}
$$

$$
\begin{aligned}
&J(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|_2^2. \\
&\nabla J(\mathbf{w}) = X^\top \left(2(\mathbf{y} - X\mathbf{w})\right). \\
&\left(\nabla_{\mathbf{x}} F(A\mathbf{x}) = A^\top \nabla_{\mathbf{u}} F(\mathbf{u}) \mid_{\mathbf{u}=A\mathbf{x}}, \nabla_{\mathbf{x}} \|\mathbf{x}\|_2^2 = 2\mathbf{x}\right)
\end{aligned}
$$

2 過学習と正則化

- 過学習
  モデルが複雑すぎて、得られた予測器の経験誤差は低いのに未知のデータに対する誤差が高くなってしまうこと。
- 未学習
  モデルが単純すぎて、与えられたデータをうまく説明できず経験誤差が高くなること
- 正則化
  モデルの複雑さを連続的に小さくすることを可能にする方法

- Ivanov 型正則化
  - $$J(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|_2^2$$
    を制約 $\|\mathbf{w}\|_2 \leq \tau$ のもとで最小化
  - $$J(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|_2^2$$
    を制約 $\|\mathbf{w}\|_1 \leq \tau$ のもとで最小化
- Tikhonov 型正則化
  - リッジ線形回帰
    $$
    \begin{aligned}
    &J(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|_2^2 + \frac{\lambda}{2}\langle \mathbf{w}, \mathbf{w} \rangle \\
    &\Rightarrow \mathbf{w}^* = \left(X^\top X + \lambda I\right)^{-1} X\mathbf{y}
    \end{aligned}
    $$
  - LASSO 回帰 (Least Absolute Shrinkage and Selection Operator)
    $$J(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda \sum_j |w_j|$$
- Ivanov 型正則化と Tikhonov 型正則化にはある種の同値性がある
  $$
  \begin{aligned}
  &\mathbf{w}_\lambda^* = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{y} - X\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \\
  &\Rightarrow -2X^\top y + 2X^\top X\mathbf{w}_\lambda^* + \lambda \mathbf{w}_\lambda^* = \mathbf{0}
  \end{aligned}
  $$

  $$
  \begin{aligned}
  &\mathbf{w}_\tau^* = \operatorname*{argmin}_{\|\mathbf{w}\|_2^2 \leq \tau} \|\mathbf{y} - X\mathbf{w}\|_2^2 \\
  &\Rightarrow \mathbf{w}_\tau^* = \operatorname*{argmin}_{\mathbf{w}} \max_{\alpha \geq 0} \|\mathbf{y} - X\mathbf{w}\|_2^2 - \alpha(\tau - \|\mathbf{w}\|_2^2) \\
  &\Rightarrow \exists \alpha^* \geq 0, \mathbf{w}_\tau^* = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{y} - X\mathbf{w}\|_2^2 - \alpha^*(\tau - \|\mathbf{w}\|_2^2) \\
  &\Rightarrow -2X^\top y + 2X^\top X\mathbf{w}_\tau^* + 2\alpha^* \mathbf{w}_\tau^* = \mathbf{0}
  \end{aligned}
  $$

3 交差検証

検証セットによる検証を行い学習の正当性を検証する。学習曲線を描いて過学習と未学習を観察し、適切なモデルを選ぶ。

- ホールドアウト法
  X^{train} と X^{valid} が全データ X の分割になるように、分割する。y も同様に分割する。訓練データ X^{train} と y^{train} の組を使って目的関数を設計、その最小解として
  $$\hat{\mathbf{w}}_\lambda$$
  を得る。
  $$\text{RMSE} = \left\|\mathbf{y}^{\text{valid}} - X^{\text{valid}}\hat{\mathbf{w}}_\lambda\right\| / \sqrt{n}$$
  が最小になるように λ を決める。
- K-分割法
  各 k に関して $X^{\text{train},k}$ と $X^{\text{valid},k}$ が全データ X の分割になるように、また $X^{\text{valid},k}\,(k = 1, \dots, K)$ が X の分割になるようにデータを分割する。各 k に関して訓練データ $X^{\text{train},k}$ と $y^{\text{train},k}$ の組（データ数を n_k とする）を使って目的関数を設計、その最小解として
  $$\hat{\mathbf{w}}_{\lambda,k}$$
  を得る。これらを用いて
  $$\bar{\text{RMSE}} = K^{-1}\sum_{k=1}^{K} \left\|\mathbf{y}^{\text{valid},k} - X^{\text{valid},k}\hat{\mathbf{w}}_{\lambda,k}\right\| / \sqrt{n_k}$$
  を計算し、これがが最小になるように λ を決める。

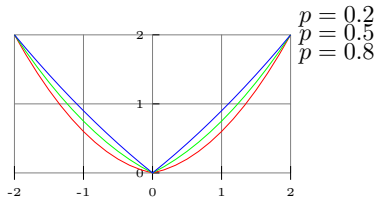

図 1　エラスティックネット正則化項

- 反復無作為抽出法
  無作為に検証データをサンプルし各 λ の性能を評価
- LOO 交差検証
  $=K = n$ で行う K-分割法。

4　正則化付き経験リスク最小化問題

正則化付き損失最小化問題 (Regularized Loss Minimization, RLM) とは以下の最小化問題

$$J(\mathbf{w}) = \sum_{i=1}^{n} \ell_i(\mathbf{w}) + \lambda r(\mathbf{w})$$

ここで r は正則化項。第一項は対して損失項という。損失項と正則化項を選ぶことで様々な問題がこれを用いて記述できる。

正則化項の例

- L_2 正則化項

$$r(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|_2^2$$

- L_1 正則化項

$$r(\mathbf{w}) = \|\mathbf{w}\|_1$$

- L_p 正則化項

$$r(\mathbf{w}) = \frac{1}{p} \|\mathbf{w}\|_p^p$$

- L_∞ 正則化項

$$r(\mathbf{w}) = \|\mathbf{w}\|_\infty$$

- L_0 正則化項

$$r(\mathbf{w}) = \|\mathbf{w}\|_0$$

- エラスティックネット正則化項

$$r(\mathbf{w}) = p \cdot \frac{1}{2} \|\mathbf{w}\|_2^2 + (1-p) \|\mathbf{w}\|_1$$

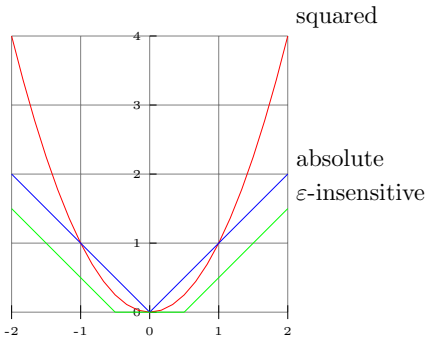

図 2　回帰問題の損失関数

4.1　回帰問題

- 二乗損失

$$\ell_i(\mathbf{w}) = (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$$

- 絶対損失

$$\ell_i(\mathbf{w}) = |\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i|$$

- ε 許容損失 (ε-insensitive loss)

$$\ell_i(\mathbf{w}) = \max\left(-(\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i) - \varepsilon, 0, (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i) - \varepsilon\right)$$

4.2　二値分類問題

$y_i \in \{+1, -1\}$ の場合、二値分類問題という。

- 識別関数

$$\hat{y}_i = \begin{cases} +1 & \langle \mathbf{w}, \mathbf{x}_i \rangle > 0 \\ -1 & \langle \mathbf{w}, \mathbf{x}_i \rangle \leq 0 \end{cases}$$

- ヒンジ損失関数

$$\begin{aligned} \ell_i(\mathbf{w}) &= \begin{cases} \max(0, 1 - \langle \mathbf{w}, \mathbf{x}_i \rangle) & y_i = +1 \\ \max(0, 1 + \langle \mathbf{w}, \mathbf{x}_i \rangle) & y_i = -1 \end{cases} \\ &= \max(0, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \end{aligned}$$

- ロジスティック損失関数

$$\ell_i(\mathbf{w}) = \log(1 + \exp(-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle))$$

分類問題の例

- サポートベクトルマシン（ヒンジ損失項＋ L2 正則化）

$$\begin{aligned} J(\mathbf{w}) &= \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) + \frac{\lambda}{2} \langle \mathbf{w}, \mathbf{w} \rangle \\ &= C \sum_{i=1}^{n} \max(0, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) + \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle \end{aligned}$$

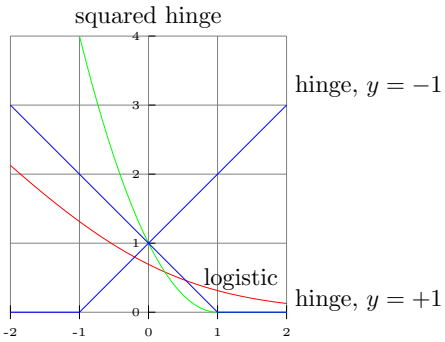

図 3　二値分類問題の損失関数

- L_1-ロジスティック回帰（ロジスティック損失項＋L1 正則化）

$$
\begin{aligned}
&J(\mathbf{w}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp\left(-y_i\left\langle\mathbf{w},\mathbf{x}_i\right\rangle\right)\right)+\lambda\sum_{j=1}^{d}|w_j| \\
&= C\sum_{i=1}^{n}\log\left(1+\exp\left(-y_i\left\langle\mathbf{w},\mathbf{x}_i\right\rangle\right)\right)+\sum_{j=1}^{d}|w_j|
\end{aligned}
$$

4.3　多クラス分類

$y\in[K]=\{1,2,\dots,K\}$ のとき K-クラス分類問題という。$\mathbf{w}_k$ をパラメータベクトルとする。

- 識別関数

$$
f(\mathbf{x})=\operatorname*{argmax}_{k\in[K]}\left\langle\mathbf{w}_k,\mathbf{x}\right\rangle
$$

- 多クラスヒンジ損失関数

$$
\begin{aligned}
\ell(\mathbf{x},y)&=\max\left(0,\max_{k\in[K]\setminus\{y\}}1-\left(\left\langle\mathbf{w}_y,\mathbf{x}\right\rangle-\left\langle\mathbf{w}_k,\mathbf{x}\right\rangle\right)\right)\\
&=\max_{k\in[K]}[\![y=k]\!]-\left(\left\langle\mathbf{w}_y,\mathbf{x}\right\rangle-\left\langle\mathbf{w}_k,\mathbf{x}\right\rangle\right)
\end{aligned}
$$

- 多クラスロジスティック損失関数

$$
\ell(\mathbf{x},y)=\log\left(\sum_{k=1}^{K}\exp\left(\left\langle\mathbf{w}_k,\mathbf{x}\right\rangle\right)\right)-\left\langle\mathbf{w}_y,\mathbf{x}\right\rangle
$$

5　カーネル法

導入として回帰問題における次のようなモデルを考える。どのように係数をえらべばよいか？

$$
\begin{aligned}
f(x)&=\alpha x^2+\beta x+\gamma\\
&=\left\langle\underbrace{\begin{bmatrix}\alpha\\\beta\\\gamma\end{bmatrix}}_{\mathbf{w}},\underbrace{\begin{bmatrix}x^2\\x\\1\end{bmatrix}}_{\boldsymbol{\phi}(x)}\right\rangle
\end{aligned}
$$

一般的に $\boldsymbol{\phi}(x)$ のことを特徴写像と呼ぶ。回帰だけでなく以下の問題で以下の議論が成立する。

$$
J(\mathbf{w})=\sum_i\ell_i\left(\left\langle\mathbf{w},\boldsymbol{\phi}(x_i)\right\rangle,y\right)+\frac{\lambda}{2}\left\langle\mathbf{w},\mathbf{w}\right\rangle
$$

- 表現定理
  微分が 0 を考える：

$$
\begin{aligned}
&\sum_i\ell_i'(\left\langle\mathbf{w}^*,\boldsymbol{\phi}(x_i)\right\rangle,y)\boldsymbol{\phi}(x)+\lambda\mathbf{w}^*=0\\
&\Rightarrow\mathbf{w}^*=-\lambda^{-1}\sum_i\ell_i'(\left\langle\mathbf{w}^*,\boldsymbol{\phi}(x_i)\right\rangle,y)\boldsymbol{\phi}(x)
\end{aligned}
$$

  よってある $(\alpha_i^*)_{i\in[n]}$ により

$$
\mathbf{w}^*=\sum_{i=1}^{n}\alpha_i^*\boldsymbol{\phi}(\mathbf{x}_i)
$$

  と書ける。したがって次の最小化問題を解けば十分

$$
\begin{aligned}
&\operatorname*{minimize}_{\alpha_i}J\left(\sum_i\alpha_i\boldsymbol{\phi}(x)\right)\\
\Leftrightarrow&\operatorname*{minimize}_{\alpha_i}\sum_i\ell_i\left(\left\langle\sum_{i'}\alpha_{i'}\boldsymbol{\phi}(x_{i'}),x_i\right\rangle,y\right)\\
&\qquad+\frac{\lambda}{2}\left\langle\sum_i\alpha_i\boldsymbol{\phi}(x_i),\sum_i\alpha_i\boldsymbol{\phi}(x_i)\right\rangle
\end{aligned}
$$

  (i,i') 要素が

$$
K_{ii'}=\left\langle\boldsymbol{\phi}(x_i),\boldsymbol{\phi}(x_{i'})\right\rangle
$$

  である行列 K を考える。すると目的関数は

$$
\operatorname*{minimize}_{\alpha_i}\sum_i\ell_i\left(\mathbf{e}_i^\top K\boldsymbol{\alpha},y\right)+\frac{\lambda}{2}\boldsymbol{\alpha}^\top K\boldsymbol{\alpha}
$$

  と書ける。
  予測器は $\left\langle\mathbf{w}^*,\mathbf{x}\right\rangle=\left\langle\sum_i\alpha_i^*\boldsymbol{\phi}(x_i),\boldsymbol{\phi}(x)\right\rangle$ という形になる。
- カーネル関数

  カーネル関数 $k(x,x')=\left\langle\boldsymbol{\phi}(x_i),\boldsymbol{\phi}(x)\right\rangle$ を考えると特徴写像がわからなくても学習、識別が可能。

$$
\begin{aligned}
K_{ii'}&=k(x_i,x_{i'})\\
\left\langle\sum_i\alpha_i\boldsymbol{\phi}(x_i),\boldsymbol{\phi}(x)\right\rangle&=\sum_i\alpha_ik(x_i,x)
\end{aligned}
$$

- カーネル関数の例
  - 線形カーネル

$$
k(\mathbf{x}_i,\mathbf{x}_{i'})=\left\langle\mathbf{x}_i,\mathbf{x}_{i'}\right\rangle=\sum_j x_{ij}x_{i'j}
$$

– 多項式カーネル

$$k(\mathbf{x}_i, \mathbf{x}_{i'}) = (\gamma \langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle + r)^d$$

– ガウスカーネル/RBF カーネル

$$k(\mathbf{x}_i, \mathbf{x}_{i'}) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2\right)$$

- Mercer の定理よりカーネル関数の正当性がわかる

> 任意の n,$(\mathbf{x}_i)_{i\in[n]}$ において
>
> $$K_{ii'} = k(\mathbf{x}_i, \mathbf{x}_{i'})$$
>
> となる行列 K が常に半正定値となる時、
>
> $$k(\mathbf{x}_i, \mathbf{x}_{i'}) = \langle \boldsymbol{\phi}(x_i), \boldsymbol{\phi}(x_{i'}) \rangle$$
>
> となる適当な特徴写像（と適当な関数空間）が存在する。

6 分類問題の指標

- 正解率/Accuracy

$$\frac{\#\{i|\hat{y}_i = y_i\}}{n}$$

- 精度/Precision

$$\frac{\#\{i|\hat{y}_i = +1, y_i = +1\}}{\#\{i|\hat{y}_i = +1\}}$$

- 再現率/Recall

$$\frac{\#\{i|\hat{y}_i = +1, y_i = +1\}}{\#\{i|y_i = +1\}}$$

- ROC 曲線/Receiver Operating Characteristic curve
  横軸：
  $$\frac{\#\{i|\hat{y}_i = +1, y_i = +1\}}{\#\{i|y_i = +1\}}$$
  縦軸
  $$\frac{\#\{i|\hat{y}_i = +1, y_i = -1\}}{\#\{i|y_i = -1\}}$$

- Precision-Recall curve
  横軸：は再現率。縦軸は精度