



DMIA Sport 2019.2.Leakages



Leakages

Что такое лики

Лики в решении задач

Лики в создании соревнований

Пробивание лидерборда

Борьба с ликами

Что такое лики (data leakages)



Что такое лик

Зависимость в данных, которой не существует в реальности

1. Одинаково влияет на прогнозы на CV и тесте
обычно мы называем это просто "ошибкой"
2. Влияет на них по-разному
обычно мы называем "ликом" именно это

Альтернативные формулировки

- if any other feature whose value would not actually be available in practice at the time you'd want to use the model to make a prediction, is a feature that can introduce leakage to your model

<http://dataskeptic.com/blog/episodes/2016/leakage>

- when the data you are using to train a machine learning algorithm happens to have the information you are trying to predict

<https://insidebigdata.com/2014/11/26/ask-data-scientist-data-leakage/>

Где они появляются

1. В создании соревнования
организаторы допустили ошибку, формируя обучающую и тестовую выборки
 1. Специально
например, чтобы захантить к себе тех, кто умеет находить лики
 2. Неспециально
обычно происходит именно так
2. В решении задачи
аналитик допустил ошибку, работая с данными
3. В работе с чувствительными данными
возможен реверс-инжиниринг обфусцированных данных

Более точное определение

	В создании соревнования	В решении соревнования
"Ошибка"	Одинаково влияет на качество на тесте и в реальной жизни	Одинаково влияет на качество на кросс-валидации и тесте
Лик	Обманчиво улучшает качество на тесте	Обманчиво улучшает качество кросс-валидации

Лики в решении задач



Ли́ки в решении задач

1. Примеры
2. Общие закономерности

Лики в решении задач. Примеры

1. Неправильное разбиение обучение/контроль

Задача: прогнозируем продажи в магазинах

Трейн: 2016 – 2018 года

Тест: 2019 год

Ещё пример: множества магазинов из трейна и теста не пересекаются

Лики в решении задач. Примеры

2. Несоответствие структуры данных в обучении и контроле

Задача: прогнозируем продажи в магазинах

Тренин: только проданные товары

Тест: все товары (в том числе те, где таргет равен нулю)

Лики в решении задач. Примеры

3. Препроцессинг данных, завязанный на таргет

Задача: прогнозируем продажи в магазинах

- Что хотим: убрать выбросы в целевой переменной
`target = np.clip(target, *np.percentile(target, [1, 99]))`
- Что хотим: сделать отбор признаков
Используем ли для этого контроль?

Лики в решении задач. Выводы

1. Разбиение обучение/контроль должно имитировать разбиение трейн/тест
2. Препроцессинг, завязанный на таргет, не должен настраиваться по контролю

Лики в создании соревнований



Лики в создании соревнований

1. Примеры
2. Общие закономерности

Лики в создании соревнований. Примеры

1. Порядок строк

Есть корреляция между номером строки и целевой переменной

Лики в создании соревнований. Примеры

2. Информация в ID

Например,

ID = datetime_of_new_user_creation + random_string

Лики в создании соревнований. Примеры

3. Метаданные файлов (особенности записи)

picture_1.jpg

picture_2.jpg

picture_3.jpg

Название файлов/время их создания/etc коррелирует с целевой переменной

Лики в создании соревнований. Примеры

4. Неконсистентный формат выгрузки данных

Например,

- 2019-01-01 для класса 1, 01-01-2019 для класса 0
- "nan" для класса 1, "None" для класса 0

Лики в создании соревнований. Примеры

5. Информация, скрытая в "форме" тестовой выборки

Задача: прогнозирование продаж в магазинах

Правильно составленная тестовая выборка:

- декартово произведение (дат, магазинов, товаров)
- никакой информации, не известной на самую раннюю в тесте дату

Ещё: задачи на "попарные" сравнения объектов, задачи с графами

Лики в создании соревнований. Примеры

5. "Лишняя" информация, скрытая в данных

Задача: классификация новостей

Проблема: даты в тексте новости

Некоторая информация может "вредить" тому решению задачи, которое хотят получить организаторы

Лики в создании соревнований. Примеры

6. Реверс-инжиниринг обфускации данных

Что сделали: нормировали целочисленный признак

Что можно сделать: частично обратить операцию

Ещё примеры: анонимизация, рандомизация данных

Лики в создании соревнований. Примеры

7. Дополнительные источники данных

Внешние данные увеличивают шанс возник

Лики в создании соревнований. Закономерности

1. Лик может быть связан со значениями предоставленного признака
2. Лик может быть связан со способом формирования тестовой выборки
3. Лик может быть связан со способом записи данных
4. Дополнительные внешние данные увеличивают шанс возникновения лика

Лики в создании соревнований. Выводы

1. Не должно быть лика информации из теста в трейн
2. Значения целевой переменной не должны быть связаны со способом формирования тестовой выборки
3. Не должно быть лика информации из будущего в прошлое
4. Обфускация данных должна быть необратимой

Пробивание лидерборда (leaderboard probing)



Пробитие лидерборда

Тестовые данные могут состоять из трех частей:
public, private, ignored.

Отправляя сабмиты, можно выяснить:

- Какие строчки входят в паблик
- Каковы в них значения целевой переменной
- Получить информацию, полезную для прогноза прайвата (иногда)

Пробитие лидерборда

1. Какой части принадлежат строчки:
public, private, ignored?

Если изменение ответа на строчке влияет на скор - это
строчка из паблика

Пробитие лидерборда

2. Выяснение среднего значения таргета в паблике

$$-L * N = \sum_{i=1}^N (y_i \ln C + (1 - y_i) \ln (1 - C))$$

$$-L * N = N_1 \ln C + (N - N_1) \ln (1 - C)$$

$$\frac{N_1}{N} = \frac{-L - \ln (1 - C)}{\ln C - \ln (1 - C)}$$

Пробитие лидерборда

3. Использование найденных значений в качестве дополнительных данных

Задача: прогнозируем продажи в магазинах

Паблик: 2019 год

Прайват: 2020 год

Тренд за 2019 год может быть весьма полезен.

Ещё: можно добавить объекты в обучающую выборку.

Пробитие лидерборда

4. Использование найденных значений в качестве дополнительных данных

Задача: прогнозируем продажи в магазинах

Паблик: часть троек (день, магазин, товар) в 2019 и 2020 году

Прайват: оставшиеся тройки

Продажи за конкретную дату могут быть не менее полезны...

Борьба с ликами



Способы поиска ликов

1. Делать EDA и исследовать "удивительные" находки подробнее
2. Критически оценивать метрики качества моделей
3. Чтобы отловить баги при выгрузке, можно сравнивать модель на выборках, выгруженных разными скриптами/людьми

Как уменьшить вероятность возникновения лика

1. При решении задач можно делать разбиение на обучение и контроль отдельным скриптом до тренировки моделей и препроцессинга данных
2. Разбивать данные на обучение/контроль также, как разбиты трейн/тест
3. Помнить о том, какие лики существуют

Ссылки

1. Лики в решении задач:
<http://www.alfredo.motta.name/cross-validation-done-wrong/>
2. Лики в создании соревнования:
<https://insidebigdata.com/2014/11/26/ask-data-scientist-data-leakage>
https://www.cs.umb.edu/~ding/history/470_670_fall_2011/papers/cs670_Tran_PreferedPaper_LeakingInDataMining.pdf
 1. Метаданные
<https://www.kaggle.com/c/the-icml-2013-whale-challenge-right-whale-redux/discussion/4865>
 2. Задачи с попарными сравнениями
<https://necromuralist.github.io/kaggle-competitions/posts/data-leakages/>
3. Пробивание лидерборда:
<https://www.kaggle.com/olegtrott/the-perfect-score-script>