Лекция 3. Обработка пропущенных значений. Кодирование категориальных признаков

Практикум по программированию, 5 семестр

Иван Евгеньевич Бугаенко,
ассистент каф. ПМиФИ

Из-за каких проблем в данных невозможно обучить модели?

Если в данных есть пропуски или категориальные признаки, то подавляющее большинство моделей обучиться не сможет

Подходы по обработке пропущенных значений

- интерпретация пропущенных значений
- статистические подходы
- машинные методы
- заполнение на основе соседних значений

Интерпретация пропущенных значений

Пропуски можно заполнять значениями, которые отвечают бизнес-правилам признакового пространства

Пример 1. Пропуск в категориальных данных

num_feature1	cat_feature1	num_feature2	target
56.09	a	7	0
79.6	NaN	4	1

num_feature1	cat_feature1	num_feature2	target
56.09	a	7	0
79.6	'NaN'	4	1

Интерпретация пропущенных значений

Пример 2. Пропуск в вещественных признаках

num_feature1	cat_feature1	num_feature2	target
56.09	a	7	0
79.6	b	NaN	1

num_feature1	cat_feature1	num_feature2	target
56.09	a	7	0
79.6	b	0 / -1 / 9999	1

Статистические подходы

Пропуски можно заполнять значениями, которые вычисляются на основе значений признаков, составляющих выборку

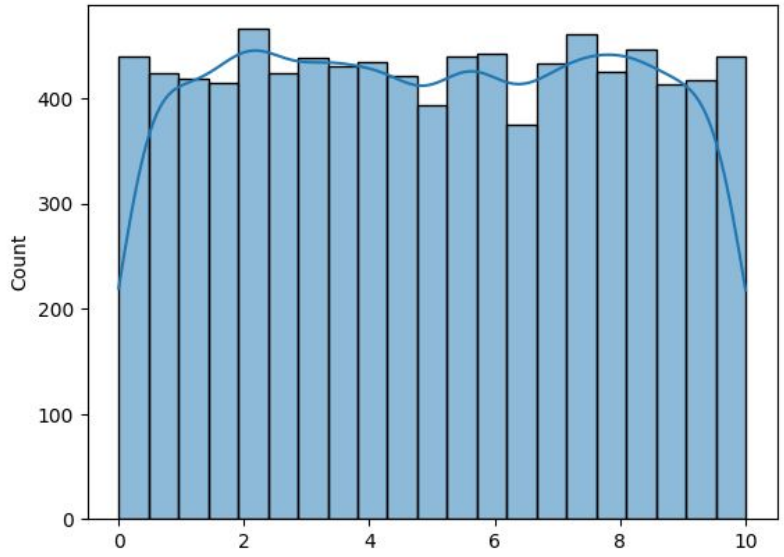
Пример 3. Пропуск в категориальных данных

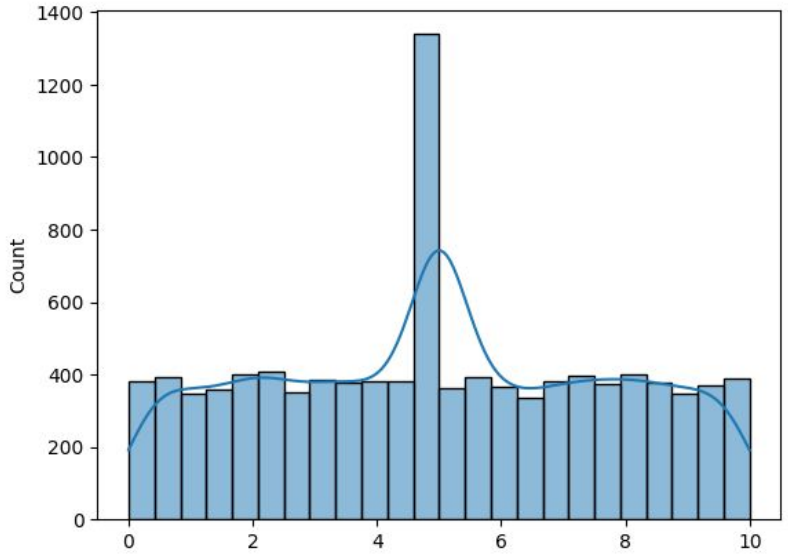
num_feature1	cat_feature1	num_feature2	target
NaN	a	7	0
86.23	b	11	1
26.76	a	15	0
66.09	a	6	1

num_feature1	cat_feature1	num_feature2	target
mean / med	a	7	0
86.23	b	11	1
26.76	a	15	0
66.09	a	6	1

Статистические подходы

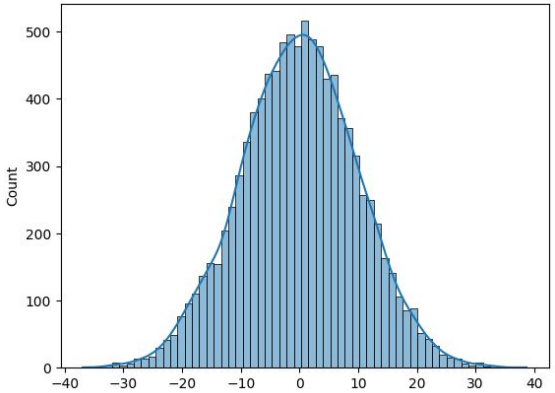
НО! Заполнение некоторой константой может кардинально изменить распределение признака. Пусть есть датасет из 10000 записей, в этом датасете в некотором признаке A 10% пропусков. Рассмотрим распределение признака до заполнения средним и после.



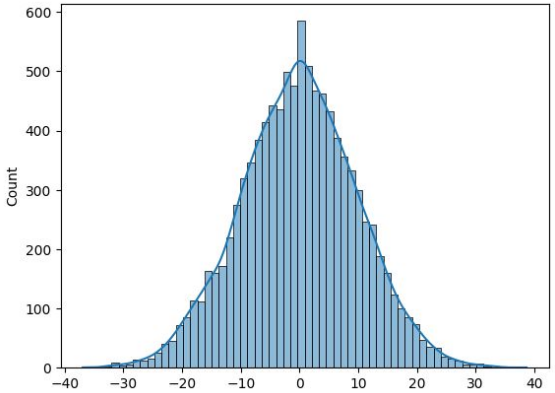


Статистические подходы

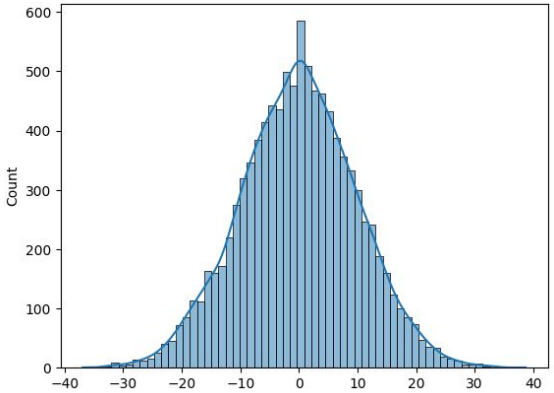
НО! Также может возникнуть ситуация, когда невозможно выбрать наиболее подходящее значение среди отобранных



Исходные данные



Заполнение средним



Заполнение медианой

Статистические подходы

Чтобы отсеять некорректное заполнение пропусков и отобрать наилучшее, можно исследовать **KL-дивергенцию** (дивергенция Кульбака-Лейблера):

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Дивергенция – это безразмерная метрика, она показывает, насколько распределение q похоже на распределение p числом от 0 до бесконечности (чем ближе метрика к 0, тем больше распределения могут быть вложены друг в друга)

Пайплайн по проверке статистических методов

1. для каждого признака, который содержит пропуски, определяется множество значений, которыми можно эти пропуски заполнить
2. для каждого значения строятся 2 гистограммы: до заполнения пропусков и после, а также считается KL-дивергенция для двух получившихся распределений
3. на основе внешнего вида распределений и дивергенции принимается решение о выборе значения

Машинные методы

Заполнять пропуски можно также при помощи методов машинного обучения. Благодаря таким алгоритмам можно улавливать нелинейные зависимости в структуре пропусков.

Основными алгоритмами являются **MICE** (Multiple Imputation by Chained Equations) и **KNNImputer**.

Суть методов заключается в том, чтобы обучить модель предсказывать столбцы с пропусками на основании оставшихся признаков (без информации о целевом).

Советы по заполнению пропусков

- если признак содержит больше 50-70% пропусков, то его можно посчитать неинформативным и отбросить
- если в строке целевое значение содержит пропуск, то такая строка подлежит удалению (восстанавливать целевое значение нельзя)
- строки, которые имеют большую концентрацию пропусков, в результате заполнения могут стать аномальными, потому они подлежат удалению (!)
- при разделении выборки на обучение и тест может возникнуть ситуация, что пропусков в train-части нет в некотором признаке, в данном случае все равно необходимо анализировать пропуски, так как они будут в test-части
- не стоит отдавать предпочтение только статистическим методам или модельным: лучше заполнить пропуски и тем, и тем способом, а потом проверить метрики качества моделей, обученных на таких данных

Какие бывают виды категориальных признаков?

Виды категориальных признаков

- равнозначные категории
- категории с отношением порядка
- смешанные категории

Виды кодировщиков категориальных признаков

Метод	Плюсы	Минусы
One-Hot	Простота, интерпретируемость	Проклятие размерности для признаков с >10 категориями
Label	Простота, не увеличивает размерность	Произвольный порядок, плохо для линейных моделей
Target	Учитывает связь с целью, сохраняет размерность	Риск переобучения, чувствительность к настройкам