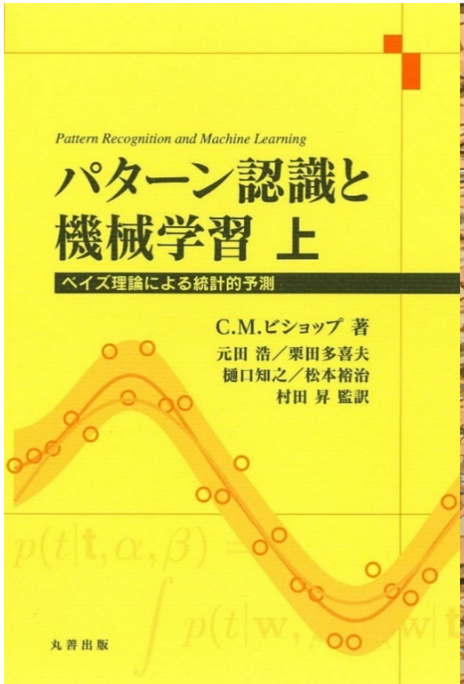
PRMLゼミ

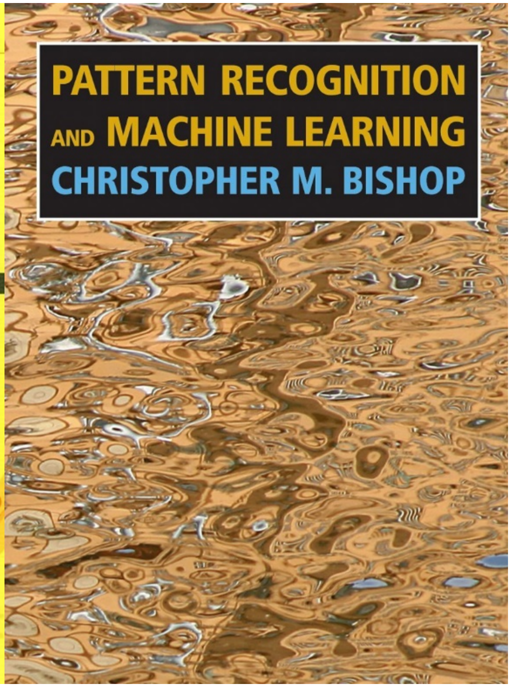
1章イントロ・1.1節・1.3節

anmitsu48

本資料について

- 本資料は、『パターン認識と機械学習　上　－ベイズ理論による統計的予測－』（丸善出版）を用いてゼミを行った際に、私が使用した発表資料を再編集したものである。
- 再編集の際は、私が持っている他の資料も利用した。
参考にした資料は最後にまとめて紹介する。



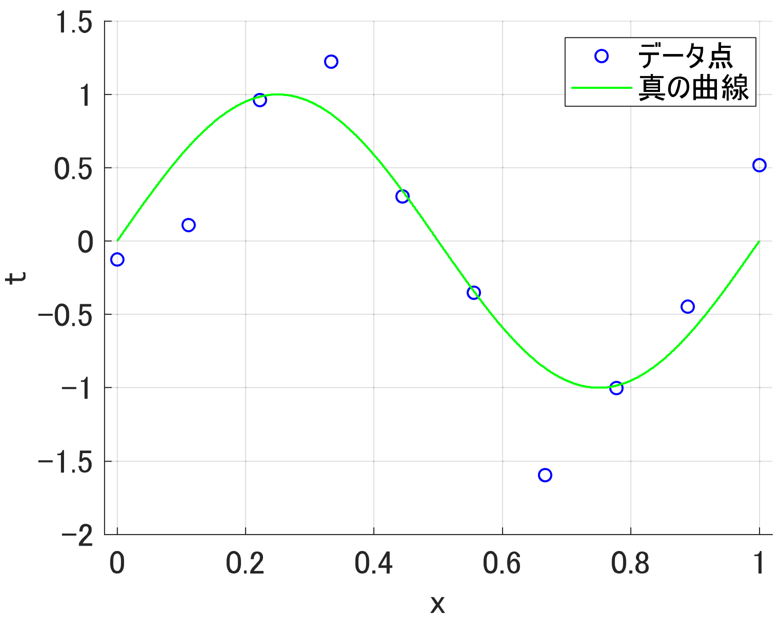


３：1.3節の紹介

- 1.3節 「モデル選択」
- PRML原著の内容を具体例とともに説明する。

1.1節を振り返る

- 訓練データ集合：N個の観測データ
  - 入力データ集合： $\boldsymbol{x} = (x_1, \ldots, x_N)^\top$
  - 目標データ集合： $\boldsymbol{t} = (t_1, \ldots, t_N)^\top$
- ゴール：訓練データ集合から、入出力の関係を推定して、新しい入力 $\hat{x}$ から目標変数 $\hat{t}$ を予測する。



- 入力データ：区間 [0, 1] から等間隔で N = 15 個の x_n を選ぶ。
- 出力データ：

$$\sin(2\pi x_n) + \varepsilon_n, \quad \varepsilon_n \sim N(0, 0.3^2)$$

ガウス分布に従うランダムノイズ

多項式曲線フィッティング

- 多項式を使って、データへのフィッティングを行う。

$$y(x,\boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

- 係数は、リッジ回帰により決定する。

$$\widetilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\Big(y(x_n,\boldsymbol{w}) - t_n\Big)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

$$\Downarrow$$

- 事前に決める必要がある多項式の次数 M や
正則化パラメータ λ は、どのように決定すればよいか？

ハイパーパラメータの決定方法

- ハイパーパラメータは、未知のデータに対して、最適な予測をするように決定すればよい。
  - 「訓練データに対する当てはまりが良い」
    ≠「未知のデータに対する当てはまりが良い」（cf. 過学習）

- 訓練データを分割し、ハイパーパラメータの各候補に対して、学習と検証を行い、最適な値を決定する。
  - 「訓練用データ」（training set）と「検証用データ」（validation set）に分割。
  - 「テストデータ」（test set）は学習と検証では利用しないことに注意。
  - MNISTなどのデータセットを使用する場合、訓練データとして提供されているデータが学習用と検証用に分けられていないことがある。このような場合は、適当に分解する必要がある。

ハイパーパラメータの決定 ①

- モデルを決定し、ハイパーパラメータの候補を用意する。
  - 今回は「多項式フィッティング＋リッジ回帰」
  - 多項式の次数の候補： M = 1,2, ⋯, 8
  - 正則化パラメータの候補： $\log \lambda = -30, -29.5, -29, \ldots, 0$

① 訓練用データを使って学習する。

- 今回、訓練用データ点 x_n は区間[0,1]を等間隔に分割する15点とし、対応する値 t_n は、$t_n = \sin(2\pi x_n) + \varepsilon_n$ で計算する。
- 検証用データ点は、区間[0,1]から上記以外の点からランダムに10個選び、対応する値を同様に計算することで取得する。

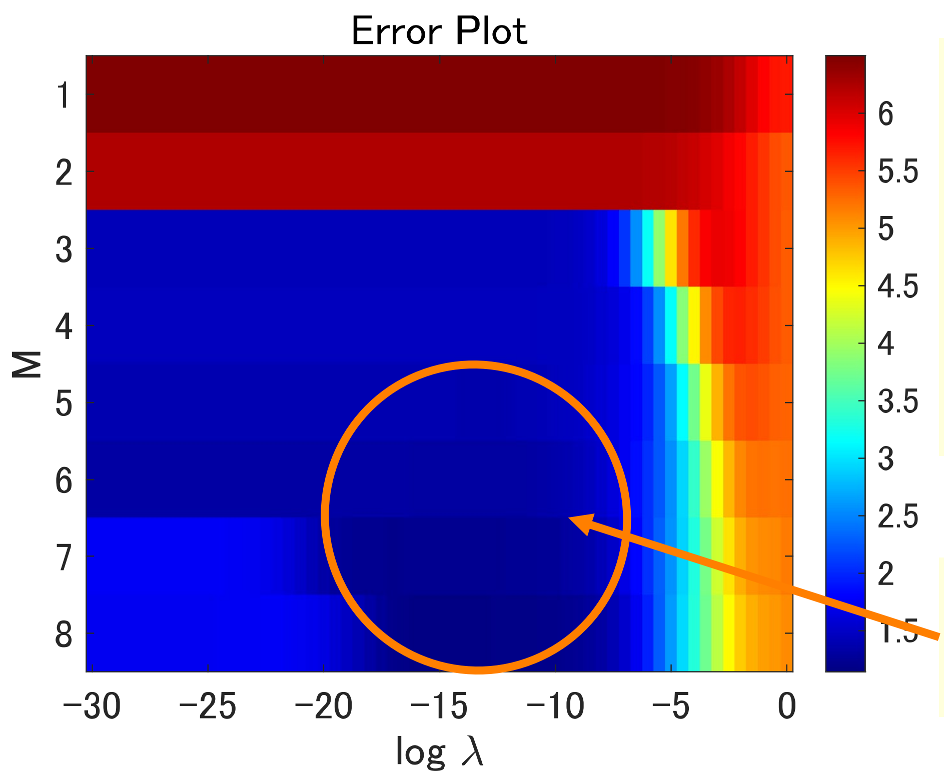
↓

各候補ごとに学習をして、検証用データに対する総二乗誤差を計算
⇒ 最適なハイパーパラメータを決定

ハイパーパラメータの決定 ②

- 各候補ごとに学習をした後に、検証用データに対する総二乗誤差を計算。

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n, \boldsymbol{w}) - t_n\right)^2$$



今回使用したデータの場合、
「次数が小さい」
or「正則化パラメータが大きい」
場合に、総二乗誤差が大きくなる。
⇒ モデルとして不適当。

今回の場合、この辺りのパラメータを使用するのが最適。

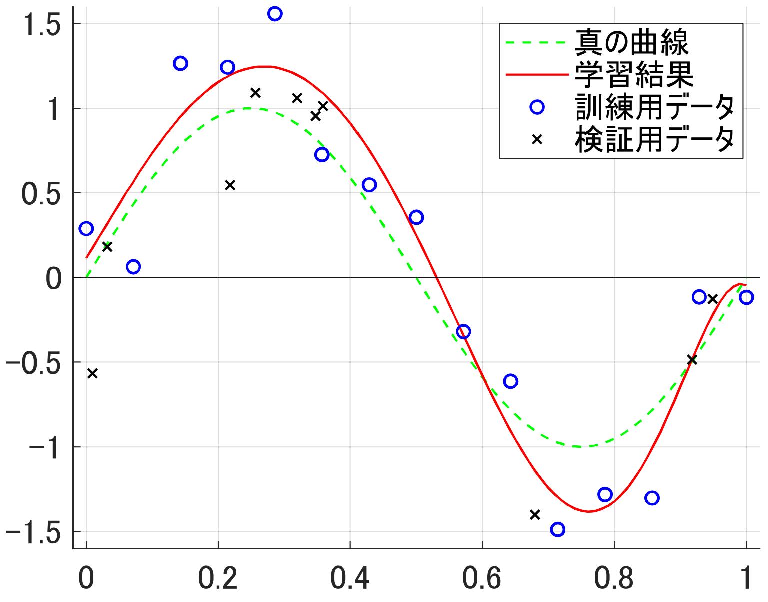
ハイパーパラメータの決定 ③

② テストデータを利用して、選択したパラメータの妥当性を確認。

- 検証用データではなく、実データやテストデータへの当てはまりが良いパラメータを採用しなければならない。
（検証用データにだけ、過剰に適合している可能性もある。）

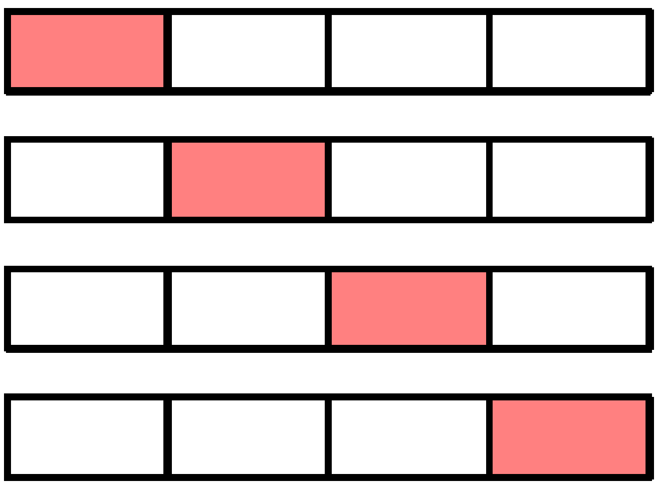
今回使用したデータでは、
$M = 8,\ \log\lambda = -15$ の場合に、
検証用データに対する
総二乗誤差が最小となった。

テストデータ（ここでは、真の曲線）への当てはまりは良くない？
⇒ 検証用データに過剰に適合していると判断してもよい。



交差確認法（Cross Validation）

- 学習に使用できる使用できるデータが少ない場合、
  できるだけ多くのデータを「訓練用」として使用するべき。
- しかし、そうすると「検証用」データが少なくなり、
  検証フェーズが意味をなさない。



交差検証法のイメージ図：
赤枠が検証用データを表す。
上図はS=4の場合を表している。
（PRML Fig1.18 より）

【交差検証法（Cross Validation）】

- データセット全体をS個に分割して、
  S−1個を訓練用データ、
  1個を検証用データとして利用する。
- これを検証用データを変えながら、S回計算して、S個の性能（誤差）の平均を計算する。
- この平均値を1つのハイパーパラメータの候補における評価値とする。

交差検証法の課題

- 使用できるデータ数が非常に少ない場合
  - S分割することで訓練用データが減ることを防ぐため、検証用データを1個とする「1個抜き法」が利用されることがある。

- 交差検証法は1つのハイパーパラメータの候補に対する評価値を得るためにS回の学習が必要となる。
  - ハイパーパラメータの数や候補数が多い場合、1回の学習に大量の時間を要する深層学習モデルの場合には、交差検証法は合わない。

情報量規準（Information Criteria）

- 複雑なモデルの過学習を避けるペナルティ項を加えて修正した指標を、モデル選択の評価に使用することがある。
  - （例）赤池情報量規準（AIC）、ベイズ情報量規準（BIC）

- 赤池情報量規準（Akaike Information Criteria, AIC）
  - $\ln p(\mathcal{D}|\boldsymbol{w}_{\mathrm{ML}})$ ：重み w が最尤推定値のときの対数尤度
  - M：調整可能なパラメータの数

$$\mathrm{AIC} = \ln p(\mathcal{D}|\boldsymbol{w}_{\mathrm{ML}}) - M$$

- AICが最大となる時のハイパーパラメータを選択
  - AICを指標とした場合、単純なモデルほど良いモデルとみなす。
  - AICの定義は同値なものがいくつか知られている。
    （例えば、「上記の式の－２倍を最小にするものを選択する」など）