

3.3 節 「機械学習の基礎と展望」

補足資料

(2024年8月30日版)

A.3.3.1 損失関数

テストデータに対する当てはまりの悪さは、**期待損失**¹⁾ (expected loss) という指標によって定量化される。一般に期待損失は、損失関数に対して、入力 の量と出力の量を確率変数とみなし、これらについて期待値をとったものである。これは、学習データだけでなく、テストデータを含む母集団から得られるデータ すべてに対する出力への当てはまりの平均的な悪さとみなすことができる。この 当てはまりの悪さである期待損失が小さくなるような機械学習モデルを、汎 化能力の高い良いモデルと考える。

ここで、損失関数として二乗損失を用いた場合の期待損失 (期待予測二乗誤 差) の性質を、少し踏み込んで見てみよう。いま、入力変数 X と出力変数 Y と の関係が、

$$Y = f(X) + \varepsilon$$

と表されているとする。ここで、 f は X と Y との真の関係を表す、未知の関数 とする。また、 ε は、 X だけでは Y を説明しきれなかった情報とし、期待値が 0、分散が σ^2 の確率変数であるとする。加えて、 ε は学習データにより学習さ れた機械学習モデル \hat{f} と独立とする。このとき、テストデータの入力 $X = x_0$ における、学習された機械学習モデル \hat{f} と出力変数 Y との期待予測二乗損失

$$E[\{Y - \hat{f}(X)\}^2 | X = x_0] \tag{A.3.1}$$

¹⁾ 期待予測誤差。リスクともよばれる。

を考える。これは、確率変数 X の実現値 x_0 が与えられた下での、確率変数 Y (または ε) に対する条件付き期待値である。

機械学習モデル \hat{f} はサンプルサイズ n の学習データに対応する出力変数 Y_1, \dots, Y_n に依存しているので、 \hat{f} も確率変数である。このことに注意して、(A.3.1) 式の期待値の中身を次のように展開する。

$$\begin{aligned} \{Y - \hat{f}(X)\}^2 &= \{f(X) + \varepsilon - E[\hat{f}(X)] + E[\hat{f}(X)] - \hat{f}(X)\}^2 \\ &= \{f(X) - E[\hat{f}(X)]\}^2 + \varepsilon^2 + \{E[\hat{f}(X)] - \hat{f}(X)\}^2 \\ &\quad + 2\varepsilon\{f(X) - E[\hat{f}(X)]\} + 2\varepsilon\{E[\hat{f}(X)] - \hat{f}(X)\} \\ &\quad + 2\{f(X) - E[\hat{f}(X)]\}\{E[\hat{f}(X)] - \hat{f}(X)\}. \end{aligned}$$

これに対して、 Y について $X = x_0$ の下での条件付き期待値をとると、最右辺の6つの項はそれぞれ次のように計算される。

- 真のモデル f は確率変数ではないため、 $\{f(X) - E[\hat{f}(X)]\}^2$ に対して $X = x_0$ の下での条件付き期待値をとると、 $\{f(x_0) - E[\hat{f}(x_0)]\}^2$ となる。
- $E[\varepsilon] = 0$ であることを利用すると、

$$E[\varepsilon^2] = E[\varepsilon^2] - E[\varepsilon]^2 = V[\varepsilon] = \sigma^2.$$

- $\{E[\hat{f}(X)] - \hat{f}(X)\}^2$ の条件付き期待値は、 $\hat{f}(X)$ の条件付分散 $V[\hat{f}(X) | X = x_0] = V[\hat{f}(x_0)]$ である。
- $f(X) - E[\hat{f}(X)]$ は $X = x_0$ の条件の下で定数なので、 $2E[\varepsilon\{f(X) - E[\hat{f}(X)]\} | X = x_0] = 2\{f(x_0) - E[\hat{f}(x_0)]\}E[\varepsilon] = 0$ 。
- $\hat{f}(X)$ と ε が独立であることから、

$$\begin{aligned} &2E[\varepsilon\{E[\hat{f}(X) | X = x_0] - \hat{f}(X)\} | X = x_0] \\ &= 2E[\varepsilon]E[E[\hat{f}(X)] - \hat{f}(X) | X = x_0] = 0. \end{aligned}$$

- 最後に、

$$\begin{aligned} &2E[\{f(X) - E[\hat{f}(X)]\}\{E[\hat{f}(X)] - \hat{f}(X)\} | X = x_0] \\ &= 2\{f(x_0) - E[\hat{f}(x_0)]\}\{E[\hat{f}(x_0)] - E[\hat{f}(x_0)]\} = 0. \end{aligned}$$

以上をまとめると、期待予測二乗損失は次で表される。

$$E\left[\{Y - \hat{f}(X)\}^2 \mid X = x_0\right] = \{E[\hat{f}(x_0)] - f(x_0)\}^2 + V[\hat{f}(x_0)] + \sigma^2. \quad (\text{A.3.2})$$

右辺の第1項は「モデルのバイアス (の2乗)」、第2項は「モデルの分散」、第3項は「観測誤差の分散」に対応する。これら3つの項目の意味を一つひとつ説明しよう。

そのために、次のような状況を考える。機械学習モデルは、母集団からの無作為抽出により観測された1つのデータセットに対して学習されるものであるが、いま、それが何セットも得られ、それぞれに対して機械学習モデルを学習したと仮定しよう。複数のデータセットはそれぞれ異なるため、データセットが100セットあれば、100個の異なる機械学習モデルが得られることになる。このとき、これら複数の機械学習モデルを平均したものと、真の関数 $f(X)$ との差、つまり偏りを計算したものが、(A.3.2)式第1項のモデルのバイアス (の2乗) になる。一方で、これら複数の機械学習モデルの散らばり具合を定量化したものが、(A.3.2)式第2項のモデルの分散である。最後に、(A.3.2)式第3項の観測誤差の分散は、観測されたデータに含まれるノイズの散らばり具合、つまり観測値そのものの精度を表すものである。はじめの2つの項であるモデルのバイアス (の2乗) と分散は、構築される機械学習モデルに応じて小さくなったり大きくなったりする。他方で、第3項は機械学習モデルに依存しないもので、どれだけ良い機械学習モデルを構築したとしても小さくすることができない。したがって、第3項を除いた、モデルのバイアスと分散、この2つを共に小さくする機械学習モデルが良いモデルとなる。

それでは、モデルのバイアスと分散が小さいのは、どのような機械学習モデルだろうか。その関係を表したのが、図 A.3.1 である。このグラフでは、横軸をモデルの複雑さ、縦軸を期待損失としている。たとえば、過度に複雑なモデル、つまり過適合したモデルに対しては、真の関数とのずれは一つひとつのモデルとは大きいかもしれないが、平均的には小さくなる傾向にある。結果として、バイアスは小さくなる。その一方でこのようなモデルは、出力の予測値のばらつきは大きくなるため、モデルの分散は大きくなる傾向にある。これは、母

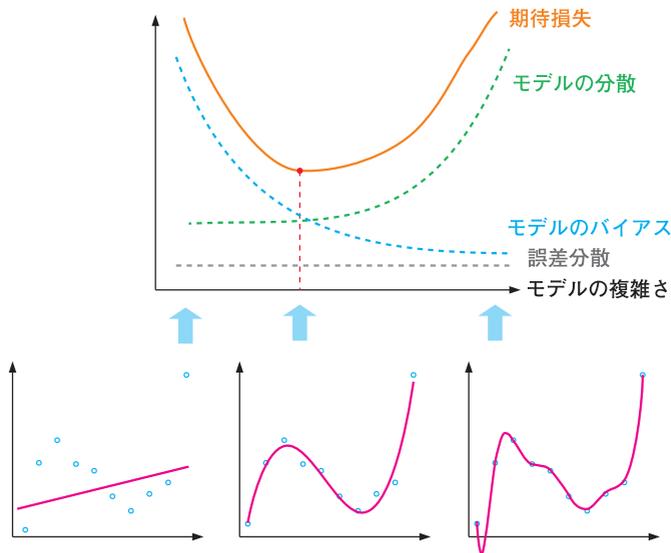


図 A.3.1 期待損失とそれを分解したもののイメージ. 上のグラフの横軸はモデルの複雑さを表しており, 左側は単純なモデルに対する期待損失, 右側は複雑なモデルに対する期待損失を表す.

集団からデータセットが得られ, それに対して機械学習モデルを学習するたびに, 大きく異なるモデルが得られていることを意味する. これは, 入力 X と出力 Y の関係を表す真の関数 f を推定しようとしていることを考えると, 望ましい性質とはいえない. 次に, 単純すぎる機械学習モデルは, X と Y の関係を十分に捉えることができず, バイアスは大きくなる. その一方で, 母集団からデータが得られるごとに学習される機械学習モデルはその単純さゆえに一つひとつが大きくばらつくことがなく, 分散は小さくなる. つまり, バイアスと分散は, 機械学習モデルの複雑さを変えてどちらかを小さくしようとするとどちらかが大きくなるというトレードオフの関係にある. 結果として, モデルの複雑さが大きすぎても小さすぎても, モデルのバイアスと分散のどちらかが大きくなり, 期待損失は大きくなってしまう. そのため, これらがバランスよく小さくなるような, つまり, モデルのバイアスと分散が総じて小さくなるような機械学習モデルを探ることが良いとされている. 図 A.3.1 では, 赤い点の位置にある複

雑さをもつモデルがこれに対応する。

A.3.3.2 RMSE と MAPE

回帰問題における予測評価指標として本文で紹介した MSE と類似した異なる評価指標として RMSE と MAPE を紹介する。

本文で紹介した**平均二乗誤差** (Mean Squared Error: MSE) は、正解のデータとモデルによる予測をそれぞれ y_i, \hat{y}_i ($i = 1, \dots, N$) としたとき、

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (\text{A.3.3})$$

と定義された。予測の評価においてもこの MSE を使っても構わないが、誤差を二乗したことでデータと単位が異なっているため、予測の評価には MSE の平方根をとった**平均二乗平方根誤差** (Root Mean Squared Error: RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} = \sqrt{\text{MSE}} \quad (\text{A.3.4})$$

が使われることが多い。

同じデータに対して複数の予測を比較するのであれば、MSE や RMSE がより小さい予測のほうが良い予測であることがわかる。しかし、われわれが許容できる予測の誤差は、予測対象のデータのスケールやちらばりの大きさによって自ずと変わってくる。そこで、MSE を正解の分散で割った値を 1 から引いた決定係数や、誤差を正解の値 y_i で割ってから平均をとる**平均絶対パーセント誤差** (Mean Absolute Percentage Error: MAPE)

$$\text{MAPE} = 100 \times \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (\text{A.3.5})$$

などが使われることもある。