東京大学
THE UNIVERSITY OF TOKYO

BY NC ND

データマイニング入門 第11回

2018年度

学習目標

- 非線形性、基底関数の導入について理解する
  - 多項式
- 過学習と正則化について理解する
  - L2正則化
- 特徴量エンジニアリングについて理解する
  - 特量選択、特徴量作成
- モデル評価と選択について理解する
  - 交差検証、混同行列
- BiasとVarianceについて理解する

モデル評価と選択

変数の多項式による線形回帰+正則化では多項式の次数や正則化項の係数がハイパーパラメータとなっていた

どのようにこれらのハイパーパラメータを決定するのがよいか？

異なるハイパーパラメータの値を持つ学習モデルのうちどれを採用するのがよいか？

クロスバリデーション

- データセットを訓練データセットと検証データセットに分ける
  - 例えば7:3の割合
- ハイパーパラメータの設定ごとに訓練データセットで学習をして学習モデル（仮説関数）をえる
- 検証データセットについて誤差が最小になるような仮説関数を与える学習モデルを採用する

モデル評価と選択

k-foldクロスバリデーション

- データセットをk個のサブセットに分割する
  - 例えばk=10
- 各学習モデルについて
  - k-1個のサブセットで学習して残り1個のサブセットで学習モデルを評価する
  - k回学習を繰り返しその誤差の平均をモデルの推定汎化誤差とする
- 推定汎化誤差が最小となる学習モデルを採用する

Validation	Train	Train	Train	誤差$_1$
Train	Validation	Train	Train	誤差$_2$
…	…	…	…	
Train	Train	Train	Validation	誤差$_k$

leave-one-out

- (k=データ数)としたk-foldクロスバリデーション

モデルの実際の運用時の性能評価については別途テストデータで評価することに注意

- 訓練データ+バリデーションデータ+テストデータ

判別の予測精度の検証

混同行列

- 2値分類の例
- 例えばh(x)=0.5以上を1と予測する

	予測（y=1）	予測（y=0）
正解ラベル（y=1）	**TP (True Positive)**	**FN (False Negative)**
正解ラベル（y=0）	**FP (False Positive)**	**TN (True Negative)**

- 正解率 (accuracy)：TP+TN/(TP+TN+FP+FN)
- 適合率 (precision)：TP/(TP+FP)
- 再現率 (recall)：TP/(TP+FN))
- F値：precision・recall/(precision+recall)
  - PrecisionとRecallはトレードオフの関係にある

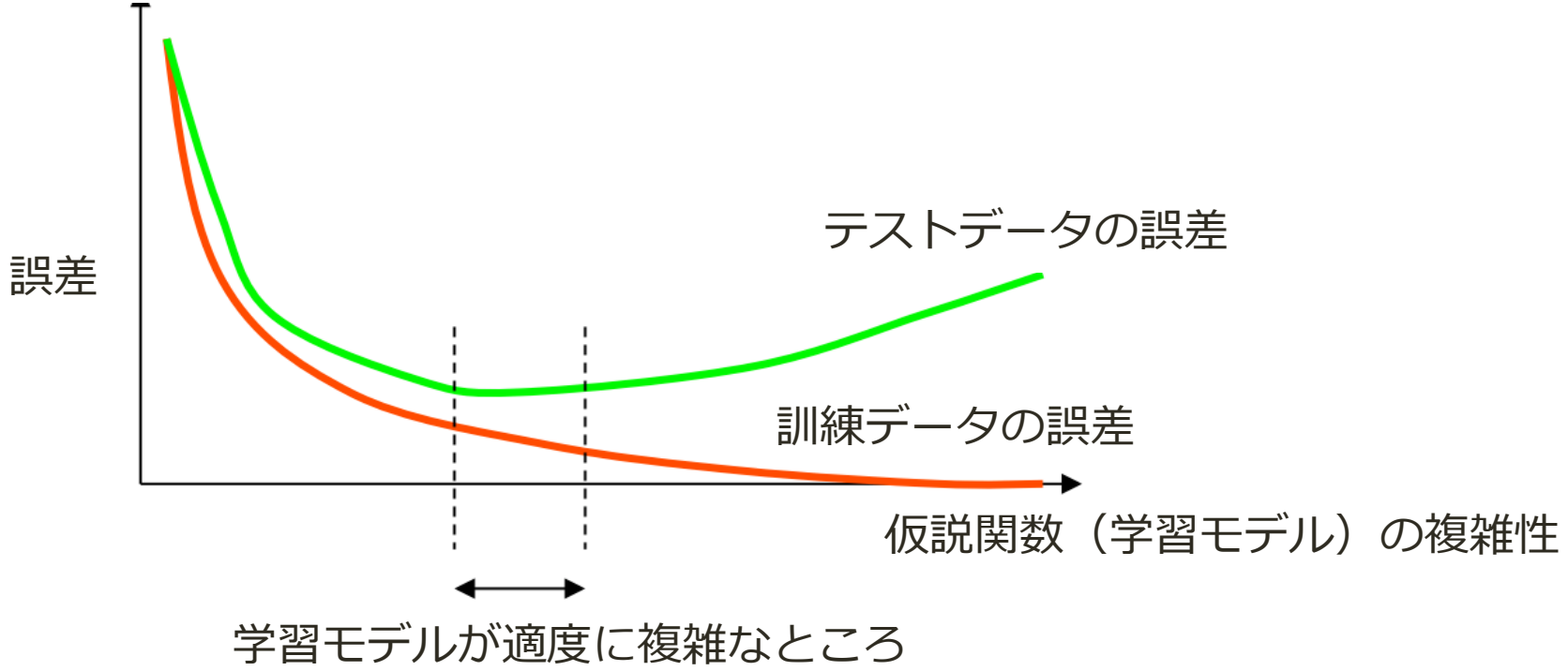
BIASとVARIANCE

Overfitting

- 学習モデルが訓練データに適合しすぎて汎化できていない(High Variance)

Underfitting

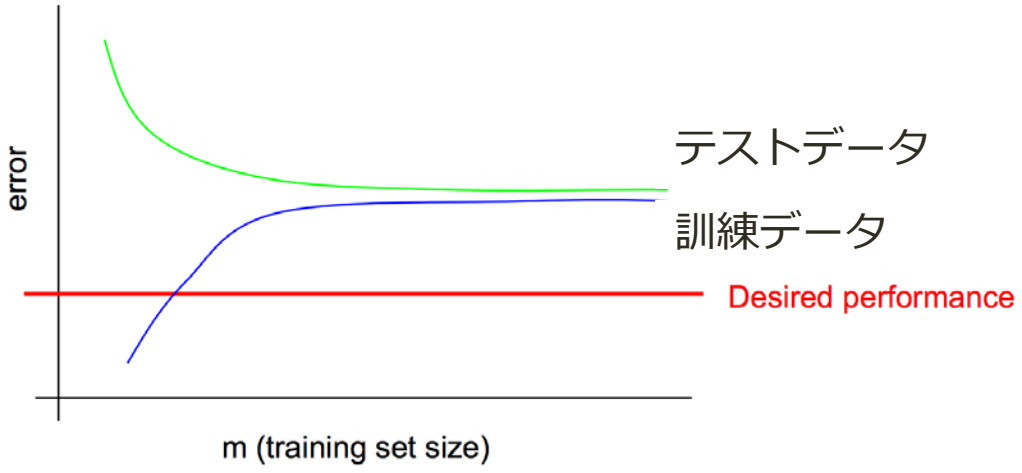
- 学習モデルが入力と出力の関係を訓練データから十分に学習できていない (High Bias)



BIASとVARIANCE

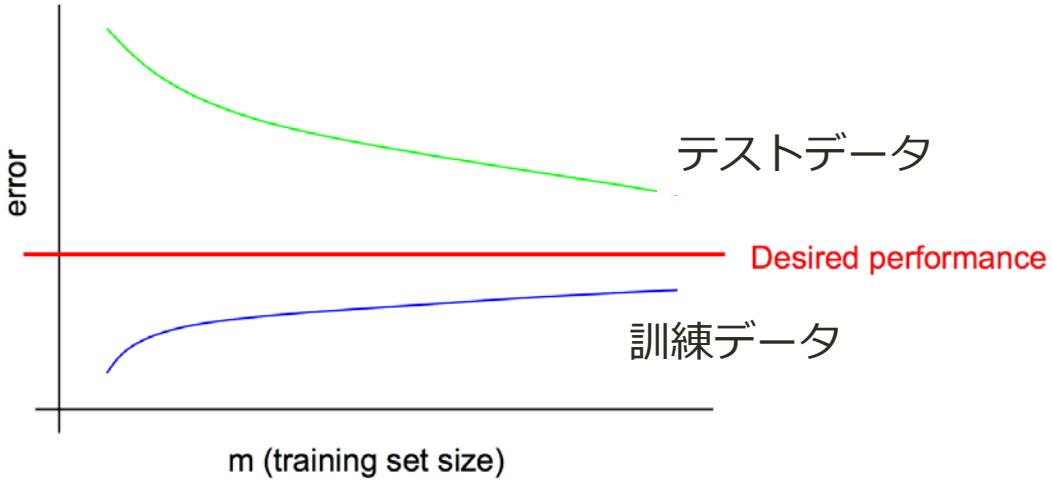
学習曲線：データセットの大きさとBias/Variance

High Bias



訓練データ自体の誤差が大きく
訓練データとテストデータの誤差の差が小さい

High Variance



訓練データとテストデータの誤差の差が大きい

BIASとVARIANCE

学習モデルのテストデータに対する誤差の期待値を考えると以下が成り立つ

- (ノイズ)+(Biasの二乗)+(Variance)

BiasとVarianceは以下のように解釈できる

- Bias
  - 学習アルゴリズムに起因する誤差
    - 異なるデータセットで学習したモデルの平均化した出力と真の出力の差
- Variance
  - 学習モデルのばらつきに起因する誤差
    - 異なるデータセットで学習したモデルの平均化した出力とあるデータセットで学習したモデルの出力の差

学習モデルの調整

High Biasの時

- モデルを複雑する
  - 特徴量を増やす
  - 多項式であれば次数を増やす
  - 正則化項係数を小さくする

High Varianceの時

- モデルを簡潔にする
  - 特徴量を減らす
  - 多項式であれば次数を減らす
  - 正則化項係数を大きくする
- データセットを増やす

学習モデルの調整

パラメータ推定の最適化手法

- 最急降下法より早く収束することが多い
- モデルが複雑であればこれらを使うほうがよい

- ニュートン法
  - コスト関数が2回微分可能であること
  - ヘッセ行列の逆行列計算
- 準ニュートン法
  - ヘッセ行列の逆行列を近似
    - BFGS法
    - L-BFGS法
- 共役勾配法

scikit-learnのLogisticRegressionで使用できる最適化
ニュートン法、L-BFGS法、SAG・SAGA（確率的勾配降下法の拡張）

学習モデルの調整

確率的・ミニバッチ勾配降下法の最適化

- 学習率の調整
  - モメンタム [Sutton 86]
  - AdaGrad [Duchi 11]
  - AdaDelta [Zeiler 12]
  - RMSProp [Hinton]
  - Adam [Kingma and Ba 15]

参考： https://arxiv.org/pdf/1609.04747.pdf

特徴量選択

特徴量数nが膨大（例えばデータ数m<<特徴量数n）な場合は特徴量数を減らし、モデルの複雑性を抑えた方がよい

前向き探索
- 特徴量セットは空集合
- 特徴量を1つずつ増やしクロスバリーデーションなどで評価することを繰り返す
- 特徴量数が所定の数に達したら探索を止める

後向き探索
- 特徴量セットはすべての特徴量集合
- クロスバリーデーションなどで評価し特徴量を1ずつ減らすことを繰り返す
- 特徴量数が所定の数に達したら探索を止める

この他にも出力に対する特徴量の情報量に基づくフィルター特徴量選択や正則化などがある

ラッソ回帰（L1正則化）は特徴量を疎にする特徴量選択と考えられる

特徴量エンジアリング

前処理
- 重複
- 外れ値

特徴量選択
- 前向き・後ろ向き探索
- 相関・情報量
  - 相関、AIC
- モデルベース
  - ランダムフォレスト、Lasso回帰（L1正則化）

特徴量作成
- 交差項、多項式
- 非線形変換
  - log, exp, など
- 次元縮約
- ドメイン知識

