機械学習とデータマイニング　レポート課題

28G23027 川原尚己

課題 A

[問 1]以下の(1)~(3)式にて省略されている計算手順を明記すること.

$$\underset{\theta\in\mathbb{R}}{\arg\min}\frac{1}{n}\sum_{i=1}^{n}(\theta-X_i)^2=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{1}$$

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

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)^{\mathbf{2}}=\frac{\mathrm{var}_\mu X}{n} \tag{3}$$

[問 2]式(4)の主張が任意の$\epsilon>0$と$\delta>0$に対して成り立つことを示すこと.

$$\boldsymbol{n}\geq\frac{\mathrm{var}_\mu X}{\epsilon}\left(\frac{1}{\delta}\right)\Rightarrow\mathbf{P}\{\mathrm{R}(\bar{X}_\mathrm{n})-\mathrm{R}^*>\epsilon\}\leq\delta \tag{4}$$

回答

[問 1]

$f(\theta)$を式(A1)のように定義する.

$$f(\theta):=\frac{1}{n}\sum_{i=1}^{n}(\theta-X_i)^2 \tag{$A1$}$$

このとき，$\frac{df}{d\theta}$は式(A2)のように表される.

$$\frac{df}{d\theta}=\frac{1}{n}\left(2n\theta-\sum_{i=1}^{n}X_i\right) \tag{$A2$}$$

$\frac{df}{d\theta}=0$となる$\theta=\theta^*$は式(A3)のようになる.

$$\theta^*=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{$A3$}$$

$f(\theta)$は下に凸な放物線であるから，$\frac{df}{d\theta}=0$となるθにおいて最小値をとる．以上より，式(1)が導出された.

次に式(2)を示す．$X_i\sim\mu$であるから，$\mathbf{E}[\mathrm{X}_i]=\mathbf{E}_\mu\mathrm{X}$が成り立つ．式(2)の左辺を直接変形することにより，以下の等式を得る.

$$\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)\right)^2=\frac{1}{n^2}\mathbf{E}\left(\sum_{i=1}^{n}\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)\right)^2$$

$$=\mathbf{E}\left(\sum_{i=1}^{n}\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)^2+2\sum_{i=1}^{n}\sum_{j>i}\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)\left(X_j-\mathbf{E}_\mu \mathrm{X}\right)\right) \qquad (A4)$$

$X_i, X_j (i \neq j)$は独立であるから，$\mathbf{E}\left(\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)\left(X_j-\mathbf{E}_\mu \mathrm{X}\right)\right)=\mathbf{E}\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)\mathbf{E}\left(X_j-\mathbf{E}_\mu \mathrm{X}\right)=0$が成り立つ．よって，式(A4)の第二項は0に等しく，式(2)が得られる．

最後に式(3)を示す．

$$\mathbf{E}\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)^{\mathbf{2}}=\mathbf{E}\left(X_i^2-2X_i\,\mathbf{E}_\mu \mathrm{X}+\left(\mathbf{E}_\mu \mathrm{X}\right)^2\right)$$

$$=\mathbf{E}_\mu\left(X_i^{\mathbf{2}}\right)-\left(\mathbf{E}_\mu \mathrm{X}\right)^2$$

$$=\mathbf{E}_\mu\left(\mathrm{X}^{\mathbf{2}}\right)-\left(\mathbf{E}_\mu \mathrm{X}\right)^2$$

であり，$\mathbf{E}\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)^{\mathbf{2}}$は分布$\mu$に対しての期待値で表せるから式(A5)が成り立つ．

$$\mathbf{E}\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)^{\mathbf{2}}=\mathbf{E}_\mu\left(X-\mathbf{E}_\mu \mathrm{X}\right)^{\mathbf{2}} \qquad (A5)$$

以上より，式(3)の左辺は式(A6)のように表される．

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}\left(X_i-\mathbf{E}_\mu \mathrm{X}\right)^{\mathbf{2}}=\frac{1}{n}\mathbf{E}_\mu\left(X-\mathbf{E}_\mu \mathrm{X}\right)^{\mathbf{2}} \qquad (A6)$$

「学習問題の具体例」の式(4)より，$\mathbf{E}_\mu\left(X-\mathbf{E}_\mu \mathrm{X}\right)^{\mathbf{2}}=\mathrm{var}_\mu X$であるから，式(3)が得られる．

[問２]

「学習問題の具体例」の 8 ページにおいて，$\mathrm{R}(\bar{X}_n)\geq 0$であり，$\mathrm{R}^*$は$\mathrm{R}(\bar{X}_n)$の下限であるから，$\mathrm{R}(\bar{X}_n)-\mathrm{R}^*\geq 0$が成り立つ．よって，Markov の不等式より式(A7)がなりたつ．

$$\mathbf{P}\{\mathrm{R}(\bar{X}_\mathrm{n})-\mathrm{R}^*>\epsilon\}\leq\frac{1}{\epsilon}\mathbf{E}[\mathrm{R}(\bar{\mathrm{X}}_n)-\mathrm{R}^*] \qquad (A7)$$

「学習問題の具体例」の 7 ページより$\mathbf{E}[\mathrm{R}(\bar{\mathrm{X}}_n)-\mathrm{R}^*]=\frac{\mathrm{var}_\mu X}{n}$が成り立つから，式(A8)が得られる．

$$\mathbf{P}\{\mathrm{R}(\bar{X}_\mathrm{n})-\mathrm{R}^*>\epsilon\}\leq\frac{\mathrm{var}_\mu X}{n\epsilon} \qquad (A8)$$

このとき，任意の$\delta>0$に対し，$n\geq\frac{\mathrm{var}_\mu X}{\epsilon}\left(\frac{1}{\delta}\right)$なる整数$n$が存在する．そのような$n$をとるとき，式(A8)より式(A9)が成り立つ．

$$\mathbf{P}\{\mathrm{R}(\bar{X}_\mathrm{n}) - \mathrm{R}^* > \epsilon\} \leq \delta \tag{A9}$$

以上より，式(4)が$\epsilon > 0$と$\delta > 0$に対して成り立つことを示せた．

課題 B

資料中の「分布の位置推定」の話に出てくる学習アルゴリズムはどのような一致性を満たすか．「学習法の一致性」の定義に出てくる$\mathrm{R}, \mathrm{R}^*_\mathrm{con}, \mathrm{H}_n$はそれぞれ何に相当し，なぜ一致性が約束できるか示すこと．

回答

「分布の位置推定」の話での学習アルゴリズムにおいては$\mathrm{R}, \mathrm{R}^*_\mathrm{con}, \mathrm{H}_n$はそれぞれ以下の事柄に相当する．

R：$\mathrm{X} \sim \mu$からの平均二乗誤差

R^*_{con}：$\mathrm{X} \sim \mu$からの平均二乗誤差の下限

H_n：nこの標本の平均

このとき，$\mathrm{R}, \mathrm{R}^*_\mathrm{con}, \mathrm{H}_n$は式(B1)~(B3)のように表される．

$$R(\theta) = \mathbf{E}_\mu(\theta - \mathrm{X})^\mathbf{2} \tag{B1}$$

$$\mathrm{R}^*_\mathrm{con} = \inf_{\boldsymbol{\theta} \in \mathbb{R}} R(\theta) \tag{B2}$$

$$\mathrm{H}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{B3}$$

この$\mathrm{R}, \mathrm{R}^*_\mathrm{con}, \mathrm{H}_n$に対して式(A8)が成り立つから，$n \to \infty$に対する極限を考えることで一致性を満たすことが示せた．

課題 C

式(5)の導出過程を明記せよ．

$$N^*_{\epsilon,\delta} \leq \frac{\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)}{\epsilon} \tag{5}$$

回答

式(C1)のように PAC 条件を考える．

$$\mathbf{P}(\mathrm{R}(\mathrm{H_n}) - \mathrm{R}^*_\mathrm{con} > \epsilon) \leq \delta \tag{C1}$$

$\mathrm{R}^*_{con} = 0$，$\mathrm{R}(h) = \mathbf{P}\{h(\mathrm{X}) \neq \mathrm{Y}\}$,「サル真似」における$n+1$回目の更新の出力を$\mathrm{H}^*_{con}$として式(C1)に代入すると，式(C2)を得る．

$$\mathbf{P}\{\mathbf{P}\{\mathrm{H}^{con}_n(\mathrm{X}) \neq \mathrm{Y}\} > \epsilon\} \leq \delta \tag{C2}$$

$\mathbf{P}\{\mathbf{P}\{\mathrm{H}^{con}_n(\mathrm{X}) \neq \mathrm{Y}\} > \epsilon\}$について評価を行うために，以下のような集合$\mathcal{H}_{bad}$を考える．

$$\mathcal{H}_{bad}(\epsilon) := \{h \in \mathcal{H} \colon \mathbf{P}\{h(\mathrm{X}) \neq \mathrm{Y}\} > \epsilon\} \tag{C3}$$

H^{con}_nが$\mathcal{H}_{bad}(\epsilon)$に含まれる確率は全ての$X_i$を正しく分類できる$h \in \mathcal{H}_{bad}(\epsilon)$よりも小さいこ

とから，式(C4)を得る．

$$P\{H_n^{con} \in \mathcal{H}_{bad}(\epsilon)\} \le |\mathcal{H}_{bad}(\epsilon)|(1-\epsilon)^n \qquad (C4)$$

さらに，$\mathrm{H}_n^{con}(\mathrm{X}) \neq \mathrm{Y}$をとる確率が$\epsilon$より大きくなる確率が$P\{H_n^{con} \in \mathcal{H}_{bad}(\epsilon)\}$より小さいことから，式(C5)を得る．

$$\begin{aligned}\mathbf{P}\{\mathbf{P}\{\mathrm{H}_n^{con}(\mathrm{X}) \neq \mathrm{Y}\} > \epsilon\} &\le P\{H_n^{con} \in \mathcal{H}_{bad}(\epsilon)\} \\ &\le |\mathcal{H}|\exp(-n\epsilon) \qquad (C5)\end{aligned}$$

式(C1)及び式(C5)より，式(C6)を満たす任意のnで PAC 条件を満たすことがわかる．

$$\begin{aligned}|\mathcal{H}|\exp(-n\epsilon) &\le \delta \\ \Leftrightarrow n &\ge \frac{\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)}{\epsilon} \qquad (C6)\end{aligned}$$

以上より，標本複雑度$N_{\epsilon,\delta}^*$は PAC 条件を満たすnの最小値であるから，式(5)を満たす．

課題 D

二値分類の場合，99%の確率で$h^*(\mathrm{X}) = \mathrm{Y}$，残り1%の確率でラベルが反転して$h^*(\mathrm{X}) \neq \mathrm{Y}$ となるような確率ノイズ Uを設計し，YとXとUとh^*を含む上記のような等式を明記した上で，その設計の妥当性を説明せよ．

回答

以下のような確率ノイズUを考える．

$$U(r) = \begin{cases} 1, & (0 \le r < 0.99) \\ -1, & (0.99 \le r < 1) \end{cases} \qquad (D1)$$

ただしrは区間[0,1)の一様分布に従う確率変数である．
このとき，$\mathrm{Y} = \mathrm{h}^*(X)U(r)$という$Y$を考えると，99% の確率で正しいラベルが出力され，1% の確率でラベルが反転して出力されるため，題意を満たす設計であり，妥当性がある．

課題 E

以下の 3 つの不等式が成り立つことを図や数式などを用いて説明せよ．

１．実数直線上の有限区間において

$$\mathrm{shatter}(3) < 2^3 \qquad (6)$$

２．二次元平面における識別線において

$$\mathrm{shatter}(4) < 2^4 \qquad (7)$$

３．$\mathbb{R}^d$における長方形において

$$\mathrm{shatter}(2d+1) < 2^{2d+1} \qquad (8)$$

回答

任意のnに対して，式(E1)が成り立つ．

$$\mathrm{shatter}(n) \leq 2^n \tag{E1}$$

式(E1)の統合が成立するのは，任意の二値ラベルの組み合わせでも入力データを分類可能な分類面が存在するときである．そのため，今回の課題においては分類不可能な入力データと二値ラベルの組を少なくとも一つ例示すればよい．

１．
実数数直線上において，図１のような３点を分類できる有限区間は存在しないため，式(6)は成立する．

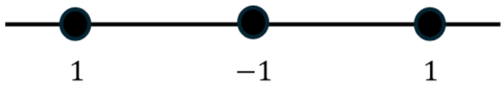

図１　実数数直線上において分類できない例

２．
二次元平面において，図２のような４点を分類できる識別線は存在しないため，式(7)は成立する．

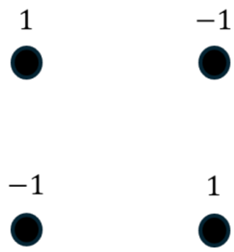

図２　二次元平面上において分類できない例

３．
どの$d+1$点も$(d-1)$次元長平面上にない場合を考えればよい．
$(2d+1)$この点のうち，ある軸において最小値または最大値であるような点は丁度$2d$個存在する．これら$2d$個の点を頂点とする領域Dを考える．Dは，d個の軸に対する最小値または最大値をすべて含んでいるため，残りの一点は必ずDに含まれることになる．Dの頂点であるようなすべての点に同一のラベルを，残りの一点にもう一方のラベルを割り当てることを考えると，このような分類は実現できない．よって，式(8)は成立する．

課題 F

以下の等式（9）と (10) がなぜ成立するか，具体的に説明すること．

$$\mathbf{E}[\mathrm{R}(\mathrm{H_T}) - \mathrm{R}(h^*)] = \mathbf{E}(\mathbf{E}[\mathrm{R}(\mathrm{H_T}) - \mathrm{R}(h^*)|\boldsymbol{Z}_T]) \tag{9}$$

$$\mathbf{E}(\mathbf{E}[\mathrm{R}(\mathrm{H_T}) - \mathrm{R}(h^*)|\boldsymbol{Z}_T]) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{E}[\mathrm{R}(\mathrm{H_t}) - \mathrm{R}(h^*)] \tag{10}$$

回答

式(9)の左辺は分布μから得られた独立標本$Z_1, \ldots, Z_T$を用いて分類器$H_1, \ldots, H_T$を学習する．その後，Uniform$\{1, \ldots, T\}$よりランダムに選択したTに対し，H_Tを用いた場合の汎化性能の期待値を表している．一方で，式(9)の右辺は学習データを固定した場合において得られたH_Tの汎化性能の期待値に対し，さらに学習データ全体において期待値をとったものである．任意の学習データは分布μに従って生成されることから，学習データを固定して学習を行ったかどうかにかかわらず，学習によって得られた分類器H_Tの汎化性能の期待値は両者の間で等しくなる．

また，式(10)の右辺のZ_Tは$\mathrm{R}(H_T) - \mathrm{R}(h^*)$とは独立した事象であるから，式(F1)のように表せる．

$$\mathbf{E}(\mathbf{E}[\mathrm{R}(\mathrm{H_T}) - \mathrm{R}(h^*)|\boldsymbol{Z}_T]) = \sum_{Z_T}\mathbf{E}[\mathrm{R}(\mathrm{H_T}) - \mathrm{R}(h^*)]P(Z_T)$$

$$= \sum_{Z_T}\sum_{t=1}^{T}\mathbf{E}[\mathrm{R}(\mathrm{H_t}) - \mathrm{R}(h^*)]P(Z_T) \tag{$F1$}$$

学習データとしてZ_Tが選ばれる確率は$\frac{1}{T}$であるから，最終的に(F2)式が得られる．

$$\mathbf{E}(\mathbf{E}[\mathrm{R}(\mathrm{H_T}) - \mathrm{R}(h^*)|\boldsymbol{Z}_T]) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{E}[\mathrm{R}(\mathrm{H_t}) - \mathrm{R}(h^*)] \tag{$F2$}$$

以上より，式(9)及び式(10)が成立することが示された．