回帰モデル・クラス分類モデルを評価・比較するための

モデルの検証
Model validation

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

“良い”回帰モデル・クラス分類モデルとは何か？

✓新しいサンプルの目的変数の値・ラベルを、正確に推定できるモデルが
良い回帰モデル・クラス分類モデル

- 回帰モデル・クラス分類モデルを構築したサンプルではないことに注意

✓そのような良いモデルを選ぶために、
いろいろなモデルを評価・比較しなければならない

✓モデルを評価・比較するための、モデルの検証の話です

データセットの呼び方

✓トレーニングデータ (キャリブレーションデータ)

- 回帰モデル・クラス分類モデルの構築に用いるデータ
- 目的変数の値・ラベルは分かっている

✓バリデーションデータ・テストデータ

- 回帰モデル・クラス分類モデルの検証に用いるデータ
- 実際には目的変数の値・ラベルは分かっているが、
  わからないものとして (目隠し・ブラインドして)モデルから推定し、
  実際と推定結果とがどれくらいあっているか確認する
  - バリデーションデータで、モデルのハイパーパラメータ (PLSの最適成分数など) を最適化する
  - テストデータで、最終的にモデルの優劣を比較する
  - バリデーションデータはなく、トレーニングデータとテストデータだけのときもある (このときのモデルのハイパーパラメータの最適化については後述)

比較指標

✓モデルの性能を評価し、比較するための指標
- 基本的には比較だけに用いるのがよく、絶対的な値に意味はない

✓トレーニングデータ・バリデーションデータ・テストデータそれぞれについて、実際の目的変数の値・ラベルと、推定された値・ラベルとが揃うと計算できる

✓回帰分析
- 決定係数 r^2
- 根平均二乗誤差 (Root Mean Squared Error, RMSE)
- 平均絶対誤差 (Mean Absolute Error, MAE)

など

✓クラス分類
- 混同行列 (confusion matrix) を計算したのちの、正解率、精度、検出率、誤検出率、Kappa係数など

回帰分析 決定係数 r^2

- ✓目的変数のばらつきの中で、回帰モデルによって説明できた割合
- ✓1に近いほど回帰モデルの”性能”が高い
  - どんな“性能”かは、r^2 を計算したデータセット・推定値による
- ✓相関係数 r を二乗したものとは異なる
- ✓異なるデータセットの間で r^2 を比較してはいけない

$$r^2 = 1 - \frac{\sum_{i=1}^{n} \left(y^{(i)} - {y_{\mathrm{EST}}}^{(i)} \right)^2}{\sum_{i=1}^{n} \left(y^{(i)} - y_{\mathrm{A}} \right)^2}$$

$y^{(i)}$：i 番目のサンプルにおける目的変数の値

${y_{\mathrm{EST}}}^{(i)}$：i 番目のサンプルにおける目的変数の推定値

y_{A}：目的変数の平均値

n：サンプル数

回帰分析 RMSE

- ✓平均的な誤差の大きさ
- ✓0 に近いほど回帰モデルの”性能”が高い
  - どんな“性能”かは、RMSE を計算したデータセット・推定値による
- ✓異なるデータセットの間で RMSE を比較してはいけない
- ✓データセットが同じであれば、r^2 が大きいほど RMSE は小さい
- ✓外れ値 (異常に誤差が大きいサンプル) があると、その値の影響を受けやすく、RMSE が大きくなりやすい

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y^{(i)} - y_{\mathrm{EST}}{}^{(i)}\right)^2}{n}}$$

回帰分析 MAE

- ✓平均的な誤差の大きさ
- ✓0 に近いほど回帰モデルの”性能”が高い
  - どんな“性能”かは、MAE を計算したデータセット・推定値による
- ✓異なるデータセットの間で RMSE を比較しないほうがよい
- ✓外れ値 (異常に誤差が大きいサンプル) の影響を受けにくい

$$MAE = \frac{\sum_{i=1}^{n} \left| y^{(i)} - y_{\mathrm{EST}}{}^{(i)} \right|}{n}$$

クラス分類 混同行列・正解率・精度・検出率

✓混同行列 (confusion matrix)

		予測されたクラス	
		1 (Positive, 陽性)	-1 (Negative, 陰性)
実際のクラス	1 (Positive, 陽性)	True Positive (TP)	False Negative (FN)
	-1 (Negative, 陰性)	False Positive (FP)	True Negative (TN)

$$\text{正解率} = \frac{TP + TN}{TP + FN + FP + TN} \qquad \text{検出率} = \frac{TP}{TP + FN}$$

$$\text{精度} = \frac{TP}{TP + FP} \qquad \text{誤検出率} = \frac{FP}{FP + TN} \quad \text{など}$$

クラス分類 Kappa係数

✓実際と予測結果の一致度を評価する指標

✓Positive(陽性)データとNegative(陰性)データの偏りがある時に有効

$$\text{Kappa係数} = \frac{\text{正解率} - \text{偶然による一致率}}{1 - \text{偶然による一致率}}$$

$$\text{偶然による一致率} = \frac{TP + FN}{A} \times \frac{TP + FP}{A} + \frac{FP + TN}{A} \times \frac{FN + TN}{A}$$

$$(A = TP + FN + FP + TN)$$

http://en.wikipedia.org/wiki/Cohen%27s_kappa

		予測されたクラス	
		1 (Positive, 陽性)	-1 (Negative, 陰性)
実際のクラス	1 (Positive, 陽性)	True Positive (TP)	False Negative (FN)
	-1 (Negative, 陰性)	False Positive (FP)	True Negative (TN)

モデルの評価・比較 ハイパーパラメータの決定

✓ハイパーパラメータ

- PLSの最適成分数
- LASSOの λ
- SVMの C、γ

など

✓良いモデル (p.1 参照) になるようにハイパーパラメータを決めたい

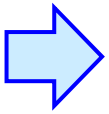
どのようなハイパーパラメータを用いるか？

✓トレーニングデータの比較指標の値がよくなるようなハイパーパラメータ

- そもそもモデルがトレーニングデータを用いて構築されているため、トレーニングデータには合うが、新しいサンプルの目的変数を推定できないようなハイパーパラメータが選ばれてしまう
- 基本的に用いられない

✓バリデーションデータの比較指標の値がよくなるようなハイパーパラメータ

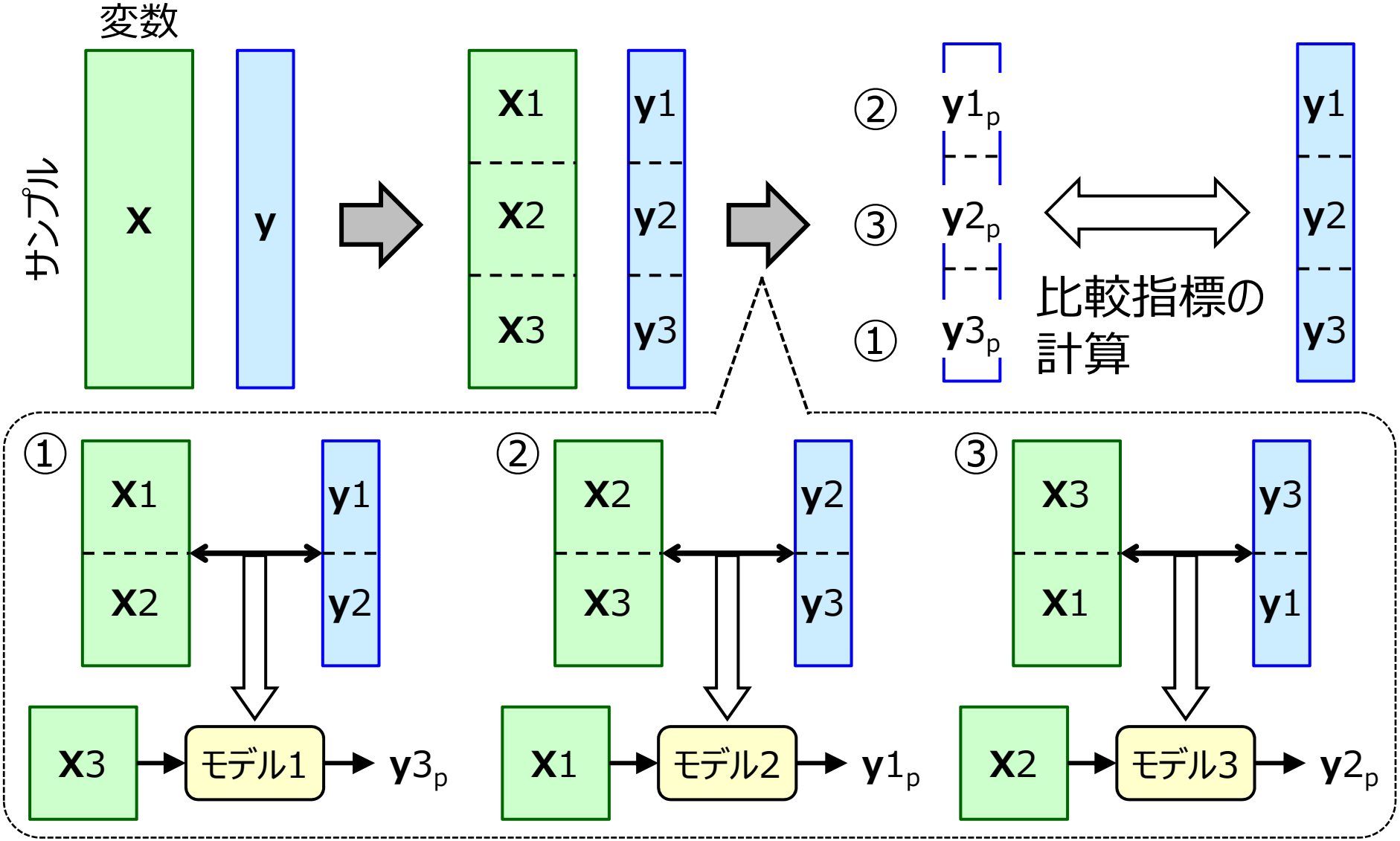
- 新しいサンプルに対する推定性能を考慮できる
- データに偏りがないようにトレーニングデータとバリデーションデータとを分けるよう注意する
- トレーニングデータが少なくなってしまう
  - ハイパーパラメータを決めた後、バリデーションデータも合わせて再度モデルを構築する
- 十分にデータ数が多いとき以外は、あまり用いられない



クロスバリデーション

クロスバリデーション

✓例) 3-fold クロスバリデーション



クロスバリデーションの補足

✓Leave-one-out クロスバリデーション

- サンプルを１つ除いて、残りのサンプルでモデルを構築し、
  除いたサンプルを推定する、ということをサンプル数だけ繰り返す
- 特にサンプル数が多いときに、すべてのサンプルでモデルを構築し、
  すべてのサンプルを推定することと似てしまうため、望ましくない

✓2-fold, 5-fold, 10-foldが一般的

✓データ数が多すぎると、計算時間がかかりすぎてしまうときは、
トレーニングデータとバリデーションデータとを分ける方法を用いる

どのようにデータセットを分けるか？

✓トレーニングデータ・バリデーションデータ・テストデータで、
サンプルに偏りがない方がよい

- 基本的にランダムに分けるのでOK

✓トレーニングデータはなるべくばらついている方がよい

- Kennard-Stone (KS) アルゴリズムにより、
トレーニングデータ・バリデーションデータ・テストデータの順に選ぶ
  1. データセットの説明変数の平均を計算
  2. 平均とのユークリッド距離が一番大きいサンプルを選択
  3. 選択されていない各サンプルにおいて、これまで選択されたすべてのサンプルとの間でユークリッド距離を計算し、その中の最小値を代表距離とする
  4. 代表距離が最も大きいサンプルを選択する
  5. 3. と 4. とを繰り返す

Y-randomization (Yランダマイゼイション)

✓特に、サンプル数が少なく説明変数 (記述子) の数が多いとき、本当は X と y の間に相関関係がなくても、r^2, r^2_{CV}の値が大きくなってしまうことがある

- たまたま X のノイズと y との間で相関がでてしまう
- 偶然の相関

✓偶然の相関かどうかを見分けるため、Y-randomizationが行われる

- Y のみ値をランダムに並べかえて、おかしなデータセットにする
- モデリングして、 r^2, r^2_{CV}の値が 0 付近になることを確認する