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

2. 2.5 Bias-Variance Tradeoff
   - 多項式回帰の例（続き）
   - バイアス-分散分解

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．

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

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

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

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，

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

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,

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

$$\begin{aligned}
\because \int_{\mathbb{R}} \frac{f(x,y)}{f_X(x)} dy &= \frac{1}{f_X(x)} \int_{\mathbb{R}} f(x,y) dy \\
&= \frac{1}{f_X(x)} f_X(x) = 1
\end{aligned}$$

excercise3

次に以下を示す.

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

Proof.

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

excercise3

Proof.

ここで,

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

excercise3

Proof.

$$\boldsymbol{\beta}^{\mathcal{G}} = \mathrm{argmin}_{\boldsymbol{\beta}} \mathbb{E}\left[\left\{g^*(\boldsymbol{X}) - \boldsymbol{\beta}^T \boldsymbol{X}\right\}^2\right]$$

であるため，$\frac{\partial}{\partial \boldsymbol{x}} g(f(\boldsymbol{x})) = \frac{\partial f}{\partial \boldsymbol{x}} \frac{\partial g}{\partial \boldsymbol{y}}$ や微分と期待値の交換 (中間値の定理，優収束定理などで示せる) などに気をつければ，

$$\left.\frac{\partial}{\partial \boldsymbol{\beta}} \mathbb{E}\left[\left\{g^*(\boldsymbol{X}) - \boldsymbol{\beta}^T \boldsymbol{X}\right\}^2\right]\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\mathcal{G}}} = \mathbf{0}$$

$$\therefore \quad 2\mathbb{E}\left[X\left\{\boldsymbol{X}^T \boldsymbol{\beta}^{\mathcal{G}}\right\}\right] = \mathbf{0}$$

よって $\spadesuit = 0$ となり，exercise2 の結果と合わせて
$\ell(g_\tau^{\mathcal{G}}) - \ell(g^{\mathcal{G}}) = \mathbb{E}\left[(g_\tau^{\mathcal{G}}(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X}))^2\right]$ が成立. □

Example 2.2 多項式回帰の例

$\mathcal{G} = \mathcal{G}_p$ は x の線形関数の集合で，$\boldsymbol{x} \in [1, u, u^2, ..., u^{p-1}]^T$ であり，
$g^*(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}^*$ である．
また，$\boldsymbol{X} = \boldsymbol{x}$ で，$Y = g^*(X) + \varepsilon(\boldsymbol{x})$，$\varepsilon(\boldsymbol{x}) \sim \mathcal{N}(0, \ell^*)$，
$\ell^* = \mathbb{E}\left[\{Y - g^*(\boldsymbol{X})\}\right]$ とする．
まず，近似誤差を考える．任意の関数 $g \in \mathcal{G}_p$ は

$$g(x) = h(u) = \beta_1 + \beta_2 u + \cdots + \beta_p u^{p-1} = [1, u, ..., u^{p-1}]^T \boldsymbol{\beta}$$

このとき $g(X)$ は $[1, U, ..., U^{p-1}]\boldsymbol{\beta}$ として分布し，$U \sim \mathcal{U}(0, 1)$ であるとする．同様に，$g^*(\boldsymbol{X})$ は $[1, U, U^2, U^3, U^4]\boldsymbol{\beta}^*$ として分布するとする．

近似誤差の計算

近似誤差は

$$\int_0^1 \left([1, U, ..., U^{p-1}]\boldsymbol{\beta} - [1, U, U^2, U^3, U^4]\boldsymbol{\beta}^*\right)^2 du$$

となる．近似誤差を最小化するため，$\boldsymbol{\beta}$ で微分して 0 になれば良い．ベクトルをベクトルで微分しており，$\boldsymbol{\beta}$ の 1 つの要素に着目して微分すれば

$$\int_0^1 ([1, u, ..., u^{p-1}]\boldsymbol{\beta} - [1, u, u^2, u^3]\boldsymbol{\beta}^*)du = 0,$$

$$\int_0^1 ([1, u, ..., u^{p-1}]\boldsymbol{\beta} - [1, u, u^2, u^3]\boldsymbol{\beta}^*)udu = 0,$$

$$\vdots$$

$$\int_0^1 ([1, u, ..., u^{p-1}]\boldsymbol{\beta} - [1, u, u^2, u^3]\boldsymbol{\beta}^*)u^{p-1}du = 0.$$

近似誤差の具体的な値

$\mathbf{H}_p = \int_0^\infty [1, u, ..., u^{p-1}]^T [1, u, ..., u^{p-1}] du$ を $p \times p$ ヒルベルト行列とする．(i, j) 要素は $\int_0^\infty u^{i+j-2} du = 1/(i + j - 1)$ で与えられる．線形方程式系は $\mathbf{H}_p \boldsymbol{\beta} = \tilde{\mathbf{H}} \boldsymbol{\beta}^*$ とまとめて表すことができる．ここで $\tilde{\mathbf{H}}$ は $\mathbf{H}_p$ の左上 $(p \times 4)$ ブロック．解は以下で表される：

$$\boldsymbol{\beta}_p = \begin{cases} \dfrac{65}{6} & (p = 1) \\ \begin{bmatrix} -\dfrac{20}{3} & 35 \end{bmatrix}^{\mathrm{T}} & (p = 2) \\ \begin{bmatrix} -\dfrac{5}{2} & 10 & 25 \end{bmatrix}^{\mathrm{T}} & (p = 3) \\ \begin{bmatrix} 10 & -140 & 400 & -250 & 0 & \cdots & 0 \end{bmatrix}^{\mathrm{T}} & (p \geq 4) \end{cases} \tag{2.18}$$

近似誤差の結果

したがって，近似誤差 $\mathbb{E}[(g^{\mathcal{G}_p}(X) - g^*(X))^2]$ は以下で与えられる：

$$\int_0^1 ([1, u, ..., u^{p-1}]\boldsymbol{\beta}_p - [1, u, u^2, u^3]\boldsymbol{\beta}^*)^2 du \simeq \begin{cases} 127.9, & p = 1, \\ 25.8, & p = 2, \\ 22.3, & p = 3, \\ 0, & p > 4. \end{cases} \tag{2.19}$$

p が増加するにつれて近似誤差が小さくなることに注目．この例では，$p = 4$ で近似誤差がゼロとなる（$p > 4$ も同様）．
一般に，近似する関数のクラス $\mathcal{G}$ がより複雑になると，近似誤差は減少する．(前に述べた)

統計的誤差

次に，統計的誤差の典型的な挙動を説明する．2 乗誤差損失に対して，統計的誤差は以下のように書ける：

$$\int_0^1 ([1, ..., u^{p-1}](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_p))^2 du = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_p)^T \mathbf{H}_p (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_p) \tag{2.20}$$

図 2.8 は，図 2.7 でテスト損失の計算に使用されたものと同じ訓練セットに対する凡化リスクの分解 (2.17) を示している．
テストロスは独立したテストデータを使用して推定されることに注意．
この場合，2 つがよく一致していることがわかる．統計的誤差の最小値は $p = 4$ 付近にあることがわかる．

図 2.8 の解釈

パラメータ数 p	リスクの構成要素
低い p	近似誤差大，統計的誤差小
高い p	近似誤差小，統計的誤差大

- 近似誤差は p が増加するにつれて $p = 4$ でゼロまで減少
- 統計的誤差は $p = 4$ 後に増加傾向
- 一般化リスクは $p = 4$ 付近で最小値

統計的誤差は推定値 $\hat{\beta}$ に依存し，これは訓練セット τ に依存することに注意．統計的誤差のより良い理解を得るため期待される挙動，すなわち多くの訓練セットでの平均的な挙動を考慮する必要がある．

バイアス-分散分解の導入

もう一度 2 乗誤差損失を使用し，一般的な $\mathcal{G}$ に対するもう一つの分解（バイアス-分散分解）を，近似誤差や統計的誤差の組み合わせで考えていく

$$\ell(g_\tau^{\mathcal{G}}) = \ell^* + \ell(g_\tau^{\mathcal{G}}) - \ell(g^*)$$

Theorem 2.1 の証明と同様の推論を使用すると：

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

ここで $D(\boldsymbol{x}, \tau) := g_\tau^{\mathcal{G}}(X) - g^*(X)$ とする.

バイアス-分散分解の展開

ランダムな訓練セット $\mathcal{T}$ に対するランダム変数 $D(\boldsymbol{x}, \mathcal{T})$ について期待 2 乗誤差は

$$\mathbb{E}[(g_\tau^{\mathcal{G}}(x) - g^*(x))^2] = \mathbb{E}\left[D^2(\boldsymbol{x}, \tau)\right] = \mathbb{E}\left[D(\boldsymbol{x}, \tau)\right]^2 + \mathbb{V}\left[D(\boldsymbol{x}, \tau)\right]$$

$$= \underbrace{(\mathbb{E}[g_\tau^{\mathcal{G}}(x)] - g^*(x))^2}_{\text{pointwise squared bias}} + \underbrace{\mathbb{V}[g_\tau^{\mathcal{G}}(x)]}_{\text{pointwise variance}} \tag{2.21}$$

学習器 $g_\tau^{\mathcal{G}}(x)$ をランダムな訓練セットの関数と見ると：

- pointwise squared bias 項は $g_\tau^{\mathcal{G}}(x)$ が $g^*(x)$ から平均的にどれだけ離れているかを測定
- pointwise variance 項は $g_\tau^{\mathcal{G}}(x)$ の期待値からの偏差を測定

期待一般化リスクの分解

X と $\mathcal{T}$ が独立であることに注意．したがって，期待一般化リスク (2.6) は以下のように書ける：

$$\mathbb{E}[\ell(g_{\mathcal{T}}^{\mathcal{G}})] = \ell^* + \underbrace{\mathbb{E}\left[\mathbb{E}[g_{\mathcal{T}}^{\mathcal{G}}(X)|X] - g^*(X)\right]^2}_{\text{expected squared bias}} + \underbrace{\mathbb{E}\left[\mathbb{V}(g_{\mathcal{T}}^{\mathcal{G}}(X)|X)\right]}_{\text{expected variance}} \quad (2.22)$$

これがバイアス-分散トレードオフの基本的な分解である．

バイアス-分散トレードオフの解釈

バイアス-分散分解 (2.22) から：

- Expected squared bias: 学習器が真の関数から系統的にどれだけ偏っているかを表す
  - 単純なモデル（線形回帰など）は高バイアス
  - 複雑なモデルは低バイアス
- Expected variance: 異なる訓練セットに対する学習器の予測のばらつきを表す
  - 単純なモデルは低分散
  - 複雑なモデル（高次多項式など）は高分散

トレードオフ: モデルの複雑さを増すとバイアスは減るが分散は増える．最適な複雑さは両者のバランスで決まる．