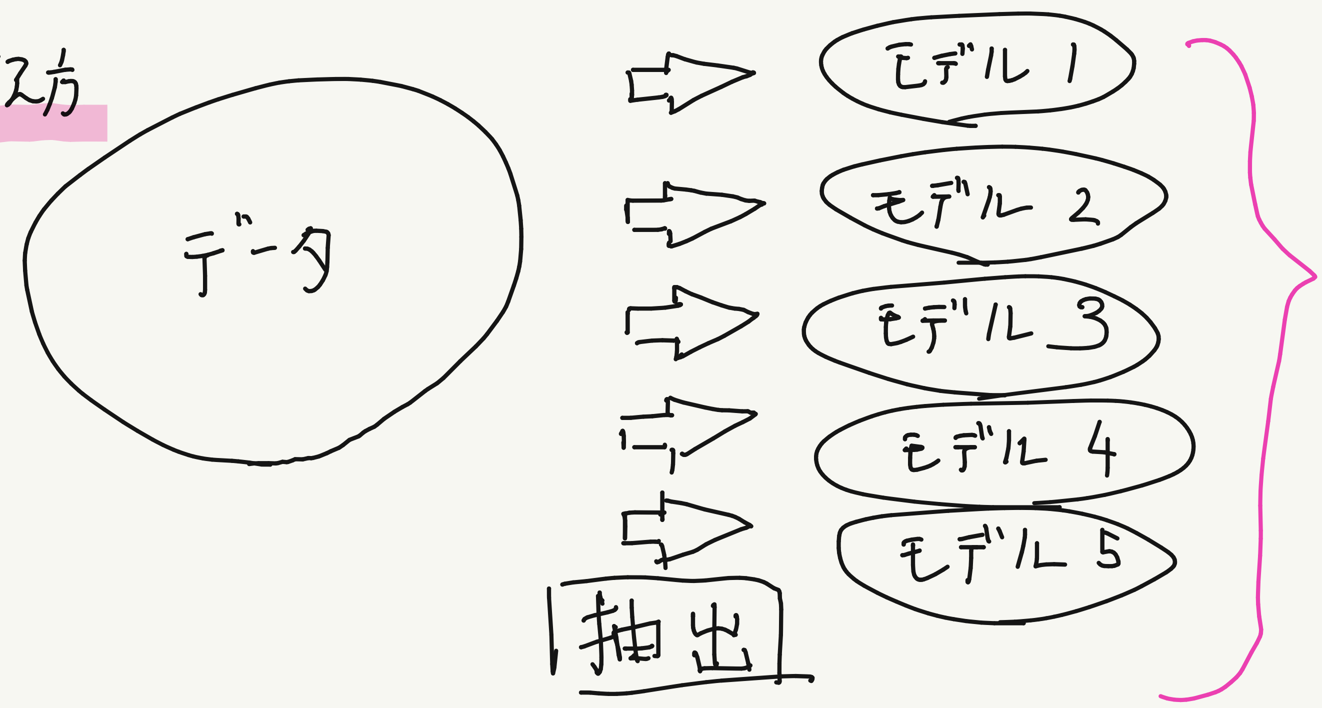
第五章 リサンプリング法

考え方



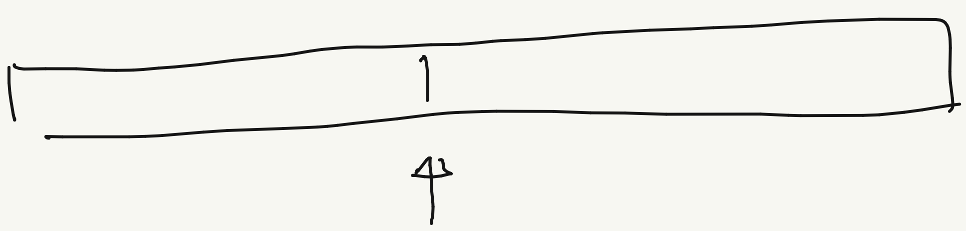
結果についての違いを検証

何か情報を得られる可能性！

① 交差検証

用途
- 方法論のテスト誤差推定
- 適切な柔軟さの設定

● ホールドアウト検証



半分にランダムで分割し、
片方を訓練、片方をテストデータとして使用

※ただし欠点がある。

① どのようにデータが分割されるかによって、テスト誤差の推定値にかなりばらつきがある。

② データが分割されることから、訓練データを用いた際のモデル予測精度が低下してしまう。

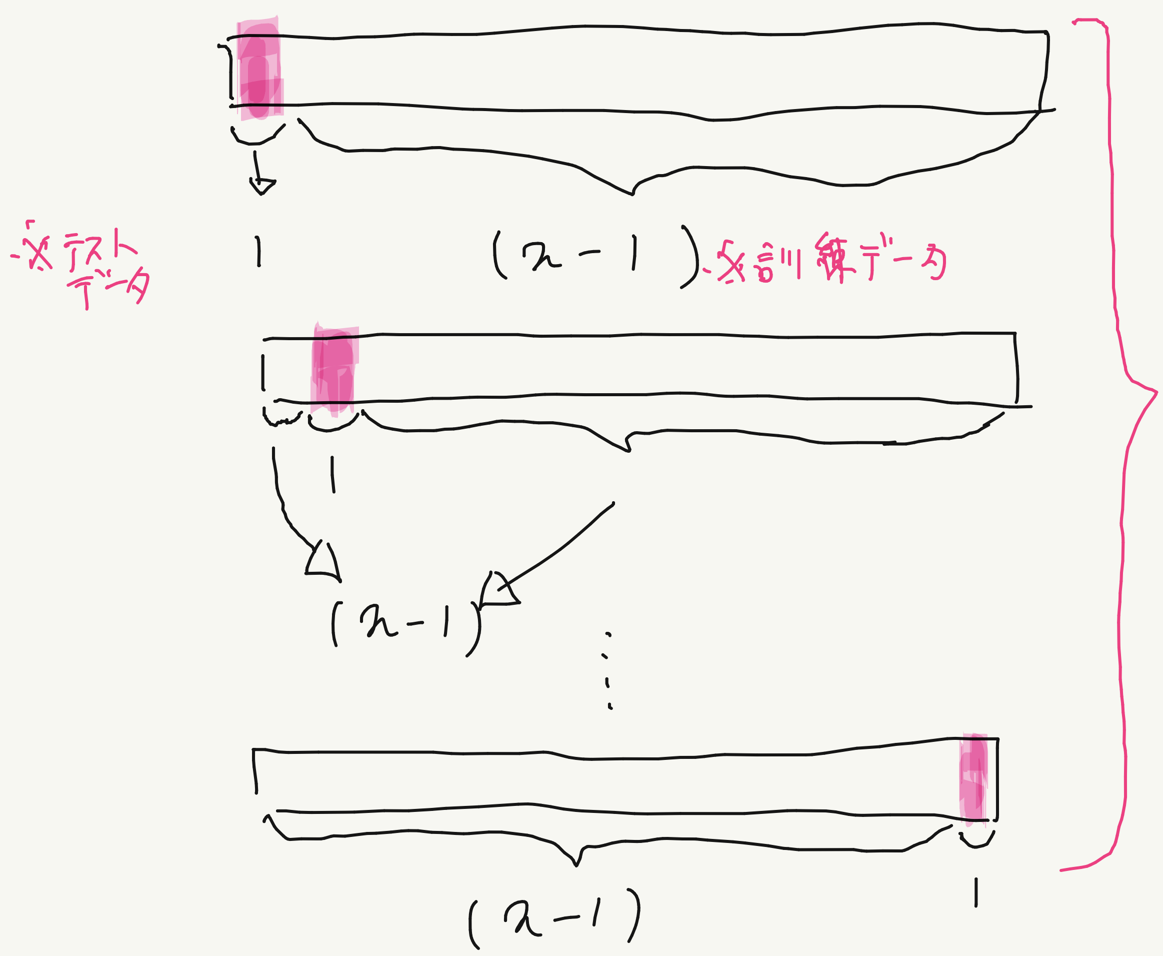
▶ ①，②の欠点を改善したものが，
「一つ抜き交差検証」と呼ばれるもの

• 一つ抜き交差検証（LOOCV）

データを分割するという考え方はホールドアウト検証と同様.
ただし，分割するのは，(n−1)：1。
つまり，訓練データ(n−1)でモデルを作成するので，
①，②の欠点を解決できる！

1の部分から生じた検証MSEの推定値の平均値をLOOCVのテストMSEの推定値とする

図解



それぞれから計算したn個の，テスト誤差の不偏推定値

$MSE_n = (y_n - \hat{y}_n)$

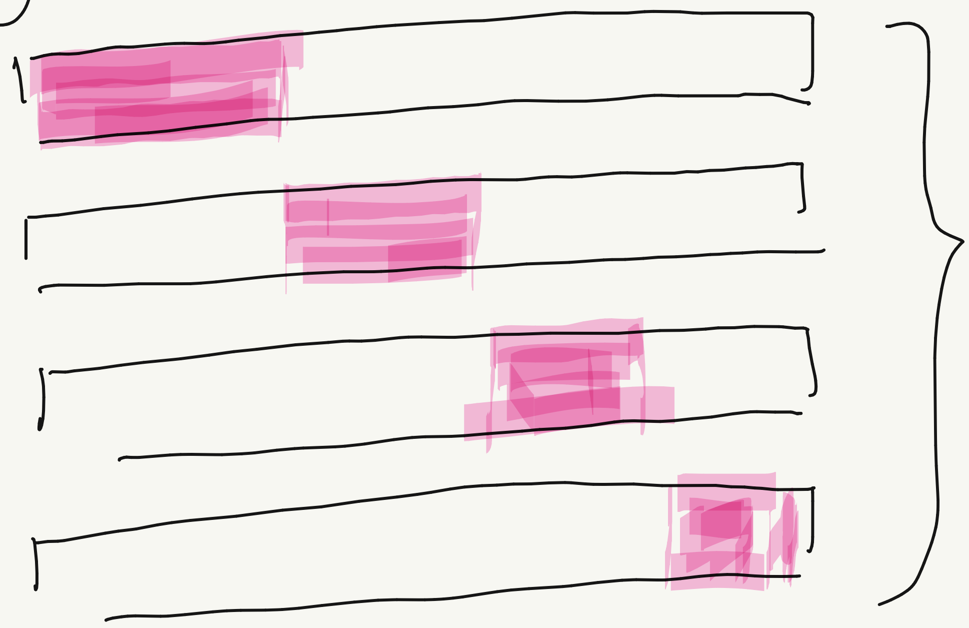
※検証MSEとも呼ぶ

これらの平均値をLOOCVにおけるテストMSEの推定値とする.

※この方法は，どのような予測モデル（ロジスティック，LDA等）にも使用が可能である.

● K分割交差検証

図解　K=4の場合



■部分が検証データとなる。

K個の■を使ってテストMSEを計算し、それらの平均値をテストMSEの推定値とする。

※実際は、K=5, K=10が用いられることが多い

② ブートストラップ

ある推定値や、統計的学習法についての不確かさを評価するのに用いられる。
(ばらつき)

※線形回帰ではSEが出力されるので、特に有用ではないが、その他ばらつきを測るのが困難な場合は有用。

→ つまり、統計的学習法において、推定されたパラメータの精度を測るために使用する。

検証に用いる標本は，元々手元にある標本から抽出し，使用する

→ 一つの標本集団より，複数のデータが得られ，そこから係数等の推定値の分布を描くことができる！

→ 真の値の区間推定などが可能に！

イメージ

n=3のデータセットの場合

	X	Y
a	4.3	2.4
b	2.1	1.1
c	5.3	2.8

Z^{*1}

	X	Y
c	5.3	2.8
a	4.3	2.4
c	5.3	2.8

Z^{*2}

	X	Y
b	2.1	1.1
c	5.3	2.8
a	4.3	2.4

⋮

Z^{*B}

	X	Y
b	2.1	1.1
b	2.1	1.1
a	4.3	2.4

ここから算出した何らかの値を基として，SEを推定する。

① 図のように，元々のデータセットより，重複を許して，n個のデータを無作為抽出にてとり出す。

② その後とり出したデータセットで，目的としている値を計算し，（※係数など）それを使ってSEを算出する。

→ これがブートストラップ推定値の標準誤差（SE）となる。

重要メモ

$\hat{f}(x_0)$ の分散とバイアス とは…？

- 分散

→ 異なる訓練データセットを使ったときに（※モデルを作成した時）
どの程度 $\hat{f}(x_0)$ が変化するかを表す量。

※小さい方がbetter.

↳ 統計的学習が柔軟であるほど
（※非線形にも柔軟に対応できるモデル）
分散は大きくなる。

- バイアス

→ 極めて複雑な現象をより単純なモデルで
近似したために生じる誤差を指す。
非線形のデータに線形の方法論を当てはめた場合、
バイアスは大きくなる。

⇒ 一方が小さくなるような方法を用いるのは容易であるが、
両方が小さくなるような方法を見つけることは難しい。

これを「バイアスと分散のトレードオフ」と呼ぶ