Утечка данных (Data Leakage)

Утечка данных в машинном обучении имеет место, когда часть признаков, присутствующих в обучающем наборе, оказываются недоступными на этапе промышленной эксплуатации модели, когда она должна делать предсказания для реальных данных из потока, генерируемого бизнес-процессом.

Типичным примером утечки является включение в число входных признаков обучающей выборки поля целевой переменной. Очевидно, что значения выходной переменной известны только на обучающем наборе и принципиально недоступны на реальных данных, поскольку именно их модель и должна предсказывать в процессе практической работы. Такое включение целевого поля в число входных как правило происходит непреднамеренно из-за желания аналитика использовать для обучения модели как можно больше информации.

Другим источником утечки может быть ошибка при организации сбора и подготовки данных для обучения модели. Например, если в обучающем наборе используется информация о ценах конкурентов, историю которых несложно собрать для построения модели, а оперативный сбор сведений о текущих ценах в процессе промышленной эксплуатации является затратным и не всегда возможным. В результате происходит их утечка.

Независимо от характера и причин утечек данных, они всегда приводят к одному и тому же негативному результату: точность предсказаний модели на обучающем наборе данных оказывается выше, чем при промышленной эксплуатации. Причины этого интуитивно понятны: в результате утечки одного или нескольких признаков количество информации, используемой моделью для предсказания на реальных данных, оказывается меньше, чем на обучающих, что негативно сказывается на точности.

В некоторых случаях неправдоподобно высокая точность модели на обучающих данных также может служить сигналом возникновения утечки. Это происходит, если в набор входных полей включается целевое поле, в результате чего между входом и выходом модели возникает сильная корреляция или даже функциональная зависимость, которую модель и научится воспроизводить. Но поскольку в реальной обстановке целевое поле окажется недоступным, зависимость, воспроизводимая моделью окажется смещенной относительно зависимости в реальных данных, что резко снизит предсказательную эффективность модели.

Иными словами, утечка данных приводит к следующему. Разработчики модели, оценив ее точность на обучающих данных как высокую, разворачивают ее у заказчика. Последний, начав практическую эксплуатацию модели, обнаруживает, что она работает совсем не так хорошо, как ему обещали.

Ситуация усугубляется тем, что само наличие утечки и ее влияние на точность модели сложно обнаружить, так как модель может делать недостаточно точные, но вполне правдоподобные предсказания. Дополнительной сложностью при борьбе с утечкой является то, что разрабатывают и эксплуатируют модель чаще всего разные люди в различных локациях, коммуникации и обратная связь между которыми по проблеме утечки ограничены.

Утечка данных является сложным, многоаспектным явлением, поэтому процесс ее формирования и воздействия на работу ML-моделей трудно формализуемы. Это препятствует выработке общих подходов к ее обнаружению и снижению негативного влияния. Тем не менее, в аналитическом сообществе выделены несколько видов утечки данных и выработаны рекомендации для борьбы с проблемой и их последствиями.

В настоящее время выделают следующие виды утечек данных:

1. **Утечка признаков (feature leakage)** — представляет собой основной вид утечки, при котором признаки обучающего набора данных, использующиеся в качестве входных при обучении модели, становятся недоступными на этапе ее практического использования.
2. **Целевая утечка (target leakage)** — связана с использованием информации, которую несет целевая переменная, в качестве входной при обучении модели. Простейшим случаем такой утечки и является включение целевого признака (или признака, связанного с ним функционально) в набор входных переменных обучающей выборки. Это аналогично случаю, если бы обучающемуся вместе с экзаменационным вопросом давали и правильный ответ, что позволило бы ему создать у экзаменатора впечатление прекрасного знания предмета и получить завышенную оценку. В литературе такие данные часто называют **нелегитимными**, а действия модели сравнивают с мошенническими.

3. **Утечка обучающего/тестового множества (train-test contamination — TTC)** — связана с процедурой предобработки данных, которая может включать нормализацию, масштабирование, сглаживание, квантование, подавление выбросов, восстановление пропущенных значений и т.д. В результате, хотя качество обучающих данных улучшается, но в них вносятся определенные искажения, которые учитываются моделью в процессе обучения. В то же время реальные данные не подвергаются предобработке и будут отражать иные зависимости, чем обучающие. Как следствие, модель потеряет в точности на реальных данных.

В настоящее время основным подходом к решению проблем, связанных с утечкой данных, является организационный. Он включает следующий набор правил и рекомендаций:

1. Не использовать в качестве входных признаки, для которых высока вероятность утечки. Недостатком подхода является сложность выявления таких признаков, а также сокращение количества информации, используемой в процессе обучения, что снижает его эффективность. Альтернативным вариантом является оценка признаков пропорционально риску их утечки так, чтобы признаки с более высокой вероятностью утечки в наименьшей степени использовались при обучении. Это позволит сделать работу модели более устойчивой к утечкам.

2. Использовать все доступные в обучающих данных признаки, но в случае утечки принять меры к восстановлению утекших признаков. При этом затраты на дополнительный сбор данных могут оказаться слишком большими.
3. Чтобы избежать TTC-утечки, следует оставить валидационный набор данных без предобработки.
4. Не допускать утекания обучающих примеров из тестового множества в обучающее, поскольку это завысит оценку точности модели по итогам обучения.
5. Оценивать, не получилась ли точность модели по результатам обучения слишком высокой, чтобы быть правдой.

Эти рекомендации не являются универсальными, и как правило проблема решается индивидуально для каждого случая. Тем не менее риск возникновения утечки данных и ее возможные последствия нельзя оставлять без внимания, поскольку использование плохо работающей модели при поддержке принятия решений в бизнесе может привести к большим потерям.

Более подробно с проблематикой утечки данных можно ознакомится в статье Утечка данных в машинном обучении.