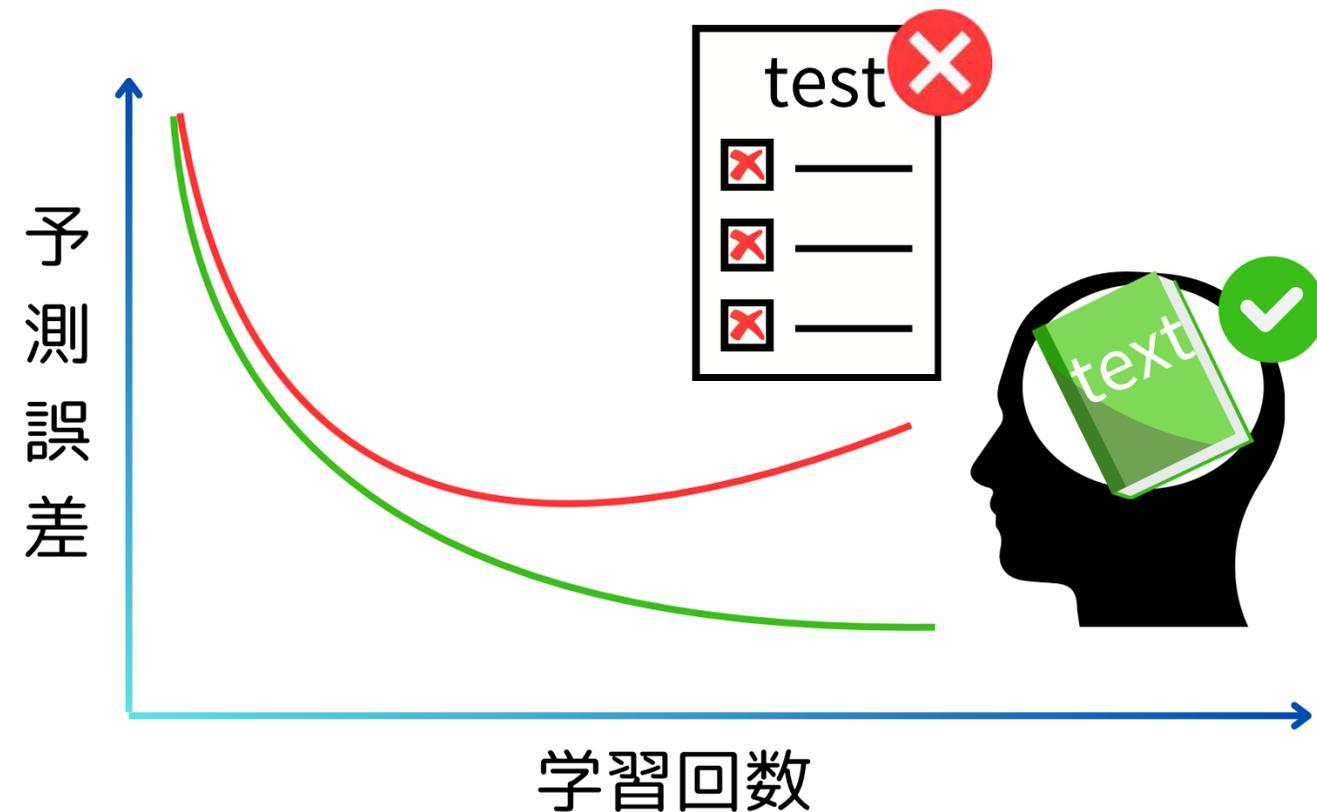


DS164 : 過学習とは何か、それがもたらす問題について説明できる



過学習とは、モデルの学習回数や複雑化に応じて訓練データの予測誤差が減少する一方、テストデータの予測誤差が増加し未知のデータに対する予測性能が低い状態を指します。



訓練データ・テストデータともに予測誤差が大きい状態は未学習と呼ばれます。

過学習の対策と確認方法

過学習の対策方法としては、以下が挙げられます。

- 学習データ量を増やす
- モデルの複雑性を減らす
- 正則化を使用する
- 早期終了を行う

過学習の確認方法として、モデルのパラメータを変更した時に、**学習データに対する精度が良くなっているのに対して、テストデータに対する精度が低下もしくは変わらない場合は過学習を疑います。**

バイアスとバリエーション

過学習に関係する概念であるバイアスとバリエーションを紹介します。

バイアス（偏り：Bias）とバリエーション（分散：Variance）

バイアスとは、**予測値と真の値とのズレ**
バリエーションとは、**推定値のばらつき**

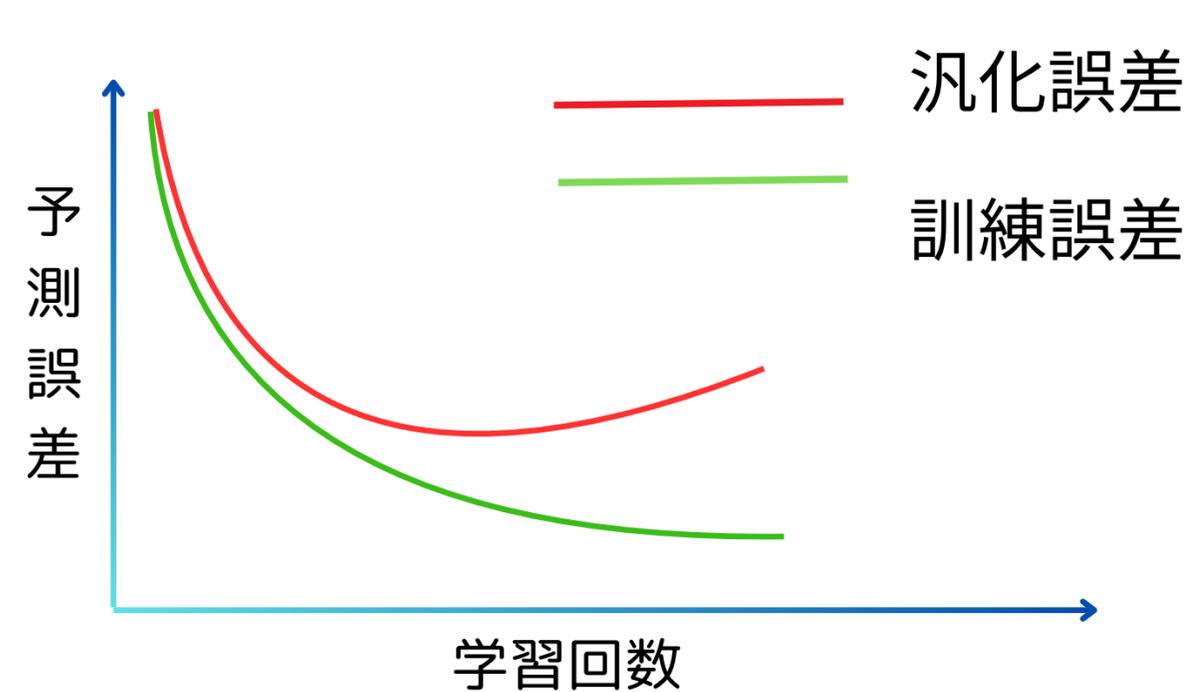
過学習になりやすいのは、低いバイアス・高いバリエーション。

理想的なモデルは、低いバイアス・低いバリエーション。

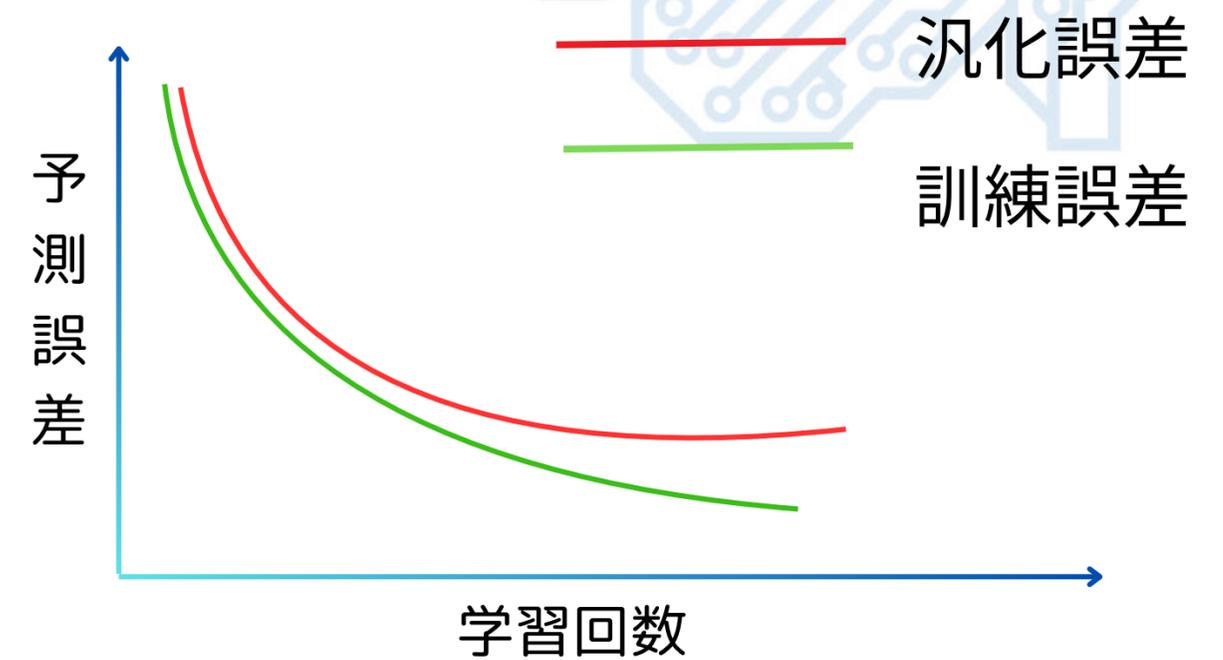
また、バイアスとバリエーションはトレードオフの関係にあります。

確認問題

① 次のうち過学習の状態を表すグラフとして適切なものはどれか。



A

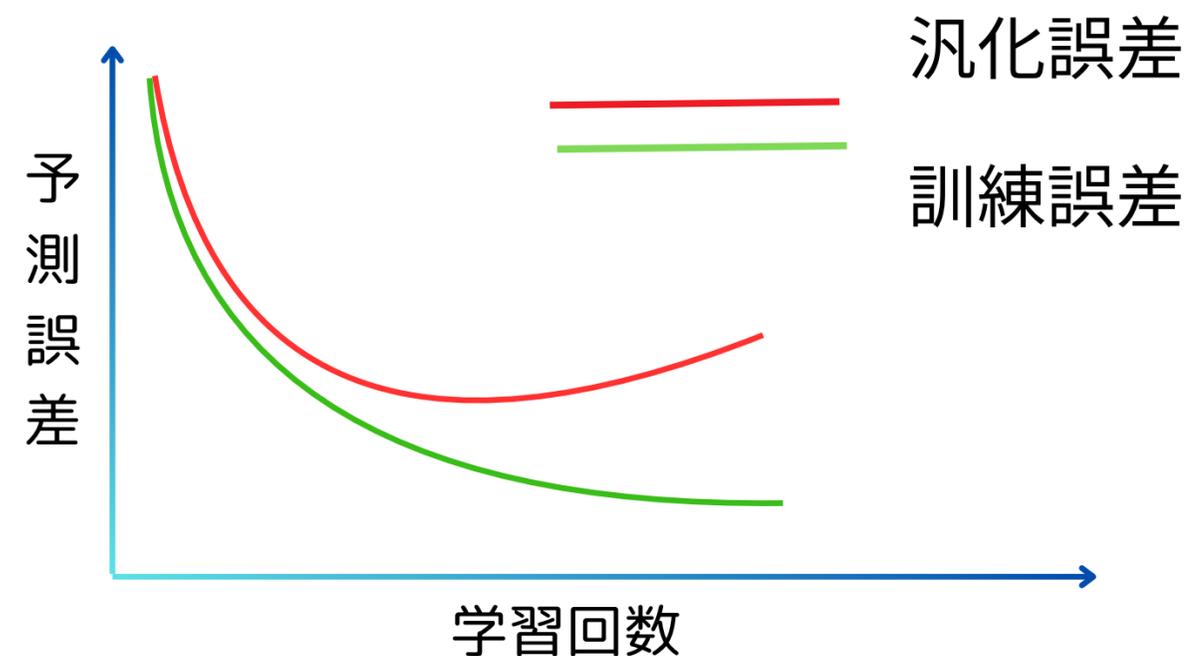


B

解説は次のスライド▶

確認問題 解説

① 次のうち過学習の状態を表すグラフとして適切なものはどれか。



A

Aが過学習したグラフです。訓練誤差が減少する一方で、汎化誤差は途中から上がっています。

DS161 : 機械学習の手法を3つ以上知っており、概要を説明できる

DS163 : 「教師あり学習」「教師なし学習」の違いを理解している

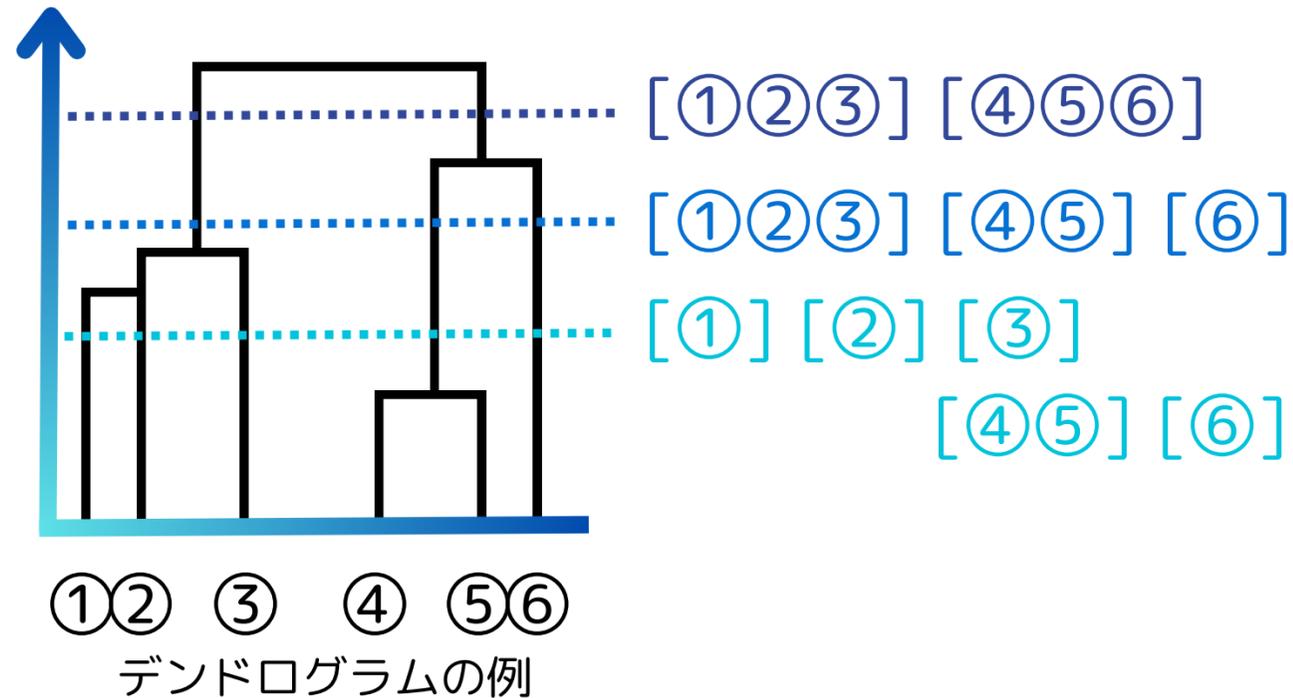
教師あり 学習	回帰	線形回帰
	分類	ロジスティック 回帰
教師なし 学習	次元削減	主成分分析
	クラスタ リング	k-means法

教師あり学習は、入力データと予測する正解データのペア（教師データ）がある機械学習手法です。教師データが量的データの場合は回帰、質的データの場合は分類タスクになります。

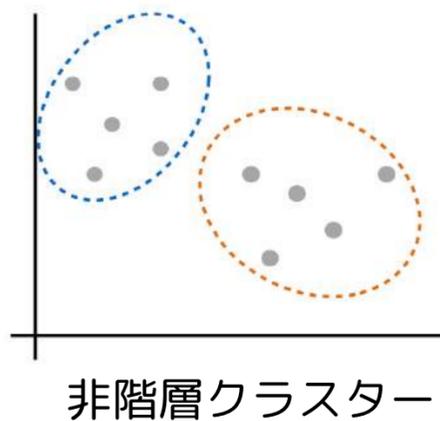
教師なし学習は、教師データがない機械学習手法であり、入力データをグループ分けするクラスタリングや情報量を圧縮する次元削減の手法があります。

DS228 : 階層クラスター分析と非階層クラスター分析の違いを説明できる

DS229 : 階層クラスター分析において、デンドログラムの見方を理解し、適切に解釈できる



特徴量の類似データ同士でグループ分けする手法「クラスター分析」には、類似データ同士を順にまとめる階層クラスター分析と階層構造を持たずにクラスター数をあらかじめ決める非階層クラスター分析があります。



階層構造を表したデンドログラムからクラスター数は変更できます。