「パターン認識と機械学習」
1.3 モデル選択

小野寺喜行

最小二乗法

- 最小二乗法で多項式曲線をあてはめる

  - 最も良い汎化を示した最適な次数の多項式がある
  - 次数はモデルの自由パラメータの数を制御し、複雑さを支配する。

正則化した最小二乗法

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{f(x_n,\mathbf{w}) - y_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$\|\mathbf{w}\|^2 = w_0^2 + w_1^2 + \cdots + w_M^2$$

- 正則化係数λ：モデルの実質的な複雑さを制御
- より複雑なモデルでは、複雑さを支配する複数のパラメータがあり得る

パラメータの値を決める

- 新たなデータに対して最も良い予測をするため

  ↓さらに

- 与えられたモデル内の複雑さパラメータ設定
- 異なる型のモデルも考慮

  → それぞれの応用毎に最も良いモデルを見つけたい

最尤アプローチ

- 過学習の問題あり
- 訓練集合に対する性能は未知データの予測性能のいい指標ではない

データが十分ある時の単純なアプローチ

手持ちのデータの一部を使っていろいろなモデルを学習、もしくは1つのモデルの複雑さパラメータの値を変える

独立なデータでそれらを比較

最も予測性能のいい物を選ぶ

確認用集合

- **確認用集合**（検証用集合; validation set）: 比較用のデータ
- 限られたサイズのデータ集合を使ってモデルの設計を繰り返す➡確認用集合にも過学習することがある
- テスト集合(test set)を別に用意しておいて、選んだモデルの性能を最終的に評価する必要がある

単純なアプローチの欠点

- 訓練とテストに使えるデータは限られている
- 得られたデータは出来るだけ多く訓練に使いたい
- 確認用集合が小さいと予測性能の推定の誤差が大きくなる

このジレンマの解決法

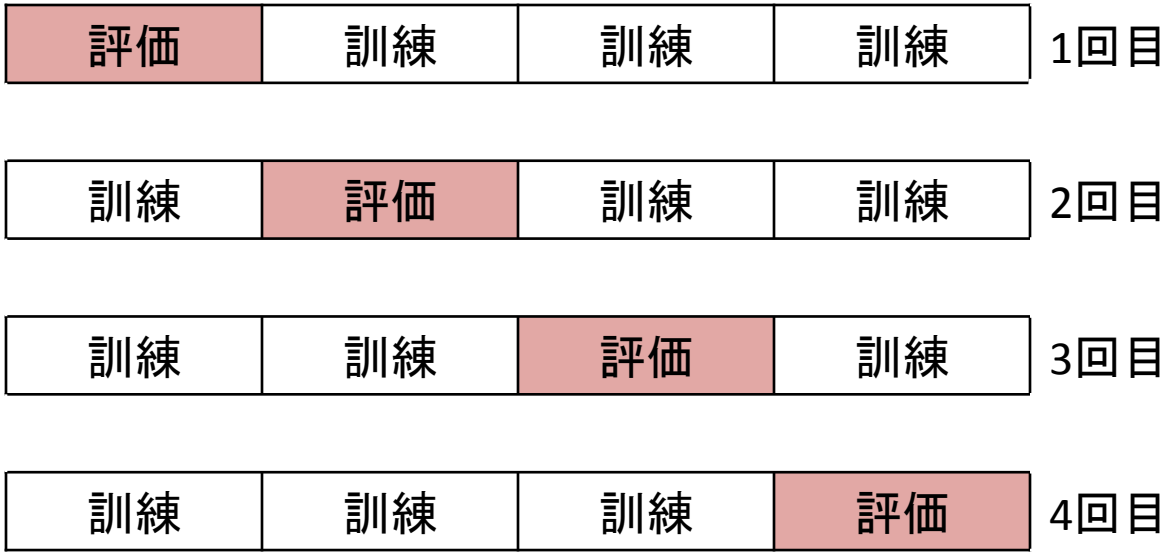
交差確認（交差検証; cross-validation）

交差確認

- 得られたデータをS個のグループに分ける
- S-1個のグループがモデル集合の訓練に使われ、残ったグループで評価を行う
- 上記手順をS個の選び方に関して繰り返し、S回の性能のスコアを平均

(S-1)/Sを訓練に使いつつ、全データを評価に使える

例)S=4のとき



この4回の性能のスコアを平均する

LOO法

- データが特に少ないとき、S=(データ点数)と考える

LOO法（1個抜き法; leave-on-out method）

交差確認の欠点

- 訓練回数がSに比例して大きくなり、訓練1回の計算量が大きくなる場合問題になる
- 単独のモデルでも複数の複雑さパラメータを持つ場合、パラメータの数に対して指数関数的に訓練回数が増える可能性がある

理想的なアプローチ

- 訓練データだけに依存
- 1回の訓練だけで複数の超パラメータとモデルのタイプを比較できるもの

訓練データだけに依存し、過学習によるバイアスを持たない性能の尺度を見つける事が必要

情報量規準

- より複雑なモデルによる過学習を避ける罰金項を足す事によって、最尤推定のバイアスを修正したい

さまざまな「情報量規準」（information criterion）と呼ばれる物が提案されて来た

赤池情報量規準 (Akaike information criterion,AIC)

- $\ln p(D \mid \mathbf{w}_{ML}) - M$
  という量が最大になるモデルを選ぶ
- $p(D \mid \mathbf{w}_{ML})$: 最尤推定を行なった場合の対数尤度
- M : モデルの中の可変パラメータの数
- 変種に**ベイズ情報量規準**（Bayesian information criterion,BIC）がある
- しかし、こうした基準はモデルパラメータの不確実性は考慮しておらず、実際は過度に単純なモデルを選ぶ傾向にある