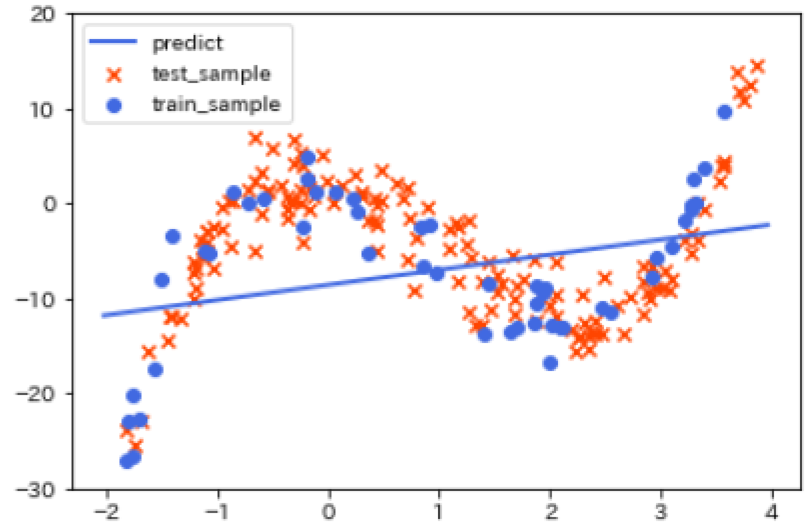
未学習と過学習

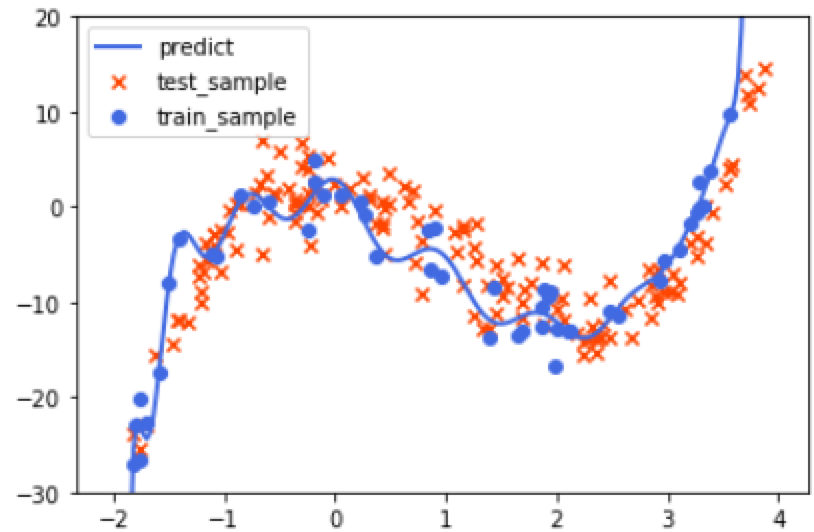
- 線形回帰

  モデルの表現能力が低いため、
  データの複雑さを表現できない
  - 未学習、学習不足



- (20次)多項式回帰

  モデルの複雑さに対して、
  データが不足しているため、
  訓練データを学習しすぎている
  - 過学習



正則化

最小化する目的関数に正則化項を加えて、
パラメータの値に制約を設けることで過学習を抑制する

$$J(\boldsymbol{W}) = \frac{1}{2}\sum_i \left(y_{\mathrm{true}}^{(i)} - y_{\mathrm{pred}}^{(i)}\right)^2 + \text{正則化項}$$

- L2正則化
  - $\boldsymbol{W}$の値が大きくなりすぎないようにする

$$J(\boldsymbol{W}) = \frac{1}{2}\sum_i \left(y_{\mathrm{true}}^{(i)} - y_{\mathrm{pred}}^{(i)}\right)^2 + \frac{\lambda}{2}\|\boldsymbol{W}\|_2^2$$

- L1正則化
  - Wの値が大きくなりすぎないように、
    かつ値を持つ要素を少なくする。

$$J(\boldsymbol{W}) = \frac{1}{2}\sum_i \left(y_{\mathrm{true}}^{(i)} - y_{\mathrm{pred}}^{(i)}\right)^2 + \lambda\|\boldsymbol{W}\|_1$$

※λ：正則化定数（制約の強さ）

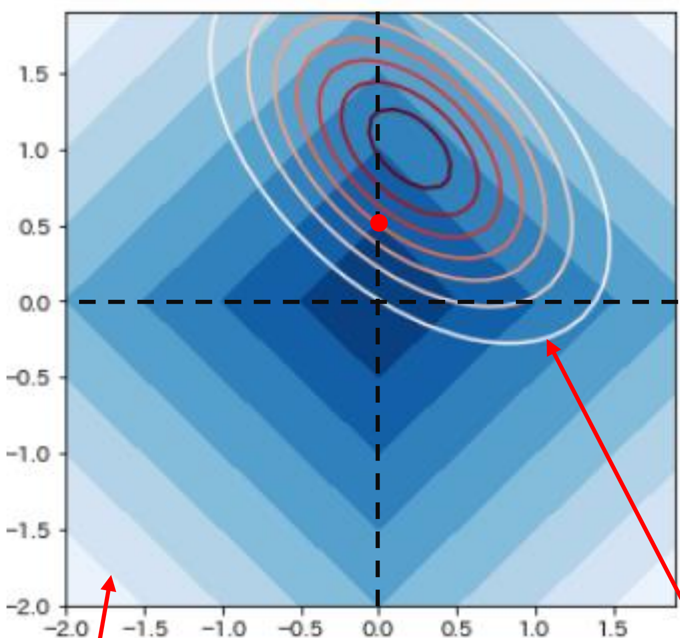
正則化のイメージ

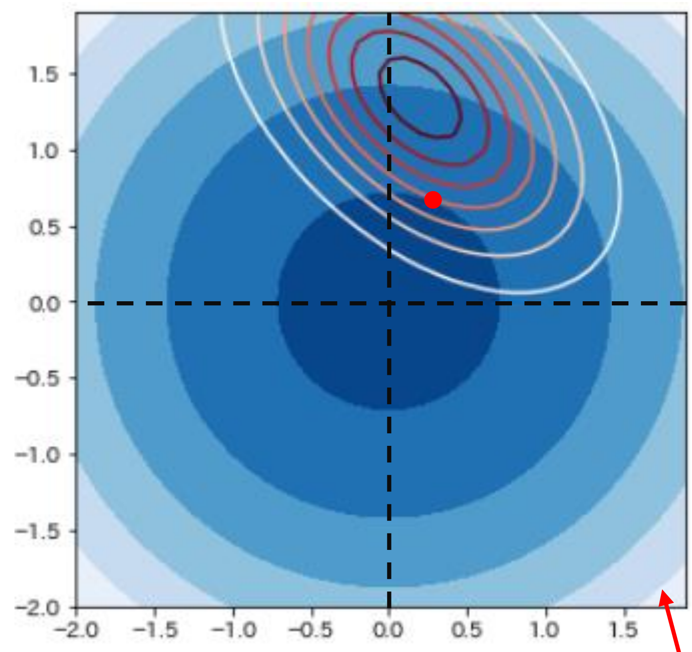
L1正則化

正則化を強めると、
軸上で接する可能性が高くなる

L2正則化

正則化を強めると値は小さくなるが、
軸上で接する可能性は低い





L1ノルムの等高線　目的関数の等高線　L2ノルムの等高線

色が濃い程良い値

次元削減

特徴量（説明変数）の次元を減らすことで、
以下のようなメリットがある。

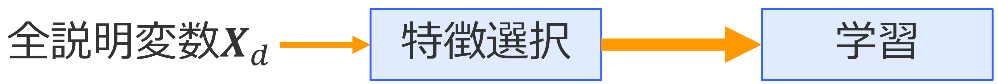
- 過学習を抑制できる
- 計算時間を軽減できる
- (3次元以下であれば)可視化できる

【主な方法】

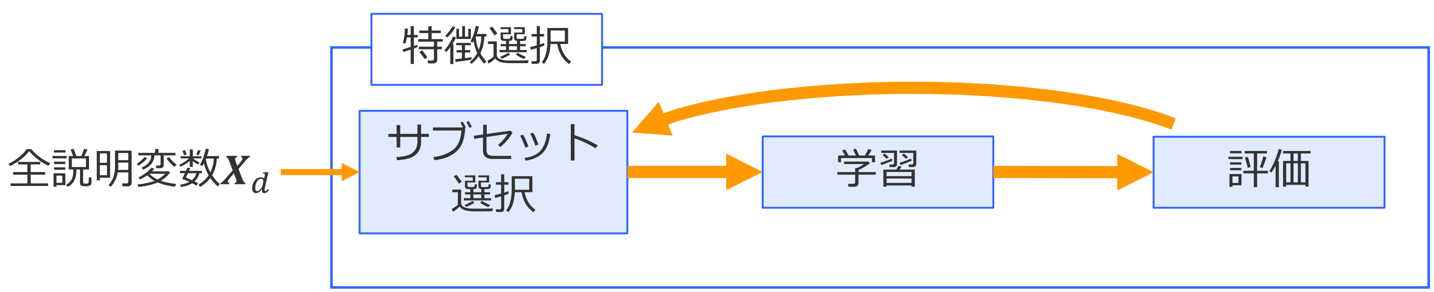
- 特徴選択
  - 特徴量全体から、一部の特徴量のみを選ぶ。
- 特徴抽出
  - 特徴量全体から、低次元の特徴部分空間を生成する。

特徴選択

- フィルター法
  説明変数と目的変数の関係性（相関など）に基づいて
  特徴選択を行う方法



- ラッパー法
  説明変数のサブセットを実際に学習し、
  評価関数がより良くなるサブセットを選択する方法



ラッパー法の例(SBS)

逐次後退選択(Sequential Backward Selection : SBS)

1. $k \leftarrow d$とする。（dは特徴量全体$\boldsymbol{X}_d$の次元数）

2. $l = 1, \ldots, k$について、以下を実行する。
   ① $\boldsymbol{X}_k$からl番目の特徴量を除いたものを$\boldsymbol{X}_l^-$とする。

   ② $\boldsymbol{X}_l^-$に関して学習・分類（回帰）を行い、評価関数Jを計算する。

3. 評価関数Jが最大（最小）となるlに関して、$\boldsymbol{X}_{k-1} \leftarrow \boldsymbol{X}_l^-$とする。

4. kが目的とする次元数になるまで、2・3を繰り返す。

特徴抽出

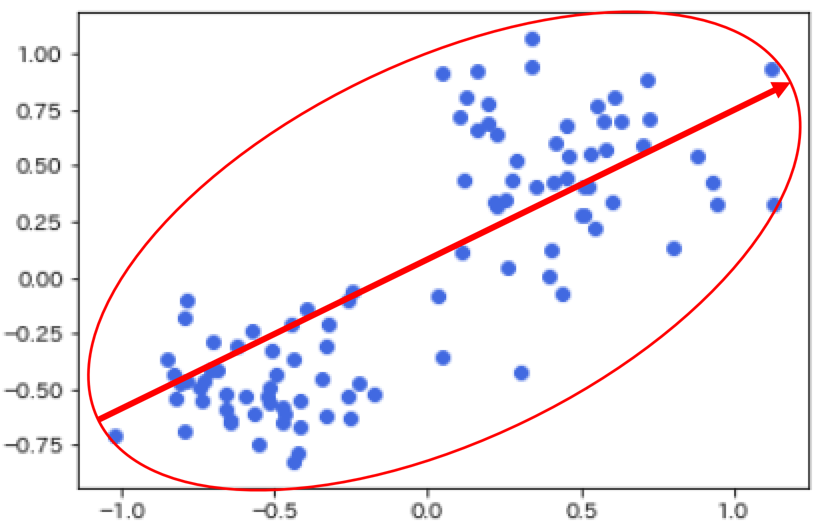
多次元の特徴量を要約するような低次元の特徴を新たに作成する

（例）

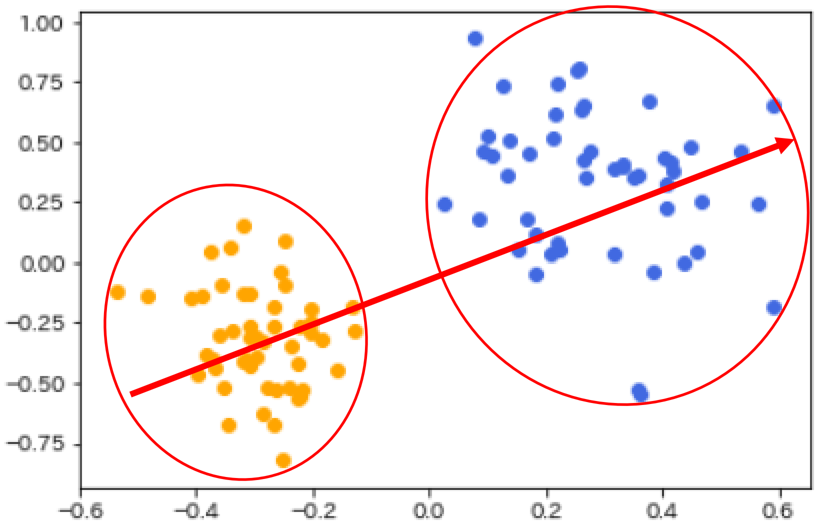
<u>直近20回のゴルフのスコア</u> ⇒ <u>平均スコア、分散</u>

20次元　　　　　　　　　　2次元

主成分分析（PCA）



線形判別分析（LDA）



データセットの分割（ホールドアウト法）

元のデータセット

- パラメータの学習：○
- 学習したパラメータの評価、ハイパーパラメータの調整：×

  元のデータセットに依存⇒過学習の可能性有

訓練データセット	テストデータセット

- パラメータの学習：○
- 学習したパラメータの評価、ハイパーパラメータの調整：○
- 調整したハイパーパラメータの評価：×

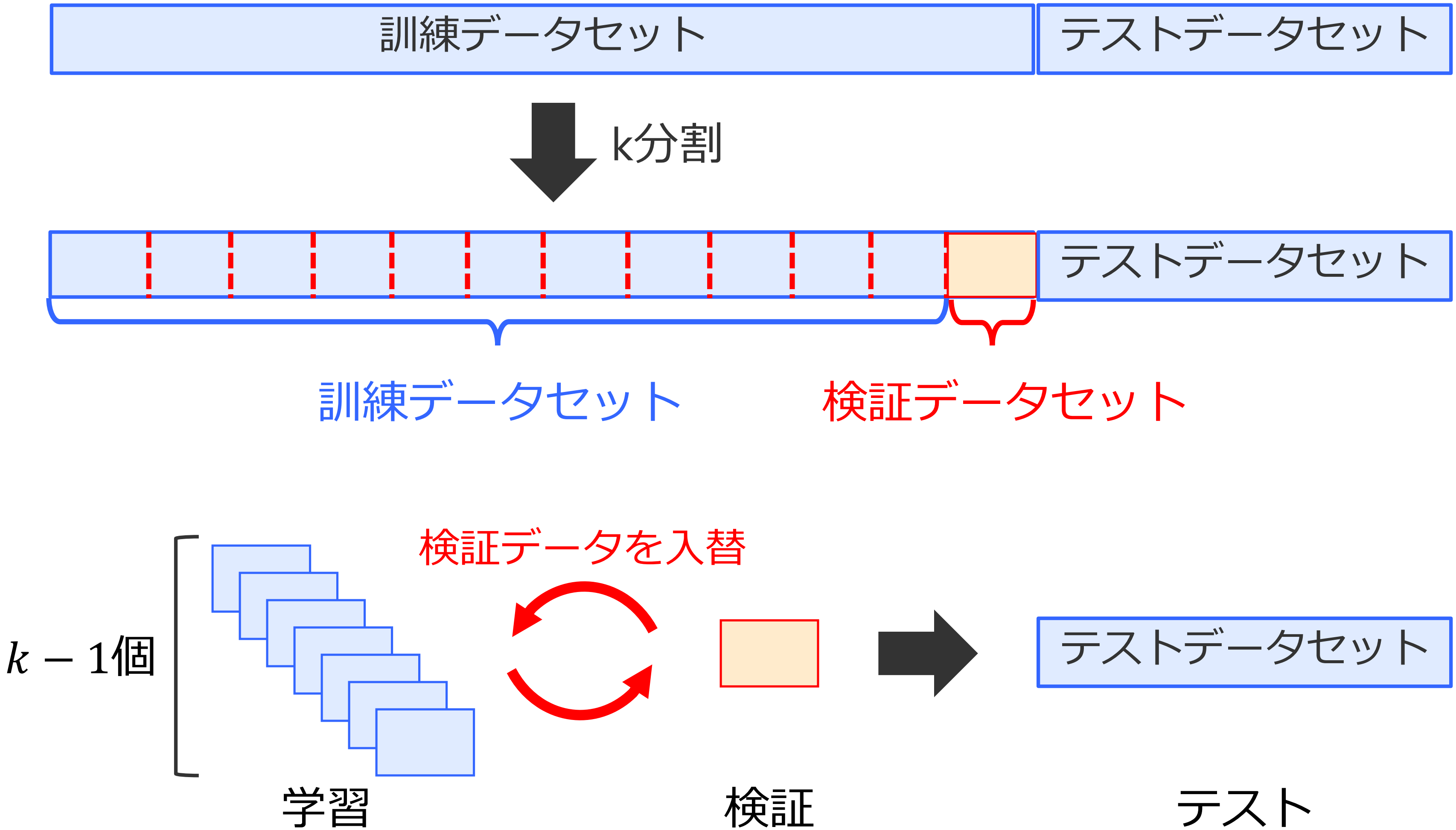
  テストデータセットに依存

訓練データセット	検証データセット	テストデータセット

- パラメータの学習：○
- 学習したパラメータの評価、ハイパーパラメータの調整：○
- 調整したハイパーパラメータの評価：○

※ハイパーパラメータ：学習とは別に、手で調整するパラメータ（正則化定数など）

k分割交差検証


訓練データセット
テストデータセット
k分割
テストデータセット
訓練データセット
検証データセット
検証データを入替
$k-1$個
テストデータセット
学習
検証
テスト