１．３　モデル選択

05T4047H　田中　洸一

内容の手順

- モデル選択の必要性
- データ集合のアプローチ
- 交差確認法
- 交差確認法の欠点
- 情報量規準

モデル選択の必要性（1）

- モデルを設計する際、例えば多項式におけるデータ集合の最もよい汎化を示した次数を設定する。
- しかし、次数を設定するということはモデルのパラメータの数を制御し、モデルの複雑さを制限することになる。

モデル選択の必要性（2）

- 我々の目的は、新たなデータに対して最もよいモデルを見つけることである。
- よって、モデルのパラメータの適切な値を設定しつつ、異なる型のモデルも考慮して、それぞれの応用ごとに最適なモデルを見つける必要がある。

データ集合のアプローチ（1）

- モデルを設計するには訓練データを用いるが、過学習の問題があるため、訓練データに対するモデルの性能が新たなデータに対する最適なモデルの性能とはならない。
- よって、訓練データの一部を使っていろいろなモデルを設計し、それらを独立なデータと比較し、最も性能のよいものを選ぶ方法がある。

データ集合のアプローチ（２）

- しかし、比較を繰り返すことによって独立データに対しても過学習が働いてしまうため、３番目のテストデータ集合を用意し、最終的にそのデータ集合のよってモデル性能を評価する事が望ましい。

交差確認法（１）

- 少ないデータに対して、できるだけたくさんの訓練データを使う方法として交差確認法がある。
- 得られたデータのうち（Ｓ－１）/Ｓの割合部分を訓練に使い、残りのデータをテストデータとして使用する。

交差確認法(2)

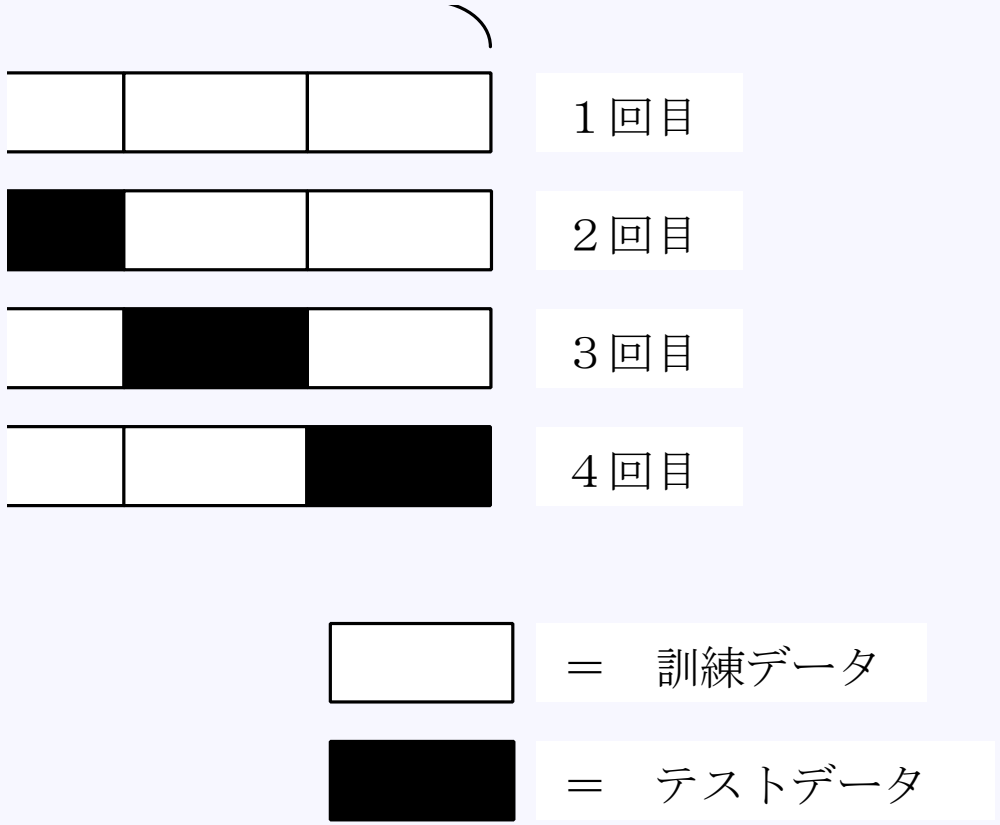

図１．Ｓ＝４の交差確認法

交差確認法の欠点（1）

- 交差確認法は、訓練を行う回数がSに比例して大きくなってしまう欠点がある。
- 一回の訓練に大きな計算量が必要な場合には、交差確認法は適した方法とはいえない。
- 理想的には、訓練データだけに依存し、一回の訓練だけで比較できるものがよい。

情報量規準（１）

- 訓練データにだけ依存し、過学習によるバイアスをもたない性能の尺度が必要。
- 「情報量規準」と呼ばれるものが存在。
- これは、過学習を避ける罰金項を足すことで、バイアスを修正しようというもの。

情報量規準（2）

- 有名なものとして赤池情報量規準（AICとも呼ばれる）がある。

$$In\,\mathrm{p}(D|\,w_{ML}) - M$$

$\mathrm{p}(D|\,w_{ML})$は最尤推定を行った場合の対数尤度

Mはモデルの中の可変パラメータの数

- この量が最大になるモデルを選ぶ。

情報量規準（3）

- その他にも、BIC（ベイズ情報量規準）がある。