DataScientist Society

集団学習(ensemble learning)

～平均, メタモデル, バギング, スタッキング, ブースティング～

今回の目標

➢ 今回の目標

前回までに「決定木」「ロジスティック回帰モデル」を学んできた. 今回は複数のモデルを使い, より高い精度を出すことについて学習する. 下記の4項目を今回の目標とする.

1. 平均, メタモデルによる集団学習を実践する
2. バギングの理論について理解し, 実践する
3. 構築したモデルによる構築データに対する予測結果には, 「答え」の情報を含んでいることを理解する
4. スタッキングの理論を理解し, 実践する

集団学習(ensemble learning)

➢ 集団学習とは?

複数のモデルの予測結果を上手く組み合わせることで, より予測精度の高いモデルを構築することを, 集団学習(ensemble learning)またはアンサンブル学習という. 今回は集団学習の手法である平均, メタモデル, バギング, スタッキング, ブースティング(紹介のみ)について解説する.

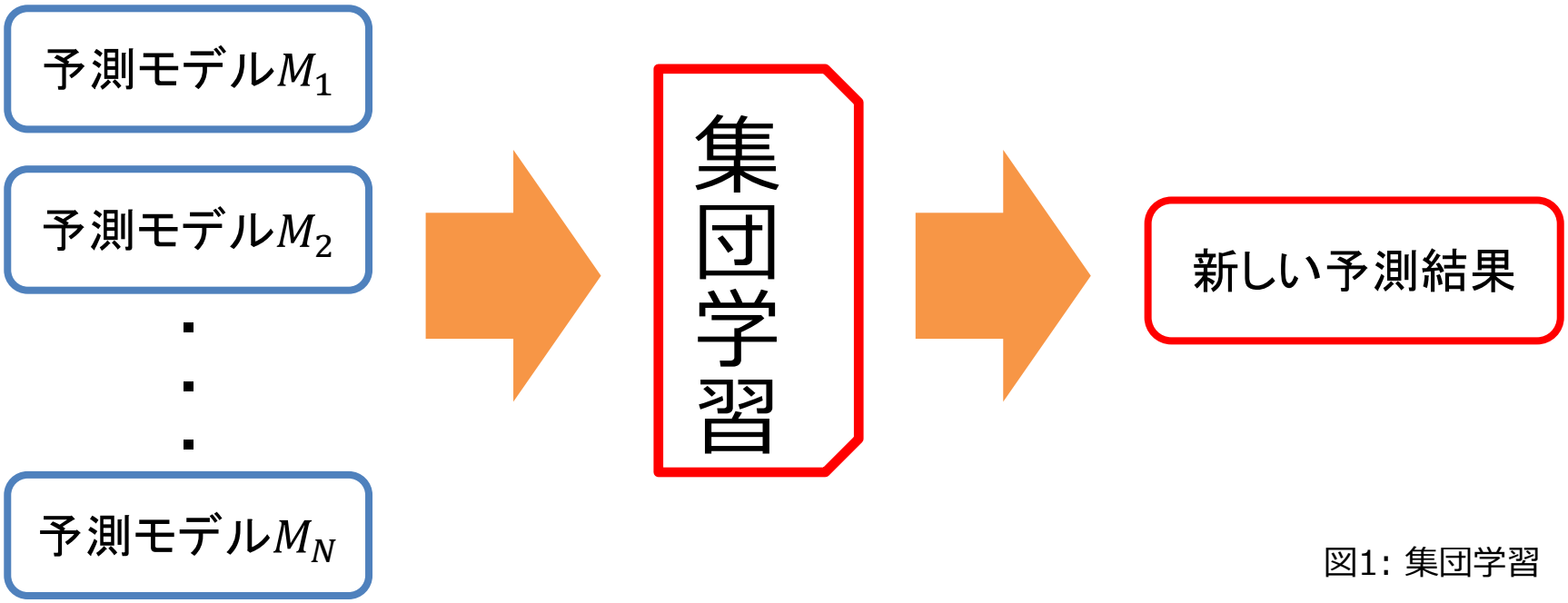

図1: 集団学習

平均

➢ 集団学習の簡単な例1

例として, 複数のモデルの予測結果に対して, 平均をとる集団学習方法が挙げられる. あるモデルの予測精度の悪い部分を, 他のモデルによって補うことができる.

表1: 集団学習の例1

予測対象	モデル1	モデル2	モデル3	モデル4	モデル5	平均	真の値
データ1	0.55	0.93	0.19	0.30	0.60	0.51	1
データ2	0.56	0.30	0.20	0.34	0.64	0.41	0
データ3	0.75	0.20	0.83	0.88	0.91	0.71	1
AUC	0.50	0.50	0.50	0.50	0.50	1	–

表1をみると, 色のついている数値は予測精度が悪いが, 他のモデルと平均をとることで, 良い修正がなされている. 実際に, AUCが0.5から1に増え, 精度が向上している.

平均による集団学習は, ランダム性のある局所最適に陥りやすいモデルにも有効な手段である(GBM, XGBoostなど).

メタモデル

➢ 集団学習の簡単な例2

複数のモデルの予測結果を変数とみなし, 学習データとして再度モデル構築を行う.

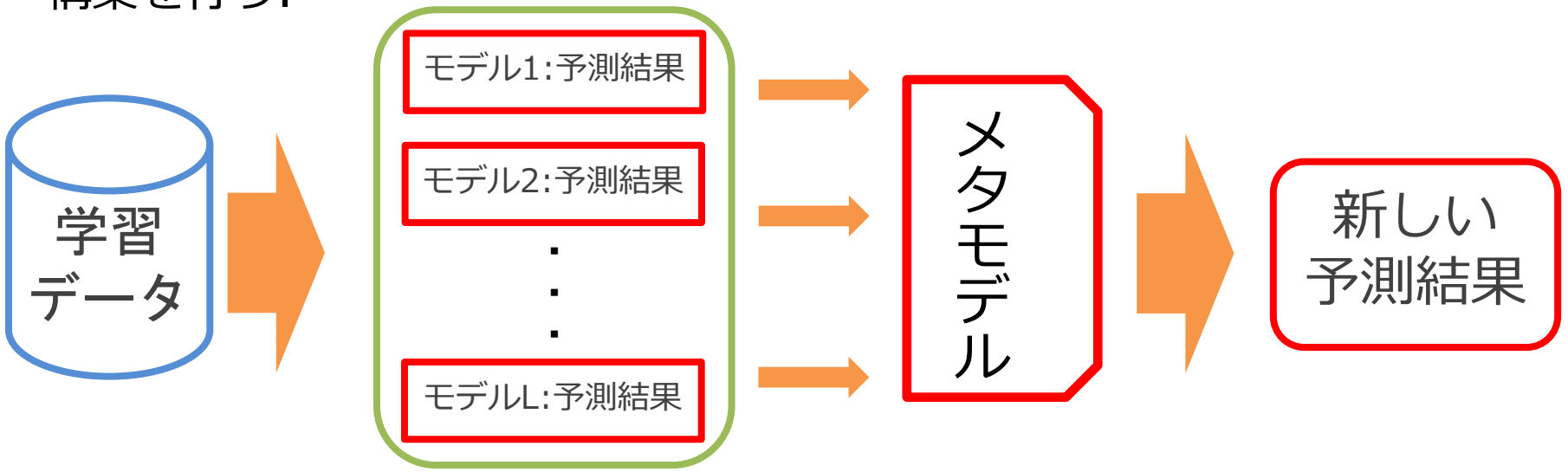

図2: メタモデル

メタモデルとして, ロジスティック回帰を行うとL個のモデルの予測結果に対して, 重みづけの平均に似たことを行ってくれる. ただし, モデルには学習データの「答え」の情報が入っているので, 過学習している場合には有効な手法ではない.

バギング(bagging)

- バギングとは?

以下のアルゴリズムによる集団学習をバギング(bagging)という.

1. N行の学習データから復元抽出でM行の学習データを取り出す
2. この学習データに対してモデルを構築する
3. 1, 2をL回繰り返してL個のモデルを構築する
4. 出力が回帰, 分類確率の場合は平均, 判別の場合は多数決により集団学習を行い, 新しい予測結果を作成する

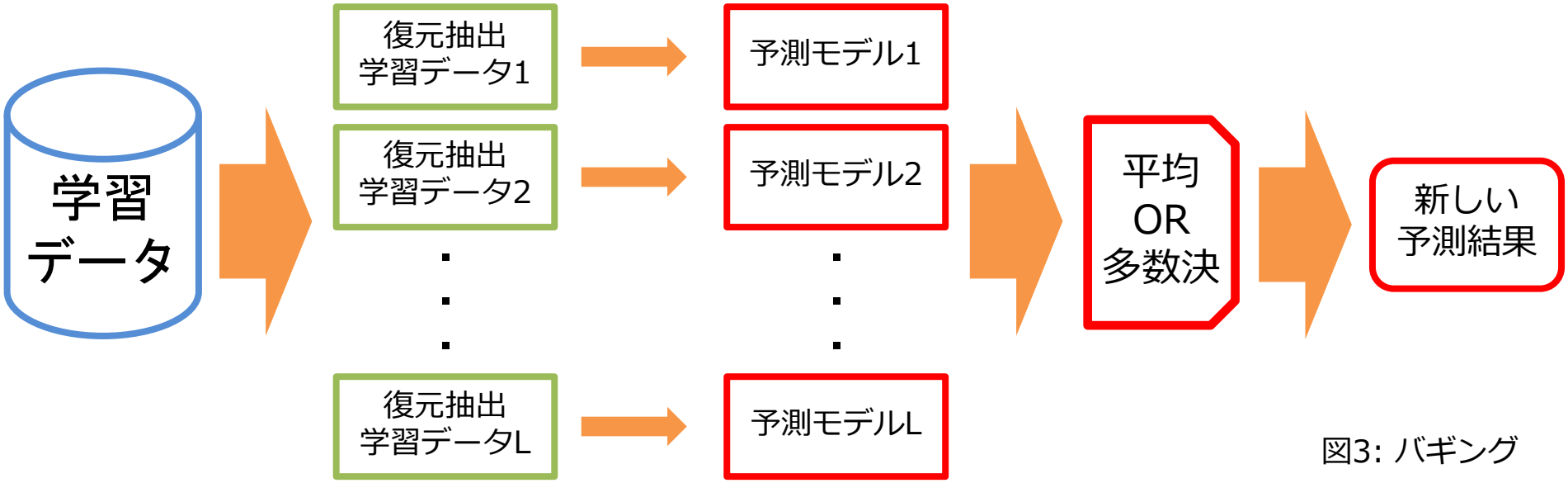

図3: バギング

バギング(bagging)

- バギングの実装

　アルゴリズムの理解を深めることを考慮して, パッケージを用いない実装例を示す(別資料参照).

- ランダムフォレストはバギングの類似手法

スタッキング(stacking)

- スタッキングとは?

以下のアルゴリズムによる集団学習をスタッキング(stacking)という.

1. 学習データをK個に分割する
2. K-1個をモデル構築に使い, 残りの1個を予測に使う
3. 2をK回繰り返すことで, 「答え」の情報を含まずに学習データ全体の予測結果を出すことができる

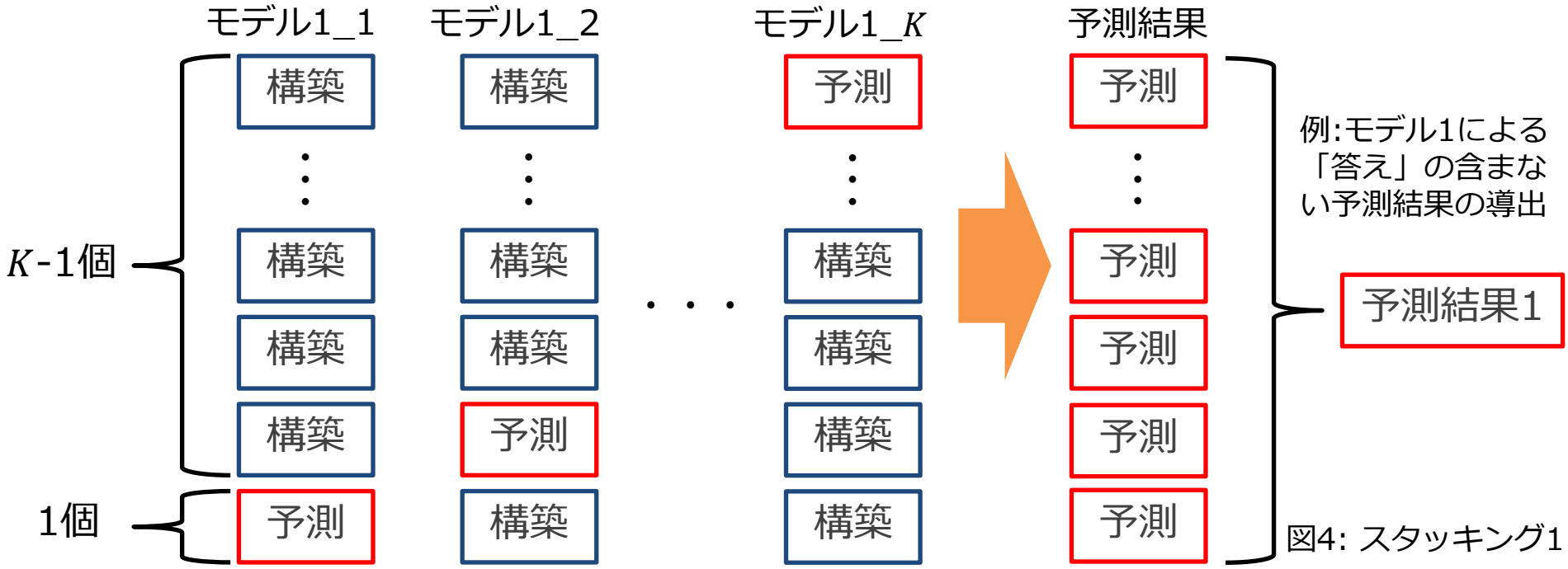

図4: スタッキング1

スタッキング(stacking)

4. 2, 3をL回繰り返すことでL個のモデルに対して「答え」の情報を含まない予測結果を得る(このとき, L×K個のモデルが出来ている)

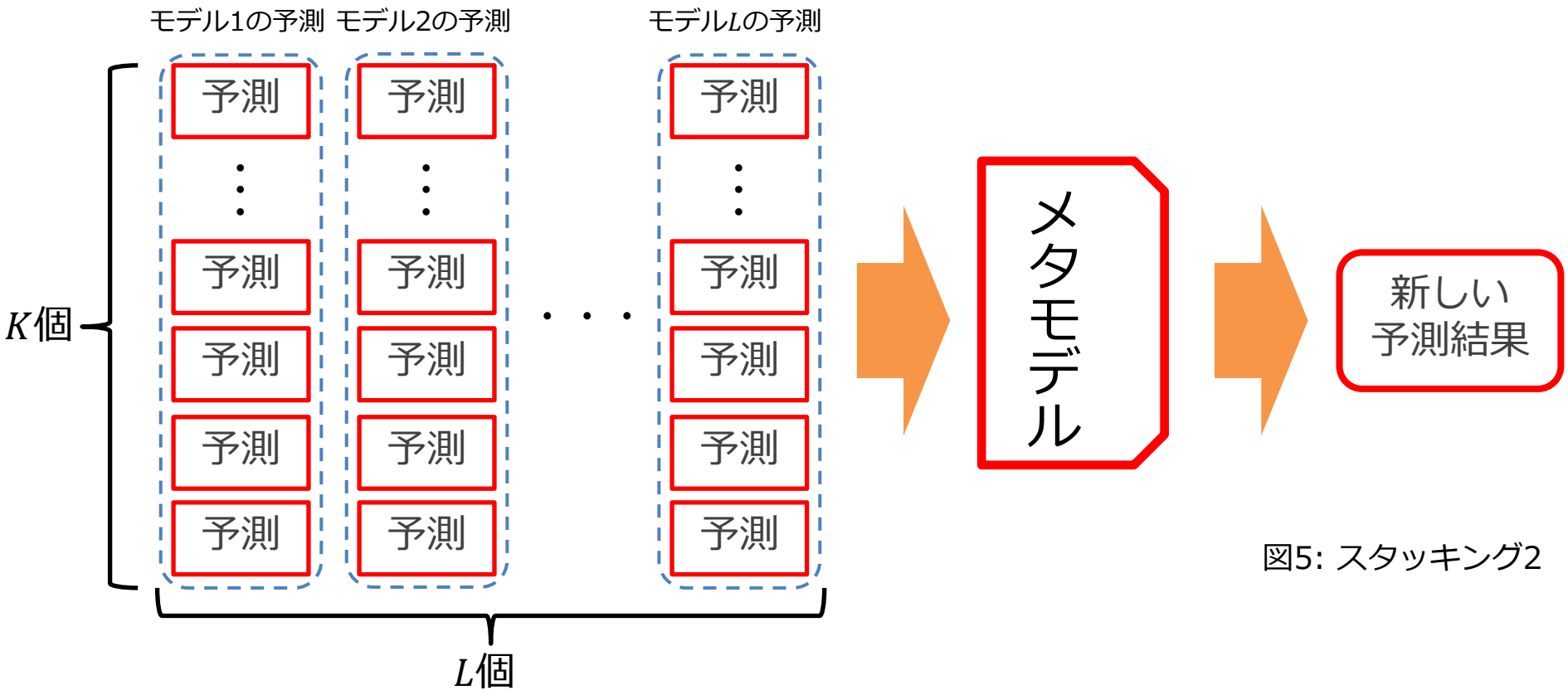

図5: スタッキング2

5. 4で出来たL個の予測結果を説明変数として, メタモデルを構築する

スタッキング(stacking)

検証データへの適用方法1:
　検証データに適用する場合は, モデル1のK個のモデル(モデル1_1～モデル1_K)に対して, 検証データ全体を適用し, 予測結果を平均してモデル1の予測結果としてメタモデルに投入する.

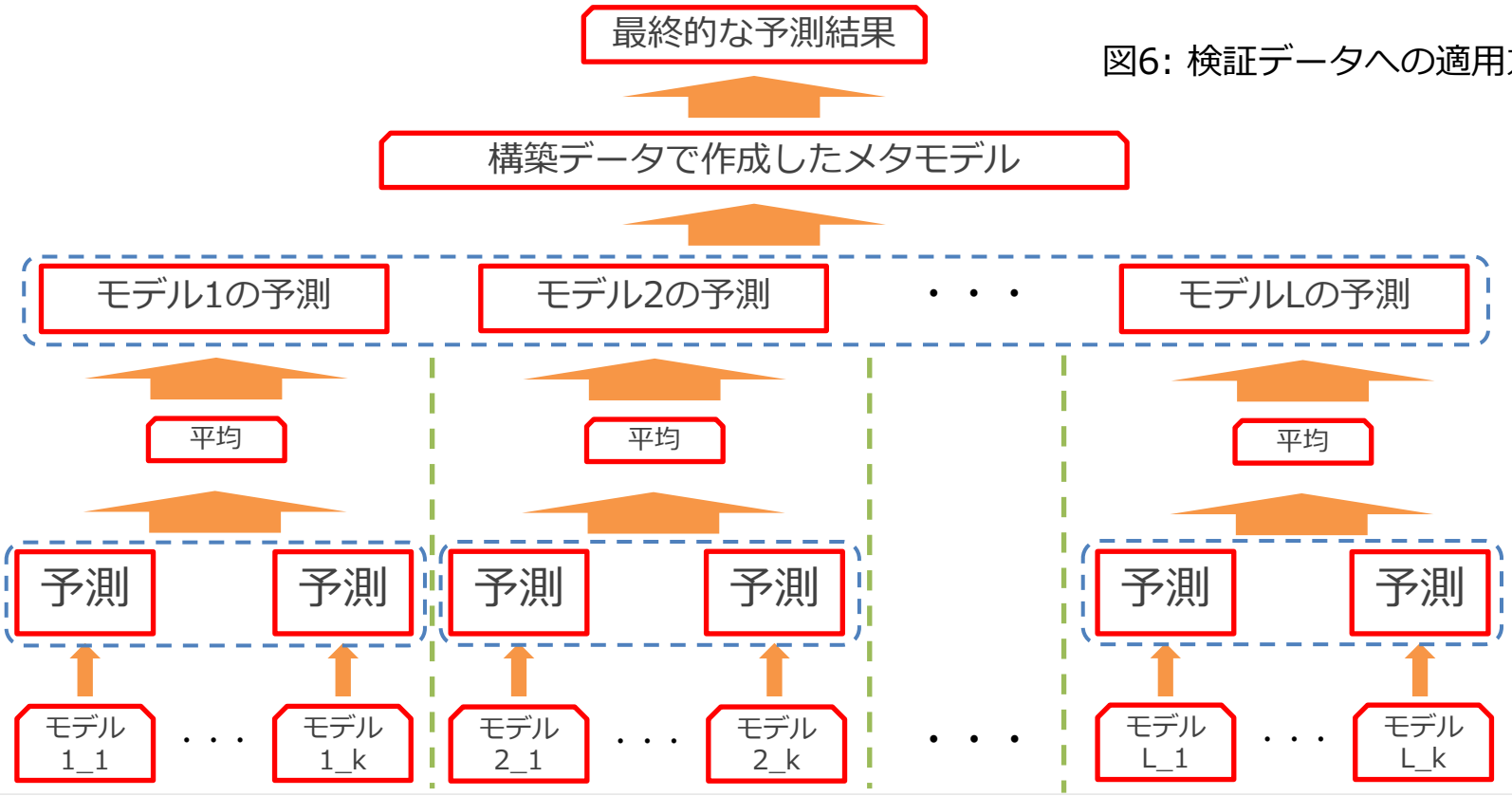

図6: 検証データへの適用方法1

スタッキング(stacking)

検証データへの適用方法2:
　モデル1のK個のモデル(モデル1_1～モデル1_K)を構築する際に, どの構築時でも同じパラメータで固定した場合, 最後に学習データ全体を使って, 固定した同じパラメータでモデル1_allを構築する. このモデル1_allに検証データ全体を適用し, その予測結果をメタモデルに投入する.

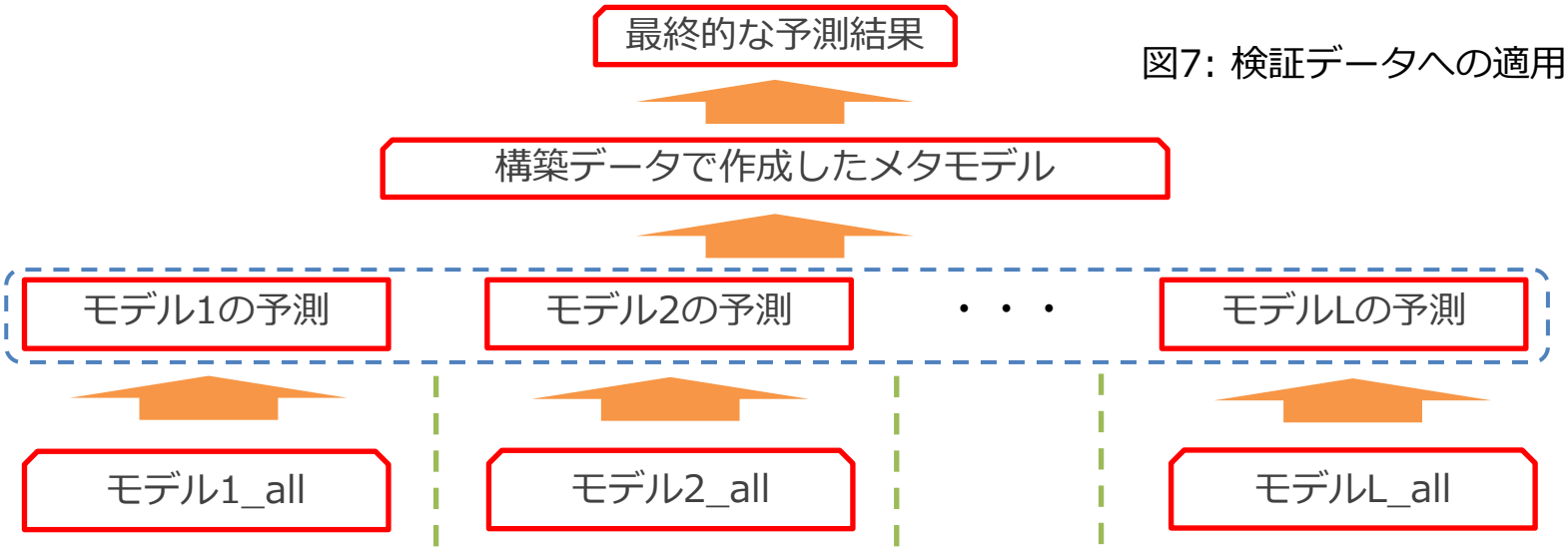

図7: 検証データへの適用方法2

- スタッキングの実装
  (別資料参照).

ブースティング(boosting)：紹介

- ブースティングとは

複数のモデルを順番に学習していく. このとき, 正答率に関連して, 予測対象ごとに重みづけなどを行い, モデルを学習する手法である.

- 有名なブースティング手法

1. アダブースト(Adaboost)
2. 勾配ブースティング(Gradient Boosting)

が挙げられる.

コンペ等で有名なGBM(Gradient Boosting Machine), XGBoostは勾配ブースティングを用いた手法である.

演習課題

1. 前回までに作成した決定木とロジスティック回帰の結果を用いて, 平均(もしくは, ロジスティック回帰でメタモデル)によるアンサンブルを行い, サブミットしてスコアを確認する(※ランダムフォレスト等の他のモデルもあれば使用する)

2. バギングの理論とプログラムを理解しながら実行する

3. スタッキングの理論とプログラムを理解しながら実行する

4. 班の中の人達のモデルを寄せ集めて,平均, メタモデル, バギング, スタッキングのいずれかを用いてアンサンブルを行いサブミットする

参考文献

[1] Trevor Hastie, Robert Tibshirani, Jerome Friedman : The Elements of Statistical Learning, 2009, Springer
[2] 金森敬文：統計的学習理論, 2015年, 講談社