1. バイアスとバリアンスについて

MSE(Mean Square Error) の式から Variance と Bias の式を導出する

母集団の真値: t，**モデルの予測値**: $y(x)$, **観測値（標本，真値＋ノイズ)**: $h(x)$
平均損失または期待損失は

$$E[L] = \int\int L(t, y(x))p(x, t)dxdt \tag{1}$$

で与えられる．回帰問題に良く使われる損失関数は二乗誤差 $L(t, y(x)) = \{y(x) - t\}^2$ である．この場合，期待損失は

$$E[L] = \int\int y(x) - t^2 p(x, t)dxdt \tag{2}$$

と書ける．目標は E[L] を最小にする y(x) を選ぶことなので，この最適解を変分法を使って求めると

$$\frac{\delta E[L]}{\delta y(x)} = 2\int \{y(x) - t\}p(x, t)dt = 0 \tag{3}$$

また，確率の加法・乗法定理を使って式変形を行う ・確率の加法定理 $P(X) = \sum_Y P(X, Y)$ ・確率の乗法定理 $P(X, Y) = P(Y|X)P(X)$ (3) 式より，

$$\begin{aligned}
\int y(x)p(x, t)dt &= \int tp(x, t)dt \\
y(x)\int p(x, t)dt &= \int tp(x, t)dt \\
y(x)p(x)dt &= \int tp(x, t)dt \\
y(x) = \frac{\int tp(x, t)dt}{p(x)} &= \int tp(t|x)dt = E[t|x]
\end{aligned} \tag{4}$$

となり，最適解 $y(x)$ は条件付き期待値になることが分かる．これは，x が与えられた下での t の条件付き平均であり，回帰関数と呼ばれる．

また，最適解 $y(x)$ が条件付き期待値なので，二乗の項を次のように展開することができる．

$$\begin{aligned}
\{y(x) - t\}^2 &= \{y(x) - E[t|x] + E[t|x] - t\}^2 \\
&= \{y(x) - E[t|x]\}^2 + \{E[t|x] - t\}^2 \\
&+ 2\{y(x) - E[t|x]\}\{E[t|x] - t\}
\end{aligned} \tag{5}$$

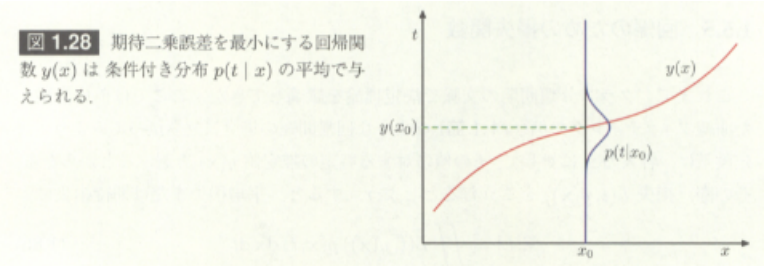

図 1　回帰関数 $y(x)$ と条件付き確率 $P(t|x)$ の関係

(5) 式を期待損失の (2) 式に代入し，t で積分するとクロス項が消え次式になる

$$\begin{aligned} E[L] &= \int\int \{y(x) - E[t|x]\}^2 p(x,t) dx dt + \int\int \{E[t|x] - t\}^2 p(x,t) dx dt \\ &= \int \{y(x) - E[t|x]\}^2 dx \int p(x,t) dt + \int\int \{E[t|x] - t\}^2 p(t|x) p(x) dx dt \\ &= \int \{y(x) - E[t|x]\}^2 p(x) dx + \int \{E[t|x] - t\}^2 p(t|x) dt \int p(x) dx \\ &= \int \{y(x) - E[t|x]\}^2 p(x) dx + \int var[t|x] p(x) dx \end{aligned} \tag{6}$$

$var[t|x]$ は x が与えられた下での t の条件付き分散で，条件付き期待値と同様に

$$var[t|x] = \int \{E[t|x] - t\}^2 p(t|x) dt$$

で表される．(6) 式で，求める関数 $y(x)$ は最初の項だけに入っており，$y(x)$ が $E[t|x]$ に等しいとき最小となり，この項は 0 になる．これは前に導いた，最適な最小二乗予測器は条件付き平均で与えられるという結果に他ならない．第 2 項は t の分散を x に関して平均したものである．これは目標データが本質的に持つ変動であり，ノイズとみなせる．これは $y(x)$ とは独立なので，損失関数の中でそれ以上減らすことのできない損失関数の最小値となる．

ここで，(4) 式の最適な予測である条件付き期待値を

$$h(x) = E[t|x] = \int t p(t|x) dt$$

と置く．また，現実には有限個のデータ D しか与えられていないため，理想的な回帰関数 $h(x)$ を厳密に求めることはできない．頻度主義的立場では，推定値の不確実性を次のように評価する．まず，分布 $p(t,x)$ に従う多数のデータ集合に従う多数のデータ集合（標本数 N で互いに独立とする）を考える．そして，与えられた任意のデータ集合 D に対して学習アルゴリズムを実行でき，予測関数 $y(x;D)$ を求められると仮定する．このとき，異なるデータ集合は異なる関数を与えるため，二乗損失の値も異なる．このような設定のもと，学習

アルゴリズムの性能は，データ集合の取り方に関する平均の意味で評価される． ここで，(6) 式の第一項の積分を考える．これは，あるデータ集合 D に対して，$\{y(x;D)-h(x)\}^2$ という形になる．この量はデータ集合 D に依存するため，この値のデータ集合の取り方に関する期待値を考える．上式の括弧内で $E_D[y(x;D)]$ を足して引き，展開すると，

$$\begin{aligned}\{y(x;D)-h(x)\}^2 &= \{y(x;D)-E_D[y(x;D)]+E_D[y(x;D)]-h(x)\}^2 \\ &= \{y(x;D)-E_D[y(x;D)]\}^2+\{E_D[y(x;D)]-h(x)\}^2 \\ &+ 2\{y(x;D)-E_D[y(x;D)]\}\{E_D[y(x;D)]-h(x)\} \end{aligned} \tag{7}$$

が得られる．そして，この式全体のデータ集合 D の取り方に関する期待値を取れば，最後の項は消える．

$$\begin{aligned}E_D[2\{y(x;D)-E_D[y(x;D)]\}\{E_D[y(x;D)]-h(x)\}] &= 2\{E_D[y(x;D)]-h(x)\}E_D[y(x;D)-E_D[y(x;D)]] \\ &= 2\{E_D[y(x;D)]-h(x)\}E_DE_D[y(x;D)-E_D[y(x;D)]] \\ &= 2\{E_D[y(x;D)]-h(x)\}E_DE_D[y(x;D)]-E_D[y(x;D)] \\ &= 0 \\ (E[A]=A, E[E[A]]=E[A],\ &when\ A\ is\ Constant)\end{aligned}$$

よって，

$$E_D[\{y(x;D)-h(x)\}^2]=\{E_D[y(x;D)]-h(x)\}^2+E_D[\{y(x;D)-E_D[y(x;D)]\}^2] \tag{8}$$

$$\begin{cases} Bias = E_D[y(x;D)]-h(x) \\ Variance = E_D[\{y(x;D)-E_D[y(x;D)]\}^2] \end{cases}$$

となる．これより，$y(x;D)$ と回帰関数 $h(x)$ の期待二乗誤差は 2 つの項の和で表されることが分かる．第一項は二乗バイアス (bias) と呼ばれ，すべてのデータ集合の取り方に関する予測値の平均が理想的な回帰関数からどのくらい離れているかを表している．第二項はバリアンス (variance) と呼ばれ，各データ集合に対する解が，特定のデータ集合の選び方に関する期待値周りでの変動の度合いを表している．したがって，第二項は関数 $y(x;D)$ の特定のデータ集合の選び方に関する敏感さを表している．

・バイアス (bias) とは、学習アルゴリズムの誤った仮定 (設計) による誤差である．高いバイアスは、特徴量とターゲットの出力間の関連性を見逃す可能性がある (underfitting)・バリアンス (variance) とは、トレーニングセットにおける小さなゆらぎ（変動）に対する感度から生じる誤差である．高いバリアンスは、アルゴリズムが意図した出力でなく、トレーニングデータのランダムノイズをモデル化する可能性がある (overfitting) Bias–variance tradeoff の wikipedia を参考

・underfitting…データの豊かさに対してモデルの自由度が小さすぎると、データの構造をとらえることが全くできない．つまり訓練誤差の値が大きすぎて、何の予測能力も得られないこと．未学習 (under learning) とも呼ばれる．・overfitting…モデルの自由度が大きすぎると、学習データのもつノイズ（統計的なゆらぎ）まで

も正確にフィッティングしてしまい、与えれた訓練データに関する訓練誤差の値はどんどん小さくなる．しかし、訓練データに適合することで未知のデータに対してどんどん予測能力を失っていく．

Q. 最小二乗法という考え方でも，等分散ガウス分布＋最尤推定法という考え方でも「二乗損失関数 $\sum_i (y_i - f(x_i))^2$ を最小にするような $f(x_i)$ を選べ」という同じ結果が得られることについて (参考) A. 最小二乗法というのは最適化手法の一つに過ぎず他の手法よりも優れていると証明できないが，データ (のノイズ) が等分散ガウス分布に従っていると仮定できるときは最小二乗法は最尤推定と同じ結果になるので，最小二乗法の正当性が示される．

参考文献

- PRML(上)
- これならわかる深層学習（入門)
- PRML の内容を分かりやすく説明しているサイト https://www.hellocybernetics.tech/entry/2017/01/24/100415
- バイアスとバリアンスについて参考にしたサイト https://nigimitama.hatenablog.jp/entry/2018/11/24/062732