DataScientist Society

過学習と精度検証法

～ホールドアウト法, クロスバリデーション法～

過学習(Over fitting)

- 過学習とは?

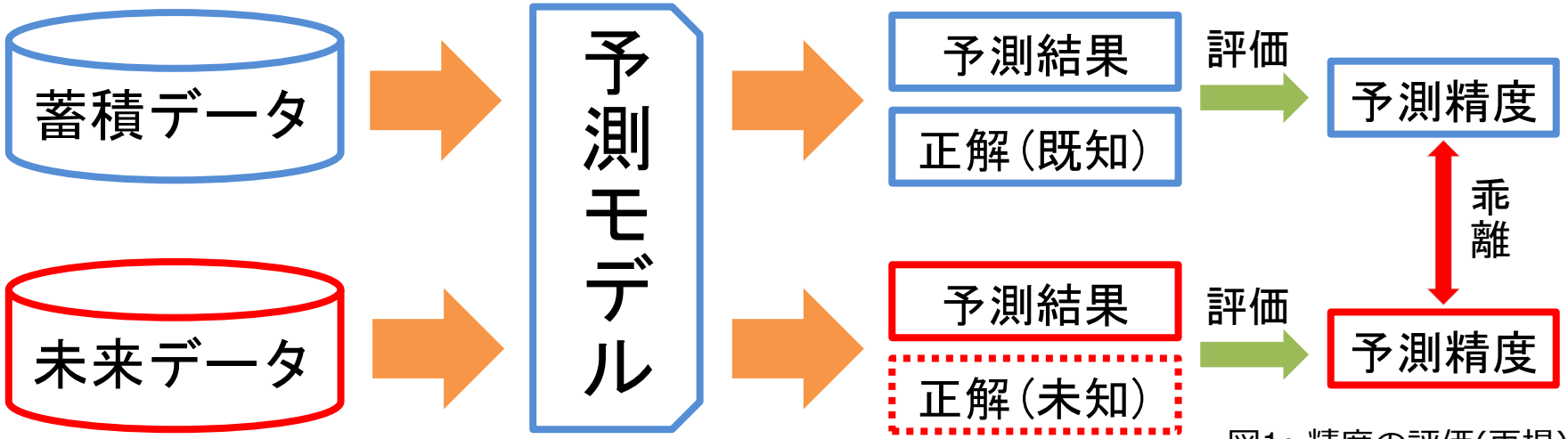


図1: 精度の評価(再掲)

図1は, 第1回の講座で扱った「モデルの精度評価」にて掲載した図である. モデルを構築した時のデータ(蓄積データ)での**予測精度が良い**のに, モデル構築に使用していないデータ(未来データ)での**予測精度が悪く**, 2つの予測精度の乖離が大きいとき,

予測モデルは**「過学習(過剰適合)」**している

という.

過学習(Over fitting)

- 過学習の例(犬猫判別)

一般に, 学習データに対してモデルが複雑であり, 多様なパターンを判別できる場合に過学習が起きやすい. 先程、ご説明した決定木だと, 学習に使用する変数が多かったり, 階層の数を増やしたり, 終端ノード数を少なくしたりすると, 図2の様に過学習を起こす要因となる.

表1: 学習データ(犬猫)

真の値	体重[kg]	エサ代[円]	飼主の身長[cm]
猫	5.8	8000	165
猫	4.6	7000	170
犬	7.8	8000	163
猫	10.4	6000	184
犬	15.5	9000	171

過学習となる要素

1. 学習データ数に対して, 決定木モデルが複雑である(最小終端ノード数が1, 階層が2)
2. 意味のない変数を学習している(犬猫を飼うことと身長は無相関という前提)

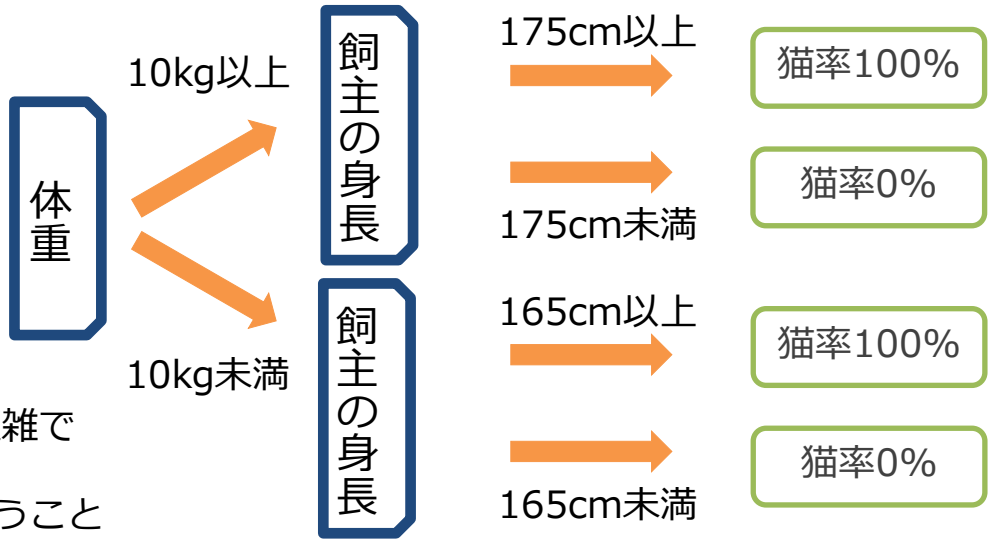


図2: 決定木モデル(過学習)1

この場合, 過学習を回避するには, 最小終端ノードを2にするか, 階層を1にすること. もしくは, 「飼主の身長」のデータを除くことなどが挙げられる.

過学習(Over fitting)

- 過学習の例(銀行の顧客ターゲティングデータ)
  本講座で扱っている「銀行の顧客ターゲティングデータ」において，決定木の階層を増やすことによって過学習を確認できる．

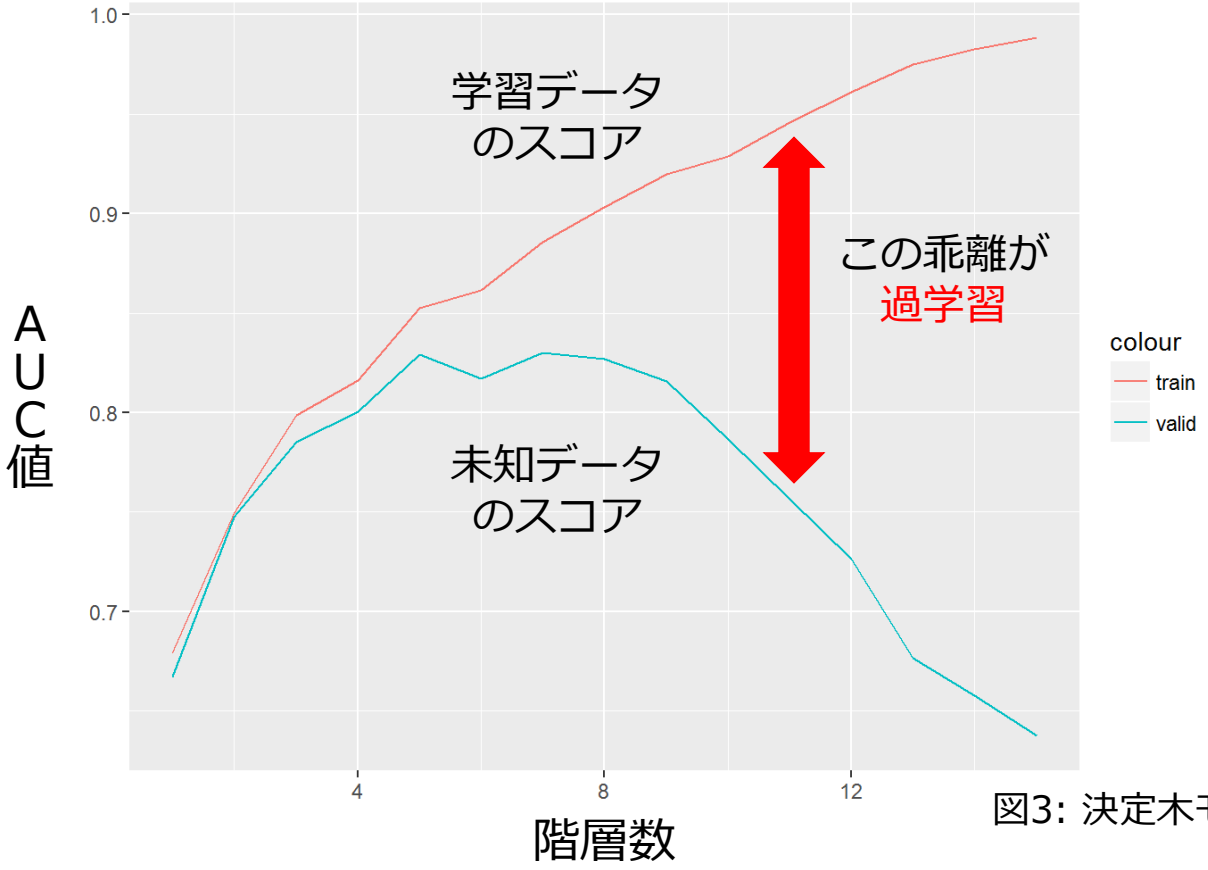


図3: 決定木モデル(過学習)2

過学習(Over fitting)

- 未知のデータに対する予測精度を上げるには?

  学習するデータに対して, 複雑なモデルを用いることにより, 多様なパターンを認識して精度の良いモデルを構築したいが, 不必要なパターンまで拾ってきて欲しくはない…

  ⇒ **最適なモデルの複雑さを推定すれば良い**

- モデルの最適な複雑さ(パラメータ)の決定方法

  既に知られているパラメータのチューニング方法として, 下記の方法が挙げられる.

  1. ホールドアウト法
  2. クロスバリデーション法(K分割交差検証法)

  次頁より, ホールドアウト法, クロスバリデーション法を紹介する.

ホールドアウト法

- ホールドアウト法とは(Hold out法)

　与えられた**答えの分かっているデータ**を任意の割合で「モデル構築データ」と「モデル検証データ」の2つに分け，「モデル構築データ」のみを用いて，モデル構築を行う．このとき，モデルにとって「モデル検証データ」は未知の存在であるから，実際の未知データに対しての予測精度を推し量ることができる．

　しかし，「モデル構築データ」を多くすると「モデル検証データ」が減り，未知のデータに対する予測精度を推し量ることができない．

　一方で，「モデル検証データ」を多くすると「モデル構築データ」が減り，モデルの精度が上がらなくなる．したがって，

十分なデータ量がある場合のみ有効な手法

である．

※十分なデータ量とは，予測対象や使用モデルにより基準が異なる．

　また，構築データと検証データの割合は7(構築):3(検証)とする場合が多い．

ホールドアウト法

- ホールドアウト法

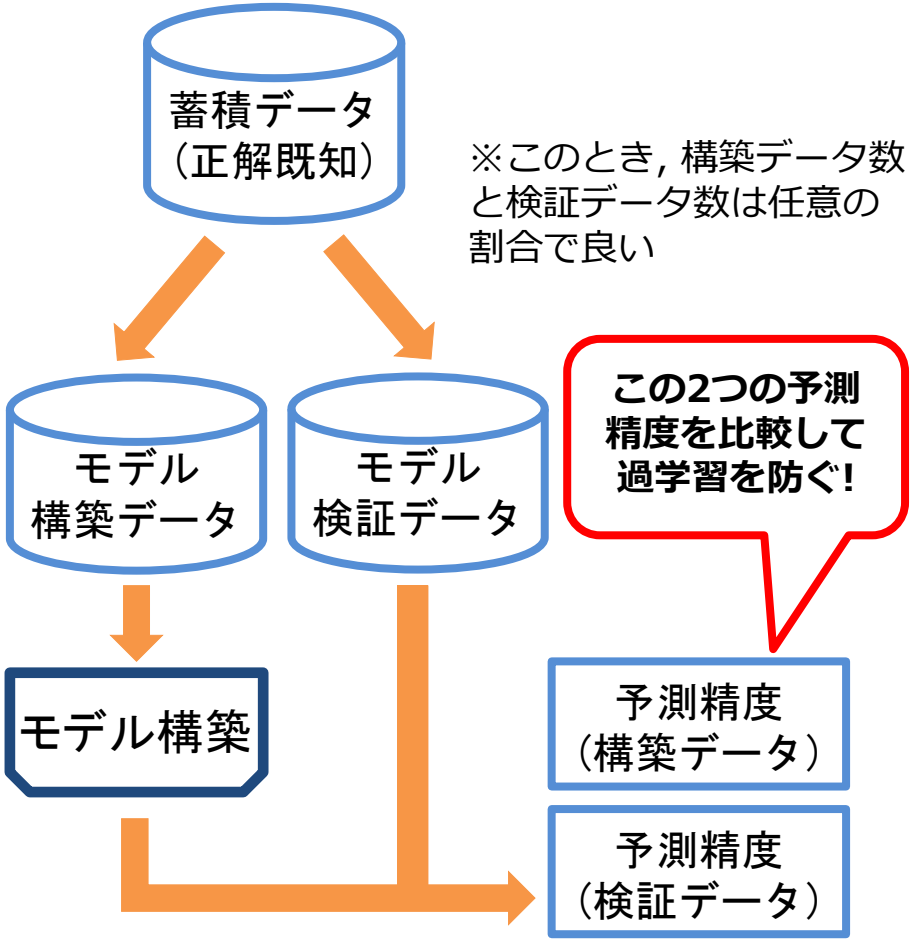
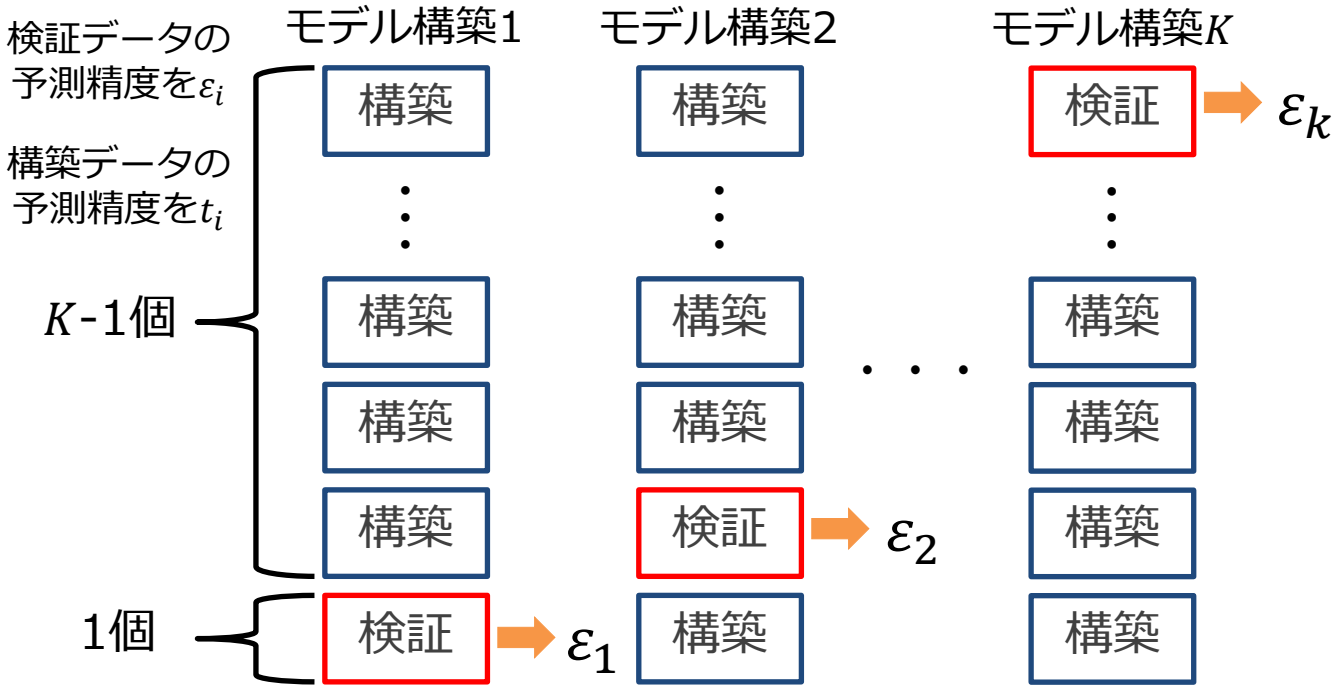


図4: Hold out法

クロスバリデーション法

- クロスバリデーション法(Cross validation法)とは

  K分割交差検証法とも呼ばれ, ホールドアウト法の欠点を補った手法である. 与えられた**答えの分かっているデータ**をK個に分割し, そのうち1つを「モデル検証データ」にし, 残りのK-1個を「モデル構築データ」とする. これによりモデルをK回構築することで, 未知データへの予測精度を推し量る.



検証データへの予測精度

$$\varepsilon = \frac{1}{K}\sum_i \varepsilon_i$$

構築データへの予測精度

$$T = \frac{1}{K}\sum_i t_i$$

この2つの予測精度を比較して過学習を防ぐ!

クロスバリデーション法

- クロスバリデーション法のサンプルプログラム
  ⇒別資料に記載.

まずは, 基本となるクロスバリデーションを理解してから手を出そう

- 様々なクロスバリデーション法(一部)
  - ✓ Group K-Fold：Groupとみなす変数が同じ値のものは構築, 検証に所属する側を一致させる
    → 例：不動産価格予測で分譲住宅など

  - ✓ Stratified K-Fold：分類問題などで, 構築と検証の各クラスの分布が等しくなるように分割を行う
  - ✓ Stratified Group K-Fold：上記2つの複合

他にも沢山の種類のCVが考案されていますので, 興味のある方は調べてみてください

参考文献

[1] 平井有三：
はじめてのパターン認識, 2012年, 森北出版株式会社
[2] Trevor Hastie, Robert Tibshirani, Jerome Friedman :
The Elements of Statistical Learning, 2009, Springer
[3] 石田基広：
Rで学ぶデータ・プログラミング入門, 2012年, 共立出版