

データリークについて

信寛昭

0. 参考にした論文

原題 : Leakage and the reproducibility crisis in machine-learning-based science

著者 : Sayash Kapoor , Arvind Narayanan

掲載誌: Patterns, Volume 4, Issue 9, 8 September 2023, 100804

DOI/URL: <https://doi.org/10.1016/j.patter.2023.100804>

目次

1. データリークとは
2. データリークが問題になる理由
3. データリークの主な種類
4. ソフトウェア工学で起きやすい例
5. 論文でチェックすべき点
6. まとめ

1. データリークとは

モデルの学習・前処理・評価に、
**実際の予測時には利用できない
情報が入り込む問題**

例：

問題集の中に、テストで使う問題が混じっている

→ **高得点でも、実力とは限らない**

2. データリークが問題になる理由

- テスト結果が**実際より**良く見える
 - 未知のデータでは**性能が下がる**
 - 論文の結論が**過大評価**される
 - 他の人では**再現できない可能性**が生じる
- **モデルが本当に予測できているか判断できなくなる**

3. データリークの主な種類

同じデータで訓練と評価を行う

→ **一度見たデータで実力を測ってしまう**

不適切な特徴量の使用

→ **答えをほぼ直接示す情報**を特徴量として使う

時間的リーク

→ **予測時点より後に得られた情報を使う**こと

訓練サンプルとテストサンプルが独立していない

→ **同じリポジトリのデータ**を用いる

4. ソフトウェア工学で起きやすい例

- 時系列を無視してデータを分割する
 - 同じプロジェクトのデータが訓練・テスト両方に入る
 - バグ修正後の情報を、バグ予測に使ってしまう
 - IssueやPull Requestの未来情報を特徴量に含める
 - 重複・類似コードが訓練・テスト両方に入る
- …etc

5. 論文でチェックすべき点

- 訓練データとテストデータの分け方
- 前処理がデータ分割後に行われているか
- 未来の情報を使っていないか
- 特徴量が目的変数の答えに近すぎないか
- 同じ対象が訓練・テスト両方に入っていないか

6. まとめ

データリークは、モデルが本来使えない情報を使ってしまう問題

- テスト性能を**実際より高く見せてしまう**
- ソフトウェア工学でも**時系列・重複・特徴量設計で起こりやすい**
- 論文を読むときは、**データ分割・前処理・特徴量・評価方法を確認する必要がある**

高い精度＝正しい研究とは限らない