

機械学習

2-4

工学院情報通信系
中原 啓貴

今日の内容

- データ前処理
 - トレーニングセットの事前調整
 - 特徴量の尺度
 - カテゴリデータの処理
- 過適合に対するパラメータチューニング
+ 上記内容の演習

欠測データについて

- 実データはサンプルの値が欠損している場合が多い
 - データ収集時の誤り
 - 不適切な測定
 - アンケート未回答
- 欠測値(missing value)をどう扱うか？
- pandas（データ読み込み処理ライブラリ）を利用
 - 欠測値を扱うライブラリを提供

欠測データの処理

- 欠測データを含む行（列）を削除
→dropnaメソッド
削除しすぎることが多く、有益な情報を失うことに
- 欠測値を補完（補完法: Interpolation）
 - 平均値補完
 - 中央値・最頻値も使われることが多い

カテゴリデータの処理

- 実データには数値以外のデータも含まれる
- **順序特徴量**
 - 順序付けや並べ替え可能
 - 服のサイズ(XL>L>M>S>XS)
- **名義特徴量**
 - 順序が定義できない
 - 服の色(Blue, Red, Green)
- scikit-learn等の機械学習ライブラリは便宜上、ラベルに数値を割り当てる
 - 名義特徴量に何らかの工夫が必要
 - 人手によるミスの可能性も

名義特徴量の one-hotエンコーディング

- Natural order エンコーディング
 - Blue → 0, Red → 1, Green → 2
 - 順序がないはずなのに、順序を考慮した学習が行われる
- One-hot エンコーディング
 - Blue → 100, Red → 010, Green → 001
 - 名義特徴量毎にダミー特徴量(上記では2個)を作成
 - scikit-learnのOneHotEncoderクラス
 - pandasのget_dummies関数だと、文字列値を持つ列のみ処理してくれる！

データセットの分割

- トレーニング(訓練)データとテスト(評価)データに分割
 - 訓練データのみに過適合(→低い汎化性能)を避けるため
 - scikit-learnのtrain_test_split関数で提供
 - ランダムに分割
 - 割り当ての比率 : 6:4, 7:3, 8:2
 - 大量のデータ(深層学習, 例1400万枚)では99:1もあり
 - Train + Test + Validation に分ける方法も
 - 訓練データとテストデータを入替えて評価する
- 交差検証(Cross-Validation)も多用されている
- 後日**k分割交差検証(k-fold cross-validation)**を紹介

特徴量のスケーリング

- 決定木学習(ランダムフォレストも)を例外として重要な処理
- 例) パーセプトロン学習の勾配降下法
 - 特定のサンプル(変数)のダイナミックレンジが大きい場合...
- 正規化(normalization): 特徴量を $[0, 1]$ にスケーリング
- 標準化(standardization): 平均値0, 標準偏差1となるように変換
 - こちらを使うことが多い(多くの線形モデルが初期パラメータを乱数で初期化するため)
- scikit-learnでどちらも利用可能

有益な特徴量の選択

- L_1 正則化によるスパース化
 - 非ゼロな特徴量が有益なものと言える
- 逐次特徴選択アルゴリズム
 - greedyなアルゴリズムで繰り返し探索
- ランダムフォレストによる特徴抽出
 - よい認識性能→重要な特徴を選択していることに

過学習(Overfitting)の傾向

- 対義語として未学習(underfitting)
- 訓練データとテストデータの結果が乖離する
(多くの場合, 訓練データのみ精度が良い)
 - 学習曲線(検証曲線)をプロットするとよい
- 原因
 - 学習データが少ないとき
 - モデルが問題に対して複雑な時
- 対策
 - データを増やす (非現実的...)
 - 正則化を追加
 - (ハイパーパラメータの少ない) 単純なモデルに変更
 - スパース化などによりデータの次元数を減らす

検証曲線による 過学習と未学習の判定

- モデルのパラメータを変化させて検証
- 例: ロジスティック回帰の正則化パラメータCの検討
正則化項に対する定数の逆数であることに注意
 - C が小さいとき, 正則化が強くなる
 - C が大きいとき, 正則化を緩める

実習

https://github.com/HirokiNakahara/MachineLearning_Lecture/tree/master/lec2_4

- 特徴量の選択と処理
- 有益な特徴量の選択
- 検証曲線による判定