4-8　データ活用実践
（教師あり学習）

東京大学 数理・情報教育研究センター
2020年5月11日

概要

- 機械学習，特に教師あり学習の種々のデータ分析手法（分類・回帰）と，それらを用いた一連のデータ分析プロセスを実行するための知識を学びます.

- データ解析例を通じてデータの前処理から分析までどのように実行されるかを体験します.

本教材の目次

機械学習

Arthur Samuel

「Field of study that gives computers the ability to learn without being explicitly programmed」(1959)

明示的なプログラムなしにコンピュータに学習能力を与える研究分野.

画像の被写体が車である事をコンピュータに判断させるようなルールを明示的にプログラムすることは非常に困難です.

▶ 代わりに機械学習では蓄積されたデータからルールを発見する学習方法自体をプログラムします.

機械学習

- データに潜むルール（パターン）を自動的に見つけます.
- 人がプログラムするのは認識の仕方ではなく学習方法です.

機械学習でのルール発見に用いられるデータを訓練データと呼びます.
訓練データにない未知データに対しても正しく予測する事（汎化）が目的です.
機械学習には大まかに教師あり学習と教師なし学習があります
※ 今回の講義では教師あり学習を取り扱います.

機械学習の流れ

1. 行いたいタスクの特定.
   やりたい事は分類，回帰，クラスタリング等のうちどれかを特定します.

2. 分析に必要なデータの確認，対象となるデータの収集.
   収集すべき教師値と予測に効きそうな特徴を検討します.

3. データの分析．機械学習による学習と評価.
   データの前処理・加工と機械学習によるデータ分析を実行します.

4. データ分析結果の共有，課題解決に向けた提案.
   データ分析結果のレポートを作成し，それを材料に次の施策を検討します.

教師あり学習

訓練データは(x, y) = (データの特徴，教師値)というペアの集まりです．

教師値はデータのラベルであり回帰では連続値，分類では離散値です．

教師あり学習による予測：
訓練データから x から y を正しく予測するルールの発見を目指します．

例えば画像認識では (x, y) = (画像データ，被写体名)

$x =$  教師あり学習による予測 → $y =$ 車（正しい予測）

その他の教師あり学習の例：
売上予測 → 店舗の売上高と店舗・地理・気候等のデータで学習する．
罹患予測 → 診断結果と体温等の受診者のデータで学習する．
成約予測 → 成約結果と商品情報・顧客情報のデータで学習する．
離反予測 → 離反情報とユーザーのサービス利用状況データで学習する．

データの収集

予測対象と関連する簡単な説明変数（特徴）の作成をし，データを収集します．

- アヤメの品種予測

教師値：アヤメの品種
　　Setosa, Virginica, Versicolor

特徴：sepal.length（がく片の長さ）
　　sepal.width（がく片の幅）
　　petal.length（花びらの長さ）
　　petal.width（花びらの幅）

特徴ベクトルx　　教師値y

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.8	4.0	1.2	0.2	Setosa
1	5.0	3.3	1.4	0.2	Setosa
2	7.1	3.0	5.9	2.1	Virginica
3	4.3	3.0	1.1	0.1	Setosa

1 行が 1 データ

教師値・特徴ベクトルは連続値・カテゴリ値（離散値）として収集します．

特徴ベクトルの設計には対象データの専門知識を有効活用しましょう．

データの設計・収集がうまくいけば，あとは汎用の機械学習で分析可能です．

訓練データが多いほど予測精度は向上します，出来るだけ収集しましょう．

データの加工・前処理

機械学習を精度良く正しく実行するにはデータの前処理・加工が必要です.

タイタニックの生存結果データ（教師値：Survived）

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp
1	0	3	Braund, Mr. Owen Harris	male	22.0	1
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1
3	1	3	Heikkinen, Miss. Laina	female	26.0	0
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1
5	0	3	Allen, Mr. William Henry	male	35.0	0
6	0	3	Moran, Mr. James	male	NaN	0

カテゴリ変数　欠損値

データクレンジング

- 外れ値検出・除去
- 欠損値除去・補完

カテゴリ値の数値への変換

- ダミー変数の追加
  例：Male, Female列を追加し0-1値に変換.

有効な前処理

- 標準化：列を平均0，分散1に変換.
- 正規化：列を$[0,1]$, $[-1,1]$に収める．列or行の平均を0にノルムを1に変換.
- 変数変換
- サンプリング：訓練データの一部を無作為に選び新たに訓練データとします.
  訓練データサイズの調整やアンサンブル学習の際に用います.

回帰と分類

教師あり学習はデータに割り当てられるラベル y が連続値・離散値（識別子，名称，カテゴリ値等）かに依って回帰・分類に分けられます.

回帰の例

- 売上予測：$y =$ 店舗の売上高，$x =$ 店舗・地理・気象・曜日情報
- 価格予測：$y =$ 住宅価格，x最寄駅の距離，築年数，間取り，設備，地理情報

分類の例

- 罹患予測 ： $y =$ 診断結果，$x =$ 受診者のレントゲン・血液検査・体温
- 成約予測 ： $y =$ 商品の成約結果，$x =$ 商品，顧客の収入・職種・家族構成
- 離反予測 ： $y =$ 離反状況，$x =$ ユーザーのサービス利用頻度・活動内容

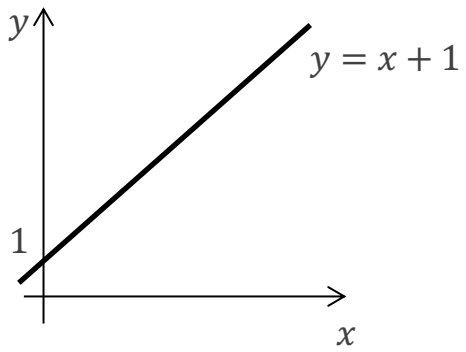
n 個の訓練データ$(x_1, y_1), \ldots, (x_n, y_n)$から写像・関数 $y = f(x)$を学習します.
f には様々なモデルがあり，モデルの選択が精度の担保に非常に重要です.

（参考）写像・関数とは？
入力xを出力yに対応付ける関係fを一般に写像といい$y = f(x)$と呼びます.
特に出力が実数値を取る対応関係は関数と呼びます.

モデル

- 関数・写像は入力xと出力yの関係を記述する：
$$y = f(x).$$

関数の例：近隣人口xに対し売上高yの増加は線形：
$$y = x + 1.$$



- 一般には正しい係数は不明なので以下の関係を想定する：
$$y = w_1 x + w_0.$$

係数w_0, w_1をパラメータと呼び，パラメータを持つ関数・写像をモデルと呼びます．

パラメータの値によってこの関数・写像の対応関係が変化しますが機械学習では訓練データ$(x_1, y_1), \ldots, (x_n, y_n)$に適合するようにパラメータを決定します．

（参考）この適合度は損失関数で計算されます．データ分析手法に応じて対応する損失関数が存在します．

種々のモデル

様々なモデルがあるのでデータに合わせ適切なものを選択します.

深層ニューラルネットも複雑なモデルの一つです.

モデルの例（赤字がパラメータ）

- $f_w(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$
- $f_w(x) = w_0 + w_1 \cos(2\pi x) + w_2 \cos(4\pi x) + w_2 \cos(8\pi x)$
- $f_w(x) = w_0 + \sum_{i=1}^{M} w_i \exp(-(x - \mu_i)^2)$

全パラメータをまとめてwと書きます. 最初の例では$w = (w_0, \ldots, w_p)$です.

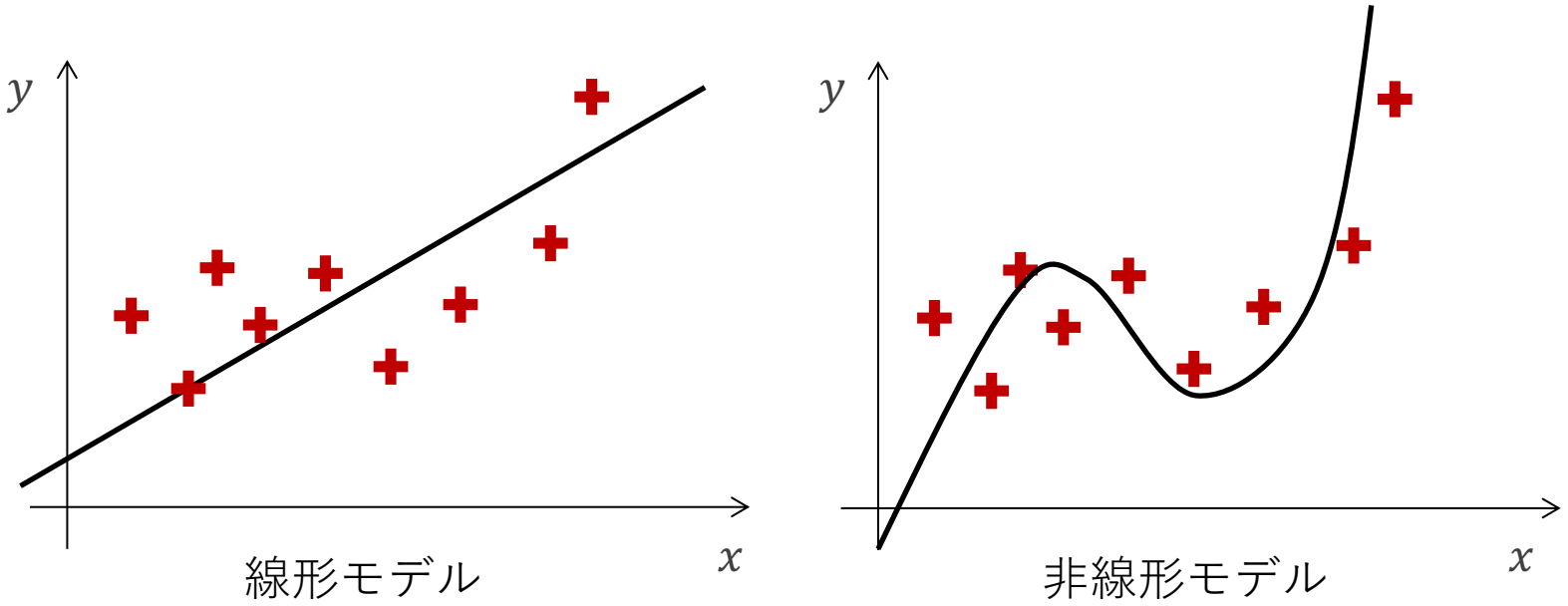
一般にパラメータwを持つモデルをf_wと書くことにします.

（参考）Σとは？共通の添字を持つ複数の数式の和を意味する記号です.
三番目の例では $w_i \exp(-(x - \mu_i)^2)$ が添字iを持つ数式で, これを$i = 1$から$i = M$まで足し上げています：
$\sum_{i=1}^{M} w_i \exp(-(x - \mu_i)^2) = w_1 \exp(-(x - \mu_1)^2) + \cdots + w_M \exp(-(x - \mu_M)^2)$.

回帰

特徴ベクトルxから連続値のラベル y を予測することを回帰と呼びます.
予測はパラメータwを持ち実数値に値を取るモデルf_wにより行います.

価格予測の例
y = 住宅価格, x = 最寄駅の距離, 築年数, 間取り, 設備, 地理情報



データの特徴が一変数の時は単回帰分析, 多変数の時は重回帰分析と呼びます.

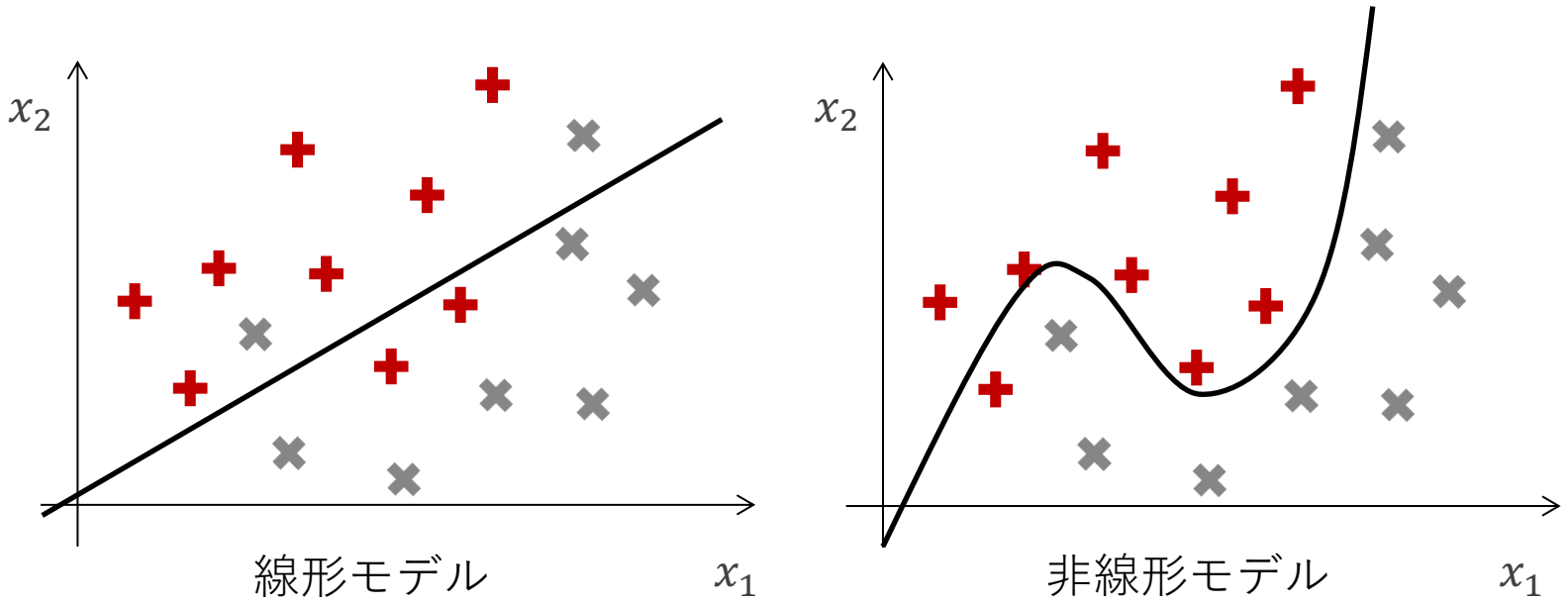
分類

特徴ベクトルxから識別子・カテゴリ等の離散値のラベル y を予測することを分類と呼びます．予測はパラメータwを持ち実数値（又は実数値ベクトル）に値を取るモデルf_wの出力を離散値に丸めて行います．

例えば診断結果等のラベルが1あるいは-1の二値で表されている時，$f_w(x)$が 0 以上であれば 1を割り当て，0 未満 であれば -1を割り当てます．

罹患予測の例

y = 診断結果，x = 受診者のレントゲン・血液・体温



損失関数

パラメータwを持つモデルf_wとデータ(x, y)の適合度合いは損失関数で定めます.
損失関数には様々な種類があり，それぞれ特有の性質を持ちます.

損失関数の例（モデルf_wは実数値に値と取るものとします．）

- 二乗損失関数

$$l(f_w(x), y) = 0.5(y - f_w(x))^2.$$

yと予測値$f_w(x)$が近い時に損失が小さく，主に回帰分析で用います.

- ロジスティック損失

$$l(f_w(x), y) = \log\left(1 - \exp\left(-yf_w(x)\right)\right).$$

yと$f_w(x)$の符号が同じで$yf_w(x)$の値が大きい程，損失は小さくなります.
ロジスティック損失を用いた分類分析をロジスティック回帰分析といいます.

経験損失最小化問題

機械学習では訓練データに平均的によく適合するパラメータを求めます．

これは次の経験損失最小化問題を解くことに帰着します:

$$\min_{w\in\Omega}\frac{1}{n}\sum_{i=1}^{n} l(f_w(x_i), y_i)\,.$$

（訓練データの損失関数値の平均）

（参考）最小値記号とは？関数がパラメータを持つとき，パラメータが取り得る範囲での関数の最小値を$\min$記号で記述します．便宜的にこの最小値を求める問題自体を表す事もあります．今回の場合は損失関数値の平均を集合Ωで動くパラメータwについて最小化する問題を表しています．

モデルの評価（回帰）

正解値$(y_i)_{i=1}^n$，予測値$(\hat{y}_i)_{i=1}^n$

- **二乗平均平方根誤差（Root mean squared error, RMSE）**

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\,.$$

- **平均絶対誤差（Mean absolute error, MAE）**

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,.$$

大きく外すデータがあると相対的にRMSEは大きくMAEは小さくなります.

- **決定係数（R^2）**：$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$，

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

平均値$\bar{y}$での予測との相対評価に他なりません.
$R^2 = 1$ は完全な予測，$R^2 = 0$ は平均値 $\bar{y}$による予測と同等，$R^2 < 0$は平均値 $\bar{y}$ による予測より劣ることを意味します.

モデルの評価（分類）

様々な指標があり目的に応じて使い分けます.

分類精度の他にクラス毎の評価値があります.

あるクラスCの予測		正解クラスがC	
		正	負
予測クラスがC	正	TP	FP
	負	FN	TN

このような分類結果の表を混同行列と呼びます.
TP, FP, FN, TNはそれぞれのケースに該当するデータ件数です.

適合率（precision） $\frac{TP}{TP+FP}$

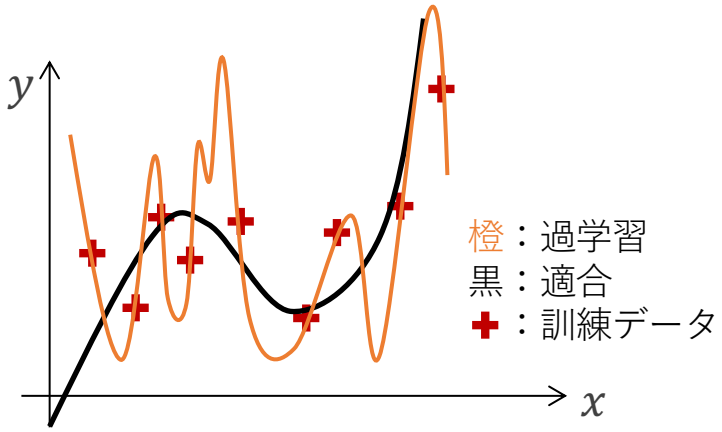
再現率（recall） $\frac{TP}{TP+FN}$

F-値 $\frac{2 \cdot \mathrm{recall} \cdot \mathrm{precision}}{\mathrm{recall}+\mathrm{precision}}$

- データのクラスバランスに偏りがある時，分類精度の意義は低くなります.
- 適合率・再現率のトレードオフは次の通りです.
  確かに正なものだけ正と予測 → 適合率：↗，再現率：↘
  少しでも正なものを正と予測 → 適合率：↘，再現率：↗
- F-値は両者の調和平均に他なりません.

過学習と正則化

訓練データに対しモデルが複雑な場合,
訓練データには適合しても未知データに
適合しない過学習という現象が起き得ます.



回帰問題での過学習の様子.

得られる関数の複雑さを抑制することで過学習を防ぐ技法を正則化と呼びます.
代表的な正則化手法として正則化付き経験損失最小化があります.
これは経験損失最小化問題に正則化項$\lambda R(w)$を加えた最小化問題です：

$$\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} l(f_w(x_i), y_i) + \lambda R(w)\,.$$

$w \in \mathbb{R}^p$はパラメータwが要素数pの任意のベクトル値を取れることを意味します.

正則化項の例：

L_2正則化 $R(w) = \|w\|_2^2 = \sum_{j=1}^{p} w_j^2$, リッジ回帰 = 二乗損失 + L_2正則化

L_1正則化 $R(w) = \|w\|_1 = \sum_{j=1}^{p} |w_j|$. LASSO回帰 = 二乗損失 + L_1正則化

交差検証

正則化係数λの大きさで訓練データへの適合度合いは変わります.

一般にλが小さいと過学習，λが大きいと訓練データにも適合しない未学習の状態になります.

交差検証は丁度良い適切なλを選ぶ手法です.

k-交差検証（Cross Validation, CV）

1．データをk個に分割します.

2．分割したデータの一つを検証用とし，残りのデータで学習します.

3．検証用データでの予測誤差を計算します.

4．手順2，3をk個の検証用データの取り方について繰り返します.

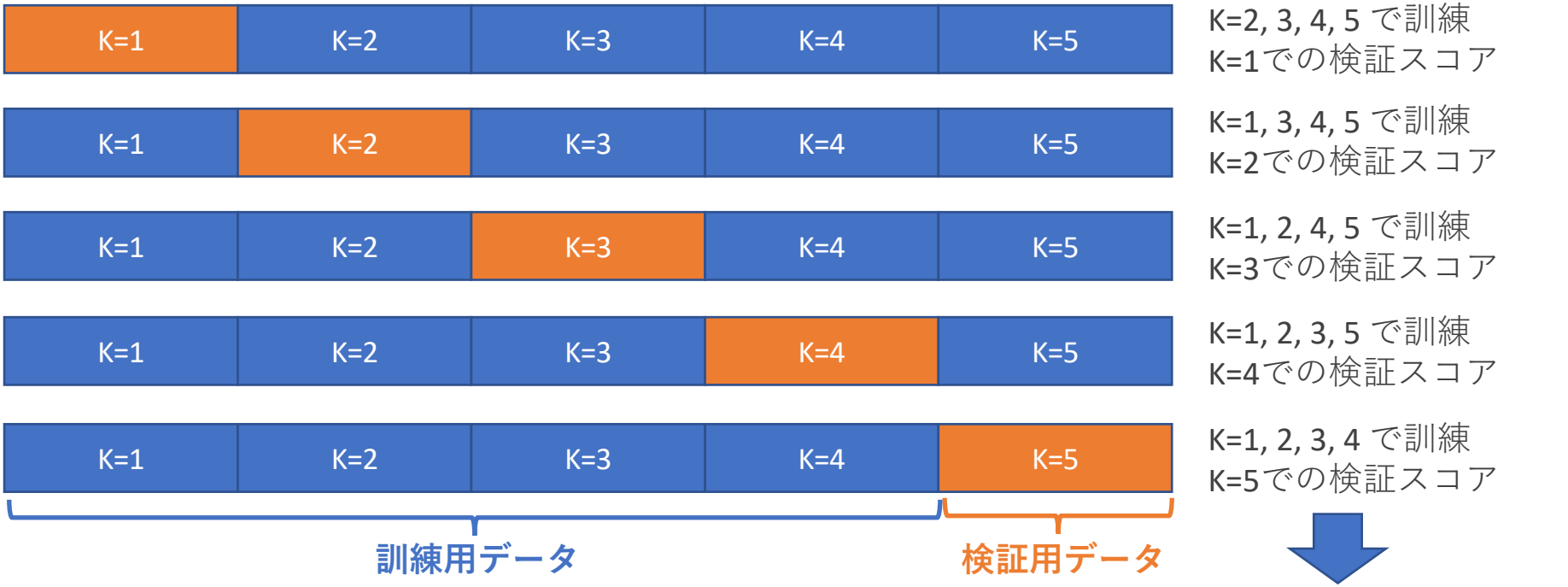
5．k個の予測誤差の平均を計算します.

ステップ5で計算される予測誤差平均をCVスコアと呼びます.

CVスコアは汎化誤差（未知データに対する誤差）の推定値であり，これを最小にするλは過学習・未学習を起こさない丁度良い値であると期待されます.

交差検証

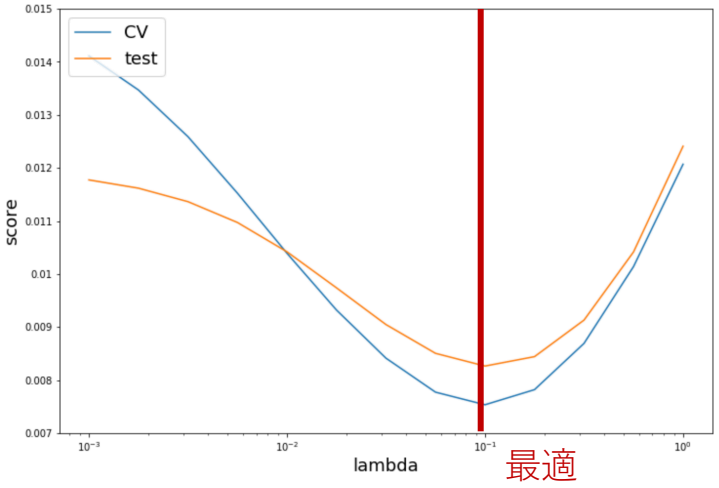
5-交差検証の実行イメージ



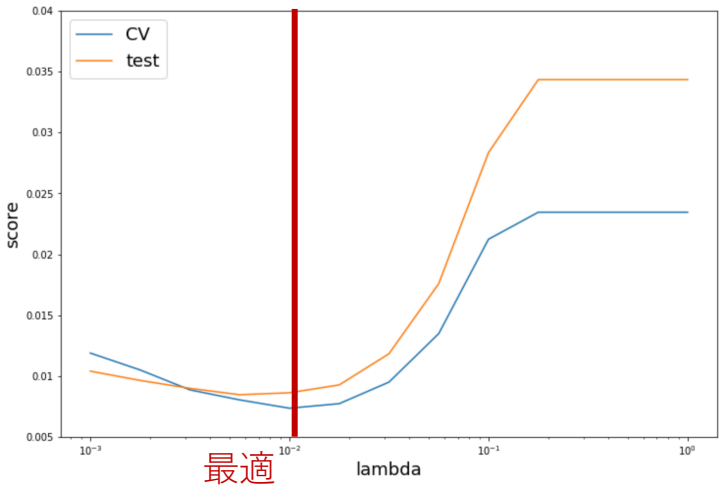
交差検証の実行例

リッジ回帰，LASSO回帰で交差検証を実行した例です．

リッジ回帰



LASSO回帰



正則化係数（lambda）に対しRMSEのCVスコア（青）とテストスコア（橙）をプロットしています．

いずれもCVスコアとテストスコアが共通の傾向を示していて，
CVスコアをもとに正則化係数を決定すれば良いテスト精度が達成されることが分かります．