機械学習とデータマイニングの基礎（大阪大学）

学習問題の具体例

Matthew J. Holland

大阪大学 産業科学研究所

目次

目次

分布の位置推定（作図用）

問題の定式化

データの形式

$\mathbb{R}$ 上の未知の確率分布 μ から n 個の標本 $\mathsf{X}_1, \ldots, \mathsf{X}_n$ を学習データとする．
（独立同分布と仮定）

意思決定の候補

$\theta \in \mathbb{R}$ を選ぶ．

最終目的

$\mathsf{X} \sim \mu$ からの平均二乗誤差を最小にしたい．

$$\min_{\theta \in \mathbb{R}} \mathrm{R}(\theta), \quad \mathrm{R}(\theta) \coloneqq \mathbf{E}_\mu (\theta - \mathsf{X})^2 \tag{1}$$

ただし $\mathsf{X} \sim \mu$ は確率的な評価用データ．

学習法のアイディア

目的関数における未知の μ に代わって $\mathsf{X}_1, \ldots, \mathsf{X}_n$ の経験分布で解いてみよう．

$$\bar{\mathsf{X}}_n \coloneqq \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (\theta - \mathsf{X}_i)^2 \tag{2}$$

$$= \frac{1}{n} \sum_{i=1}^n \mathsf{X}_i \tag{3}$$

二乗誤差なので，計算は至って簡単．

学習法の評価

注意点

固定した θ に対しては $\mathrm{R}(\theta)$ は単なる実数だが，$\mathrm{R}(\bar{\mathsf{X}}_n)$ はランダムである．

比較基準

$$\mathrm{R}^* \coloneqq \inf_{\theta \in \mathbb{R}} \mathbf{E}_\mu (\theta - \mathsf{X})^2 = \mathbf{E}_\mu (\mathbf{E}_\mu[\mathsf{X}] - \mathsf{X})^2 = \mathrm{var}_\mu \mathsf{X} \tag{4}$$

目的関数値

$$\mathrm{R}(\bar{\mathsf{X}}_n) = \mathbf{E}_\mu \left(\bar{\mathsf{X}}_n - \mathsf{X} \right)^2 \tag{5}$$

$$= \left(\bar{\mathsf{X}}_n - \mathbf{E}_\mu \mathsf{X} \right)^2 + \mathrm{var}_\mu \mathsf{X} \tag{6}$$

学習法の評価 1

平均的な挙動

前掲の $\mathrm{R}(\bar{\mathsf{X}}_n)$ と R^* を踏まえて

$$\mathbf{E}\left[\mathrm{R}(\bar{\mathsf{X}}_n) - \mathrm{R}^* \right] = \mathbf{E} \left(\bar{\mathsf{X}}_n - \mathbf{E}_\mu \mathsf{X} \right)^2 \tag{7}$$

$$= \mathbf{E} \left(\frac{1}{n} \sum_{i=1}^n (\mathsf{X}_i - \mathbf{E}_\mu \mathsf{X}) \right)^2 \tag{8}$$

$$= \frac{1}{n^2} \sum_{i=1}^n \mathbf{E} (\mathsf{X}_i - \mathbf{E}_\mu \mathsf{X})^2 \tag{9}$$

$$= \frac{\mathrm{var}_\mu \mathsf{X}}{n} \tag{10}$$

という，ある種の「性能保証」を得る．
この差異の分布についてほかに言えることは？

学習法の評価 2

Markov の不等式を使う

まず，任意の実数 $\varepsilon > 0$ について，以下が成り立つ．

$$\varepsilon \, \mathbf{1}\{\mathrm{R}(\overline{\mathsf{X}}_n) - \mathrm{R}^* > \varepsilon\} \leq \mathrm{R}(\overline{\mathsf{X}}_n) - \mathrm{R}^* .$$

期待値をとってから ε で割る．

$$\mathbf{P}\left\{\mathrm{R}(\overline{\mathsf{X}}_n) - \mathrm{R}^* > \varepsilon\right\} \leq \frac{1}{\varepsilon} \mathbf{E}\left[\mathrm{R}(\overline{\mathsf{X}}_n) - \mathrm{R}^*\right] = \frac{\mathrm{var}_\mu \mathsf{X}}{n\varepsilon}$$

整理整頓

$$n \geq \frac{\mathrm{var}_\mu \mathsf{X}}{\varepsilon} \left(\frac{1}{\delta}\right) \implies \mathbf{P}\left\{\mathrm{R}(\overline{\mathsf{X}}_n) - \mathrm{R}^* > \varepsilon\right\} \leq \delta \qquad (11)$$

目次

二値分類（作図用）

問題の定式化

データの形式

- 学習データ：$(\mathsf{X}_1, \mathsf{Y}_1), \ldots, (\mathsf{X}_n, \mathsf{Y}_n)$
- 評価データ：(X, Y)
- 入力は $\mathsf{X}_i \in \mathbb{R}^d$ とする．
- ラベルは $\mathsf{Y}_i \in \{-1, +1\}$ とする．

意思決定の候補

- スコア関数 $s : \mathbb{R}^d \to \mathbb{R}$（許容されるスコア関数の集合を $\mathcal{S}$ と記す）
- 分類は $x \mapsto \mathrm{sign}(s(x))$ で行う．
- 候補の全体：$\mathcal{H} := \{\mathrm{sign}(s(\cdot)) : s \in \mathcal{S}\}$

最終目的

誤答してしまう確率をできるだけ下げたい．

$$\min_{s \in \mathcal{S}} \mathrm{R}(s), \quad \mathrm{R}(s) := \mathbf{P}\{\mathrm{sign}(s(\mathsf{X})) \neq \mathsf{Y}\} \qquad (12)$$

攻略法の方向性

愚直に攻める場合

経験分布で近似することはもちろんできるが，

$$\mathsf{R}_n(s) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\mathrm{sign}(s(\mathsf{X}_i)) \neq \mathsf{Y}_i\right\} \tag{13}$$

という目的関数は使い勝手が悪い．

（勾配情報は無意味，凸性はない，連続性もない，…）

攻略法の方向性

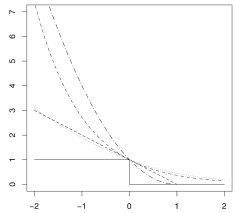

図: Bartlett et al. [1] の Fig. 1 から抜粋．

攻略法の方向性

緩和してみよう

都合の悪い目的関数を上から抑えてくれる好都合の関数を導入する．

$$\ell(s; x, y) \coloneqq \log_2\left(1 + \mathrm{e}^{-s(x)y}\right)$$

新しい目的関数の案

$$\bar{\mathsf{L}}_n(s) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell(s; \mathsf{X}_i, \mathsf{Y}_i) \tag{14}$$

わかっていること

- $\mathsf{Y}s(\mathsf{X}) > 0 \iff \mathrm{sign}(s(\mathsf{X})) = \mathsf{Y}$
- $\bar{\mathsf{L}}_n(s) \geq \mathsf{R}_n(s)$

手法と解析

緩和後の解

$$\mathsf{S}_n \in \arg\min_{s \in \mathcal{S}} \bar{\mathsf{L}}_n(s) \tag{15}$$

計算量を考慮する

時間 T をかけて入手できる解を $\mathsf{S}_n^{(T)}$ と表記する． (16)

不等式を活用する

$$\begin{aligned}\mathrm{R}(\mathsf{S}_n^{(T)}) &= \mathsf{R}_n(\mathsf{S}_n^{(T)}) + \mathrm{R}(\mathsf{S}_n^{(T)}) - \mathsf{R}_n(\mathsf{S}_n^{(T)}) \\ &\leq \mathsf{R}_n(\mathsf{S}_n^{(T)}) + \sup_s |\mathsf{R}_n(s) - \mathrm{R}(s)| \\ &\leq \bar{\mathsf{L}}_n(\mathsf{S}_n^{(T)}) + \sup_s |\mathsf{R}_n(s) - \mathrm{R}(s)|\end{aligned}$$

性能評価を「最適化の良し悪し」と「統計的な推定精度」に分解できそうである．

目次

実数値を当てつづけるゲーム（作図用）

問題の定式化

データの形式
実数値 $x_1, x_2, \ldots, x_T$ を学習データとする．

意思決定の制約
以下の手順を繰り返す（$t = 1, 2, \ldots, T$）．

- $s_t \in [u, v] \subset \mathbb{R}$ を選ぶ
- $x_t \in [u, v] \subset \mathbb{R}$ を与えられる
- 罰則 $(s_t - x_t)^2$ を食らう

最終目的
罰則の合計を最小限に抑えるように $\mathbf{s}_T := (s_1, \ldots, s_T)$ をうまく決めたい．

$$\min_{\mathbf{s}_T} \mathrm{R}_T(\mathbf{s}_T), \quad \mathrm{R}_T(\theta_1, \ldots, \theta_T) := \sum_{t=1}^{T} (\theta_t - x_t)^2 \tag{17}$$

学習法の評価

最良の s_T は自明
当然，$s_1 = x_1, \ldots, s_T = x_T$ ならば，$\mathrm{R}_T(\mathbf{s}_T) = 0$．あまりヒントにはならない…

中間的な比較基準を導入
全データが既知の場合，最良の値を一つだけ選ぶ．

$$\begin{aligned} s_T^* &:= \operatorname*{arg\,min}_{u \le s \le v} \mathrm{R}_T(s, \ldots, s) \\ &= \frac{1}{T} \sum_{t=1}^{T} x_t \end{aligned} \tag{18}$$

この「理想値」なら，これを過去のデータから近似する案は一応浮かんでくる．

学習法のアイディア

直感

時点 $t > 1$ では R_t はわからないが，R_{t-1} は既知である．これを活用しよう．

過去の理想値を追いかける

$$\begin{aligned} s_t &= \operatorname*{arg\,min}_{u \le s \le v} \mathrm{R}_{t-1}(s, \ldots, s) \\ &= \frac{1}{t-1} \sum_{i=1}^{t-1} x_i \end{aligned} \tag{19}$$

何となく前掲の (18) に近いようだが，その差異をもっと正確に突き止めたい．

学習法の評価（再び）1

表記を整える

$$\bar{x}_t \coloneqq \frac{1}{t} \sum_{i=1}^{t} x_i, \quad t \in [T] \tag{20}$$

初期値として，任意の $\bar{x}_0 \coloneqq s_1 \in [u, v]$ を選ぶ．

比較対象の整理

- 前掲の (19) に従って $\mathbf{s}_T = (\bar{x}_1, \bar{x}_1, \ldots, \bar{x}_{T-1})$ とする．
- 比較対象は $\mathbf{s}_T^* = (s_T^*, \ldots, s_T^*) = (\bar{x}_T, \ldots, \bar{x}_T)$ である．

学習法の評価（再び）2

罰則の合計を比較する

$$\begin{aligned} \mathrm{R}_T(\mathbf{s}_T) - \mathrm{R}_T(\mathbf{s}_T^*) &= \sum_{t=1}^{T} \left[(\bar{x}_{t-1} - x_t)^2 - (\bar{x}_T - x_t)^2 \right] \\ &\le \sum_{t=1}^{T} \left[(\bar{x}_{t-1} - x_t)^2 - (\bar{x}_t - x_t)^2 \right] \end{aligned} \tag{21}$$

ここで，(21) は以下の事実から導かれる．

$$\mathrm{R}_T(\bar{x}_1, \ldots, \bar{x}_T) \le \mathrm{R}_T(\bar{x}_T, \ldots, \bar{x}_T) \tag{22}$$

この (22) は二乗誤差 $(\cdot)^2$ に限らず，実は一般的な知見である．

（証明の詳細は本資料の付録を参照すること）

学習法の評価（再び）3

(21) の右辺の t 番目の項を見て，二乗誤差の性質を利用していく．

$$\begin{aligned} (\bar{x}_{t-1} - x_t)^2 - (\bar{x}_t - x_t)^2 &= \bar{x}_{t-1}^2 - 2x_t(\bar{x}_{t-1} - \bar{x}_t) - \bar{x}_t^2 \\ &= \bar{x}_{t-1}(\bar{x}_{t-1} - \bar{x}_t) - 2x_t(\bar{x}_{t-1} - \bar{x}_t) + \bar{x}_t(\bar{x}_{t-1} - \bar{x}_t) \\ &\le |\bar{x}_{t-1} - 2x_t + \bar{x}_t||\bar{x}_{t-1} - \bar{x}_t| \\ &\le 2(v-u)|\bar{x}_{t-1} - \bar{x}_t| \\ &= 2(v-u) \left| \left(\frac{1}{t-1} - \frac{1}{t} \right) \sum_{i=1}^{t-1} x_i - \frac{x_t}{t} \right| \\ &\le 2(v-u) \left[\left| \frac{1}{t(t-1)} \sum_{i=1}^{t-1} x_i \right| + \frac{|x_t|}{t} \right] \end{aligned}$$

学習法の評価（再び） 4

両端をつなぎ合わせると，以下の不等式を得る．

$$(\bar{x}_{t-1} - x_t)^2 - (\bar{x}_t - x_t)^2 \leq 2(v-u)\left[\frac{|v-u|}{t} + \frac{|v-u|}{t}\right] = \frac{2(v-u)^2}{t}$$

この不等式を (21) に当てはめると，以下の上界を得る．

$$\begin{aligned}\sum_{t=1}^{T}\left[(\bar{x}_{t-1} - x_t)^2 - (\bar{x}_t - x_t)^2\right] &\leq 2(v-u)^2 \sum_{t=1}^{T}\frac{1}{t} \\ &= 2(v-u)^2\left[1 + \sum_{t=2}^{T}\frac{1}{t}\right] \\ &\leq 2(v-u)^2\left[1 + \int_1^T (1/t)\ \mathrm{d}t\right] \\ &= 2(v-u)^2\left[1 + \log(T)\right]\end{aligned}$$

整理整頓

単純な学習法でも成り立つ保証

$$\mathrm{R}_T(\mathbf{s}_T) - \mathrm{R}_T(\mathbf{s}_T^*) \leq 2(v-u)^2\left[1 + \log(T)\right] \tag{23}$$

仮定の強弱

- どんな学習データでも構わない（敵対的な相手でも OK）．
- 区間 $[u, v]$ が無限大になってしまうと，別の仮定が必要になる．

目次

まとめ

ここで典型的な「教師あり学習」と「オンライン学習」と言える問題例を見てきた．

注目して欲しい要素

- 意思決定の候補の良し悪しに数値的な罰則を与えている
- 学習法のアウトプットを評価する方法

次は，これらの具体例を含む学習問題の体系を整備し，その相違点を論じていく．

参考文献

[1] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

学習法の評価（再び）

(22) の証明

$\ell(s;x) = (s-x)^2$ とおいて，(18) を以下のように拡張する．

$$s_t^* := \arg\min_{s\in\mathbb{R}} \mathrm{R}_t(s,\ldots,s) = \bar{x}_t \tag{24}$$

この表記のもと，まず，以下のことは自明である．

$$\ell(s_1^*;x_1) \le \ell(s_1^*;x_1) \tag{25}$$

学習法の評価（再び）

(22) の証明（続き）

次に，$T>1$ について，以下の不等式が成り立つと仮定しておく．

$$\sum_{t=1}^{T-1} \ell(\bar{x}_t;x_t) \le \sum_{t=1}^{T-1} \ell(\bar{x}_{T-1};x_t) \tag{26}$$

この (26) から導き出したいのは (22)，つまり以下の不等式

$$\sum_{t=1}^{T} \ell(\bar{x}_t;x_t) \le \sum_{t=1}^{T} \ell(\bar{x}_T;x_t) \tag{27}$$

上記の左辺と右辺それぞれの T 番目の項は共通なので，差し引ける．

$$(27) \iff \sum_{t=1}^{T-1} \ell(\bar{x}_t;x_t) \le \sum_{t=1}^{T-1} \ell(\bar{x}_T;x_t) \tag{28}$$

学習法の評価（再び）

(22) の証明（続き）

嬉しいことに，(28) の右側の不等式は (26) によって簡単に証明できる．

$$\sum_{t=1}^{T-1} \ell(\bar{x}_t;x_t) \le \sum_{t=1}^{T-1} \ell(\bar{x}_{T-1};x_t) \le \sum_{t=1}^{T-1} \ell(\bar{x}_T;x_t)$$

つまり，

- (26) から (27) が証明できた．
- $T=2$ の場合は (25) によって既に証明済み．
- 任意の T に対して (27) が成り立つ（数学的帰納法）．

(22) の証明終わり．