Data Science and Machine Learning

照屋 佑喜仁

June 11, 2025

1. 2.4 Tradeoffs in Statical Learning
   - 教師あり学習の技術
   - Tradeoff
   - irreducible risk, approxiation error, statistical error
   - approximation-estimation tradeoff
   - excercise2
   - exercise3

教師あり学習

- 教師あり学習における機械学習の技術
    - generalization risk(2.5) あるいは expected generalization risk(2.6) をできるだけ小さくする
    - できるだけ少ない計算リソースで
- これを達成するために，適切な予測関数の集合 $\mathcal{G}$ を選ぶ必要がある．この選び方は下のような要因によって決まる．
    - 集合の複雑さ (最適な予測関数 g^* を適切に近似，あるいは含むのに十分に複雑 (豊か) か？)
    - (2.4) の最適化によって学習者を訓練する容易さ
    - 集合 $\mathcal{G}$ において，training loss(2.3) が risk(2.1) をどれだけ正確に推定するか
    - 連続なのか，分類なのか……

Tradeoff

- 集合 $\mathcal{G}$ の選択は，通常トレードオフを伴う
  - 単純な $\mathcal{G}$ からの学習器は早く訓練できるが，上手く近似できない可能性
  - g^* を含むような豊かな $\mathcal{G}$ からの学習器は多くの計算リソースを必要とする可能性
- モデルの複雑さ，計算の単純さ，推定の制度の関係を見るために 2 つの tradeoff について考えていく
  - the approxiation-estimation tradeoff(近似-推定トレードオフ)
  - the bias-variance tradeoff(バイアス-分散トレードオフ)

今，generalization risk(2.5) を 3 つの要素に分解して考える．

$$\ell(g_{\mathcal{T}}^{\mathcal{G}}) = \underbrace{\ell^*}_{\text{irreducible risk}} + \underbrace{\ell(g^{\mathcal{G}}) - \ell^*}_{\text{approximation error}} + \underbrace{\ell(g_{\mathcal{T}}^{\mathcal{G}}) - \ell(g^{\mathcal{G}})}_{\text{statistical error}} \tag{2.16}$$

irreducible risk, approximation error

$$\ell(g_\tau^{\mathcal{G}}) = \underbrace{\ell^*}_{\text{irreducible risk}} + \underbrace{\ell(g^{\mathcal{G}}) - \ell^*}_{\text{approximation error}} + \underbrace{\ell(g_\tau^{\mathcal{G}}) - \ell(g^{\mathcal{G}})}_{\text{statistical error}} \tag{2.16}$$

- ℓ^* は $\ell(g^*)$ で定義される irreducible risk(還元不能リスク)．どの学習器も ℓ^* より小さいリスクで予測することはできない．
- $g^{\mathcal{G}}$ は $\operatorname{argmin}_{g\in\mathcal{G}} \ell(g)$ で定義される，$\mathcal{G}$ 内で最も最良の学習器．
- $\ell(g^{\mathcal{G}}) - \ell^*$ は approximation error(近似誤差)．irreducible risk と $\mathcal{G}$ のなかで最良の予測関数の risk の差を見ている．
  - 適切な $\mathcal{G}$ を選び，その上で $\ell(g)$ を最小化するのは，単純に数値解析と関数解析の問題となる (ここで訓練データ τ は登場しないから)
  - $\mathcal{G}$ が g^* を含まなければ近似誤差は任意に小さく出来ず risk を大きくする要因となる
  - 近似誤差を減らす唯一の方法は，$\mathcal{G}$ を大きくしてより多くの関数を含めること

statical (estimation) error

$$\ell(g_\tau^{\mathcal{G}}) = \underbrace{\ell^*}_{\text{irreducible risk}} + \underbrace{\ell(g^{\mathcal{G}}) - \ell^*}_{\text{approximation error}} + \underbrace{\ell(g_\tau^{\mathcal{G}}) - \ell(g^{\mathcal{G}})}_{\text{statistical error}} \tag{2.16}$$

- $\ell(g_\tau^{\mathcal{G}}) - \ell(g^{\mathcal{G}})$ は statistical(estimation) error(統計的 (推定) 誤差)．訓練セット τ に依存．特に，学習器 $g_\tau^{\mathcal{G}}$ が $\mathcal{G}$ の最良の予測関数 $g^{\mathcal{G}}$ をどれだけ上手く推定しているかに依存している．
  (良い予測器なら) この誤差は訓練サイズが無限大に近づくにつれて (確率的に，または期待値として)0 に収束するはずである．

approximation-estimation tradeoff

approximation-estimation tradeoff(近似-推定トレードオフ) は，２つの相反する要求を対立させる．

- $\mathcal{G}$ が十分にシンプルで，統計的誤差が大きくなりすぎない必要がある．（推定しやすい？）
- $\mathcal{G}$ が十分に充実して，近似誤差が小さいことを保証する必要がある．（g^* をできれば見つけたい？）

2 乗誤差損失でのリスクを解釈してみる

2 乗誤差損失のリスクは $\ell(g_{\mathcal{T}}^{\mathcal{G}}) = \mathbb{E}\left[(Y - g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{X}))^2\right]$ となる．このとき，最適な予測関数は $g^*(\boldsymbol{x}) = \mathbb{E}\left[Y \mid \boldsymbol{X} = \boldsymbol{x}\right]$ で与えられるのであった．(theorem2.1)
このとき，分解 (2.16) は以下のように解釈できる．

- $\ell^* = \mathbb{E}\left[(Y - g^*(\boldsymbol{X}))^2\right]$ は還元不能誤差であり，これより小さい期待 2 乗誤差の予測関数はない．
- 近似誤差 $\ell(g^{\mathcal{G}}) - \ell(g^*)$ は $\mathbb{E}\left[g^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X})^2\right]$ に等しい (excrcise2)．つまり，最適予測値と $\mathcal{G}$ 内の最適予測値との間の 2 乗誤差の期待値として解釈できる．

2 乗誤差損失でのリスクを解釈してみる

- 統計的誤差 $\ell(g^{\mathcal{G}}_{\mathcal{T}}) - \ell(g^{\mathcal{G}})$ に関しては $\mathcal{G}$ が線形関数の集合である場合を除いて，期待 2 乗誤差 (mean squared error 平均 2 乗誤差ともいう) としての直接的な解釈は存在しない．
  $\mathcal{G}$ が線形関数集合の場合，関数 $g(\boldsymbol{x})$ はあるベクトル $\boldsymbol{\beta}$ に対して $g(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ と表すことができ，統計的誤差は $\mathbb{E}\left[(g^{\mathcal{G}}_{\mathcal{T}}(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X}))^2\right]$ となる (excercise3).

以上より，2 乗誤差損失を用いる場合，線形関数集合 $\mathcal{G}$ に対する generalization risk は

$$
\begin{aligned}
\ell(g^{\mathcal{G}}_{\mathcal{T}}) &= \mathbb{E}\left[(g^{\mathcal{G}}_{\mathcal{T}}(\boldsymbol{X}) - Y)^2\right] \\
&= \ell^* + \underbrace{\mathbb{E}\left[g^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X})^2\right]}_{\text{近似誤差}} + \underbrace{\mathbb{E}\left[(g^{\mathcal{G}}_{\mathcal{T}}(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X}))^2\right]}_{\text{統計的誤差}}
\end{aligned}
$$

統計的誤差だけが訓練データに依存する唯一の項であることに注意.

excercise2

保留していた証明を行う．まず,

$$\ell(g^{\mathcal{G}}) - \ell(g^*) = \mathbb{E}\left[(g^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X}))^2\right]$$

を示す.

Proof.

$$\begin{aligned}
\ell(g^{\mathcal{G}}) &= \mathbb{E}\left[(Y - g^{\mathcal{G}}(\boldsymbol{X}))^2\right] \quad \text{(2 乗誤差を採用したときの定義)} \\
&= \mathbb{E}\left[\left\{Y - g^*(\boldsymbol{X}) + g^*(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X})\right\}^2\right] \\
&= \underbrace{\mathbb{E}\left[\left\{Y - g^*(\boldsymbol{X})\right\}^2\right]}_{\ell(g^*)\text{ の定義}} + \mathbb{E}\left[\left\{g^*(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X})\right\}^2\right] \\
&\quad - \underbrace{2\mathbb{E}\left[\left\{Y - g^*(\boldsymbol{X})\right\}\left\{g^*(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X})\right\}\right]}_{\bigstar}
\end{aligned}$$

excercise2

Proof.

ここで，前々回の Theorem2.1 の証明と同様の議論により $\bigstar = 0$ となるので，

$$\ell(g^{\mathcal{G}}) - \ell(g^*) = \mathbb{E}\left[\left(g^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X})\right)^2\right]$$

を得る．□

Remark (tower property, taking out what is known)

tower property:

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$$

taking out what is known:

$$\mathbb{E}[XY|X] = X\mathbb{E}[Y|X]$$

参考 (前々回のスライド)

Proof.

定義より,

$$
\begin{aligned}
\mathbb{E}[Y - g^*(\boldsymbol{X})|\boldsymbol{X}] &= \int_{\mathbb{R}} (y - g^*(\boldsymbol{x}))f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})dy \\
&= \int_{\mathbb{R}} yf_{Y|\boldsymbol{X}}(y|\boldsymbol{x})dy - \int_{\mathbb{R}} g^*(\boldsymbol{x})f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})dy \\
&= \mathbb{E}[Y|\boldsymbol{X}] - \mathbb{E}[Y|X] \int_{\mathbb{R}} \frac{f(x,y)}{f_X(x)}dy \\
&= 0 \\
\because \int_{\mathbb{R}} \frac{f(x,y)}{f_X(x)}dy &= \frac{1}{f_X(x)} \int_{\mathbb{R}} f(x,y)dy \\
&= \frac{1}{f_X(x)} f_X(x) = 1
\end{aligned}
$$

excercise3

次に以下を示す.

$$\ell(g_{\mathcal{T}}^{\mathcal{G}}) - \ell(g^{\mathcal{G}}) = \mathbb{E}\left[(g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X}))^2\right]$$

Proof.

$$\begin{aligned}
\ell(g_{\mathcal{T}}^{\mathcal{G}}) &= \mathbb{E}\left[\left\{Y - g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{X})\right\}^2\right] \\
&= \mathbb{E}\left[\left\{Y - g^*(\boldsymbol{X}) + g^*(\boldsymbol{X}) - g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{X})\right\}^2\right] \\
&= \ell(g^*) + \underbrace{\mathbb{E}\left[\left\{g^*(\boldsymbol{X}) - g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{X})\right\}^2\right]}_{\clubsuit}
\end{aligned}$$

excercise3

Proof.

ここで,

$$
\begin{aligned}
\clubsuit &= \mathbb{E}\left[\left\{g^*(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X}) + g^{\mathcal{G}}(\boldsymbol{X}) - g^{\mathcal{G}}_{\mathcal{T}}(\boldsymbol{X})\right\}^2\right] \\
&= \mathbb{E}\left[\left\{g^*(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X})\right\}^2\right] + \mathbb{E}\left[\left\{g^{\mathcal{G}}(\boldsymbol{X}) - g^{\mathcal{G}}_{\mathcal{T}}(\boldsymbol{X})\right\}^2\right] \\
&\quad - 2\mathbb{E}\left[\left\{g^*(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X})\right\}\left\{g^{\mathcal{G}}(\boldsymbol{X}) - g^{\mathcal{G}}_{\mathcal{T}}(\boldsymbol{X})\right\}\right]
\end{aligned}
$$

□