Машинное обучение
семинар 5

10 марта 2020

Кирилл Лунев

Работа с пропущенными значениями

Пропущенные значения: мотивация

На практике нередко встречаются данные с пропусками:

- Сломался сенсор
- Ошибка парсинга
- Респондент не ответил на вопрос
- Не всё залогировалось
- ...

Пропущенные значения: типы пропусков

Missing completely at random (MCAR) — вероятность пропуска не зависит от значений наблюдаемых и пропущенных данных:

$$P(M|X_{observed}, X_{missing}) = const$$

Пример: у случайной части пациентов не измерили вес

Пропущенные значения: типы пропусков

Missing at random (MAR) — вероятность пропуска зависит от значений наблюдаемых, но не от значений пропущенных данных:

$$P(M|X_{observed}, X_{missing}) = f(X_{observed})$$

Пример: измеряем вес только у пациентов с повышенным давлением

Пропущенные значения: типы пропусков

Missing not at random (MNAR) — вероятность пропуска зависит от значений наблюдаемых и пропущенных данных:

$$P(M|X_{observed}, X_{missing}) = f(X_{observed}, X_{missing})$$

Пример: измеряем вес только у пациентов, страдающих ожирением

Пропущенные значения: методы восстановления

Удалить объекты с пропущенными значениями

Плюсы:

- Не портим данные

Минусы:

- Можем сильно уменьшить выборку
- Можем внести смещение (если пропуски не MCAR)

Пропущенные значения: методы восстановления

Удалить признаки с пропущенными значениями

Плюсы:

- Не портим данные

Минусы:

- Можем потерять полезный сигнал

Пропущенные значения: методы восстановления

Заменить специальным значением (-1, 0, 99999, ...)

Плюсы:

- Не теряем данные

Минусы:

- Вносим смещение

Пропущенные значения: методы восстановления

Заменить средним/медианой/модой

Плюсы:

- Не теряем данные
- Учитываем известные данные

Минусы:

- Вносим смещение

Можно агрегировать по другим факторам

Пропущенные значения: методы восстановления

Last Observation Carried Forward (LOCF)

First Observation Carried Backward (FOCB)

№	…	User Id	Time	Missing feature
1	…	1	1	None
2	…	1	2	None
3	…	2	3	1
4	…	1	3.5	None
5	…	2	4	None
6	…	3	5	5
7	…	3	5.8	None
…	…	…	…	…

Пропущенные значения: методы восстановления

Last Observation Carried Forward (LOCF)

First Observation Carried Backward (FOCB)

№	…	User Id	Time	Missing feature	LOCF
1	…	1	1	None	None
2	…	1	2	None	None
3	…	2	3	1	1
4	…	1	3.5	None	1
5	…	2	4	None	1
6	…	3	5	5	5
7	…	3	5.8	None	5
…	…	…	…	…	…

Пропущенные значения: методы восстановления

Last Observation Carried Forward (LOCF)

First Observation Carried Backward (FOCB)

№	…	User Id	Time	Missing feature	LOCF	LOCF + FOCB
1	…	1	1	None	None	1
2	…	1	2	None	None	1
3	…	2	3	1	1	1
4	…	1	3.5	None	1	5
5	…	2	4	None	1	5
6	…	3	5	5	5	5
7	…	3	5.8	None	5	5
…	…	…	…	…	…	…

Пропущенные значения: методы восстановления

Last Observation Carried Forward (LOCF)

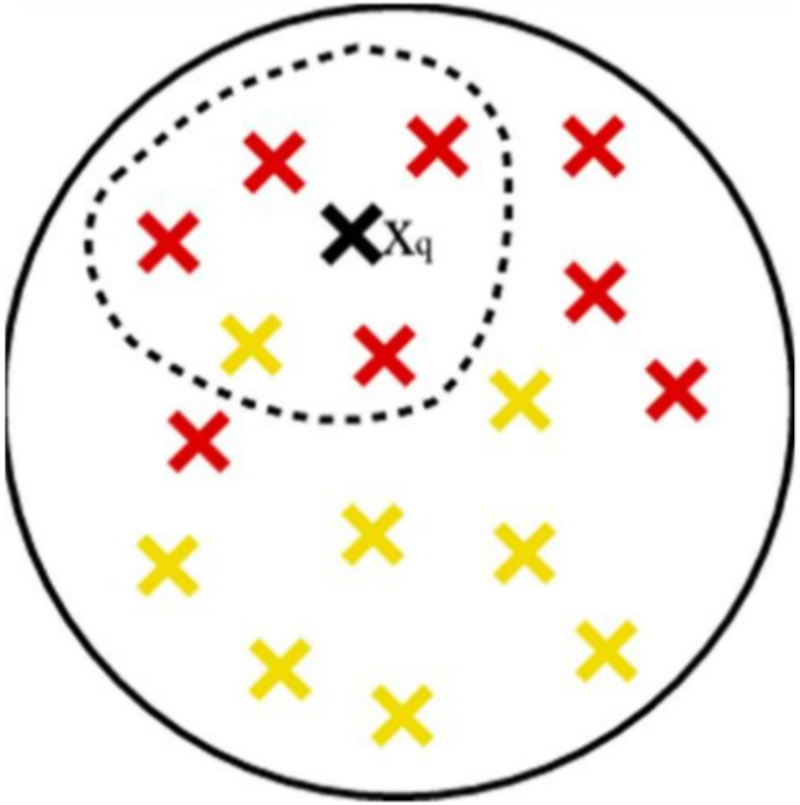
First Observation Carried Backward (FOCB)

№	…	User Id	Time	Missing feature	LOCF	LOCF + FOCB
1	…	1	1	None	None	8
2	…	1	2	None	None	8
3	…	1	3.5	None	None	8
4	…	1	8.5	None	None	8
5	…	1	10.2	8	8	8
6	…	2	3	1	1	1
7	…	2	4	None	1	1
…	…	…	…	…	…	…

Пропущенные значения: методы восстановления

K-NN

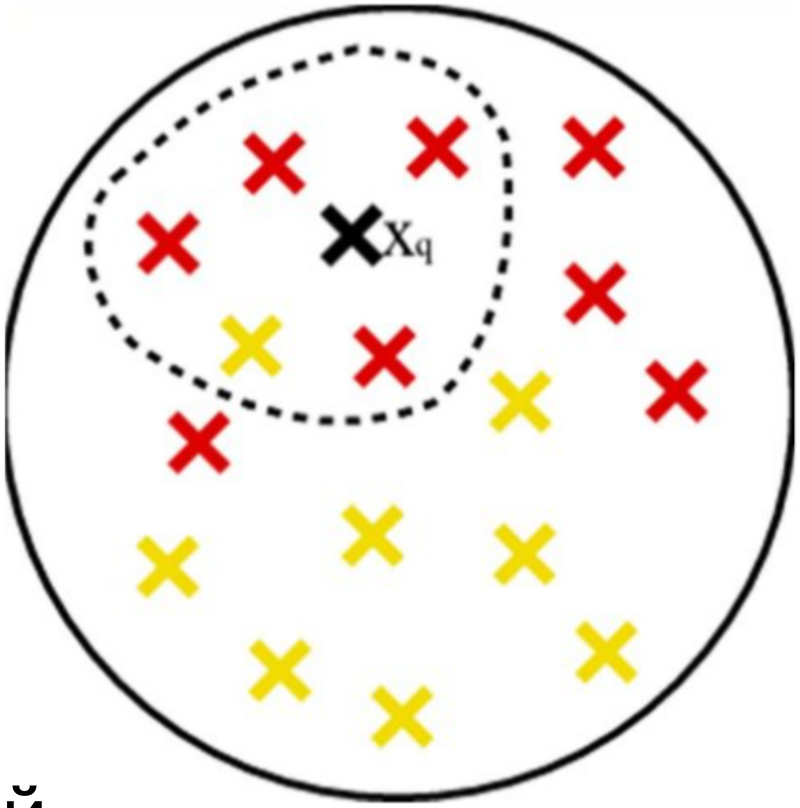
Предположение: близкие по известным признакам объекты близки в признаках с пропусками



- Для объекта с пропуском найдем ближайшие k без пропусков
- Заменим пропуск на взвешенное расстоянием среднее по соседям

Пропущенные значения: методы восстановления

K-NN



Плюсы:

- Высокая точность
- Обобщается на категориальные признаки

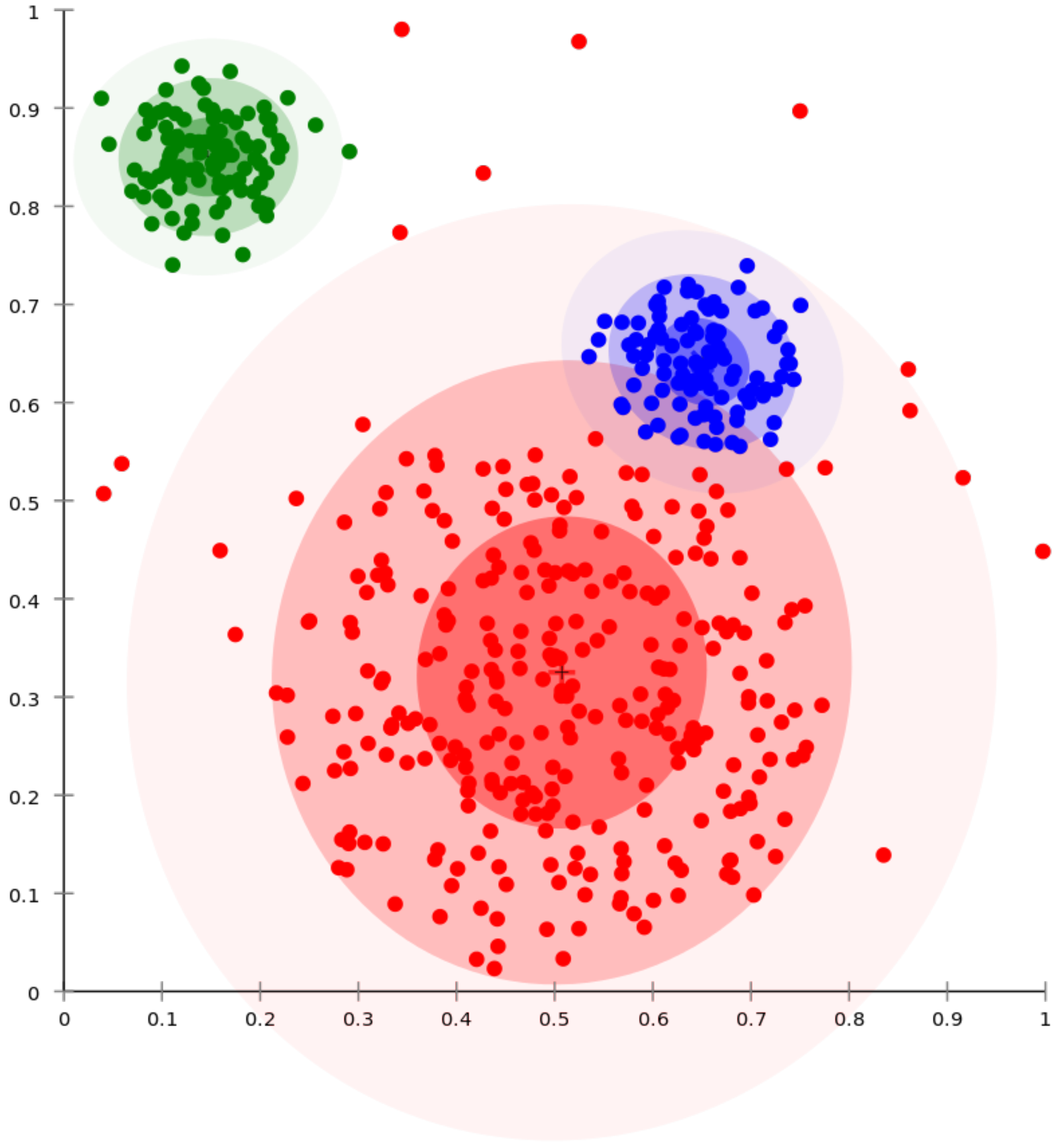
Минусы:

- Нужно настраивать: расстояние, число соседей
- Нужно много данных без или почти без пропусков

Пропущенные значения: методы восстановления

Кластеризация

- K-Means, Fuzzy K-Means, etc.
- Аналогично K-NN, но усредняем по объектам из кластера
- Нужно настраивать кластеризацию



Пропущенные значения: методы восстановления

Предсказание пропущенных значений

- Используем признак с пропущенными значениями как целевую переменную
- Обучаемся на объектах с известными значениями
- Предсказываем для объектов с неизвестными

Пропущенные значения: методы восстановления

Multiple Imputation by Chained Equations (MICE)

- Используем признак с пропущенными значениями как целевую переменную
- Обучаемся на объектах с известными значениями
- Предсказываем для объектов с неизвестными
- Повторяем предыдущие пункты для всех признаков с пропусками
- Повторяем предыдущий пункт до сходимости

Пропущенные значения: методы восстановления

Multiple Imputation by Chained Equations (MICE)

Плюсы

- Высокое качество

Минусы

- Нужно настраивать модели
- Может долго работать

Пропущенные значения: методы восстановления

Singular value decomposition (SVD)

Плюсы

- Используем все данные для восстанавливаемых значений

Минусы

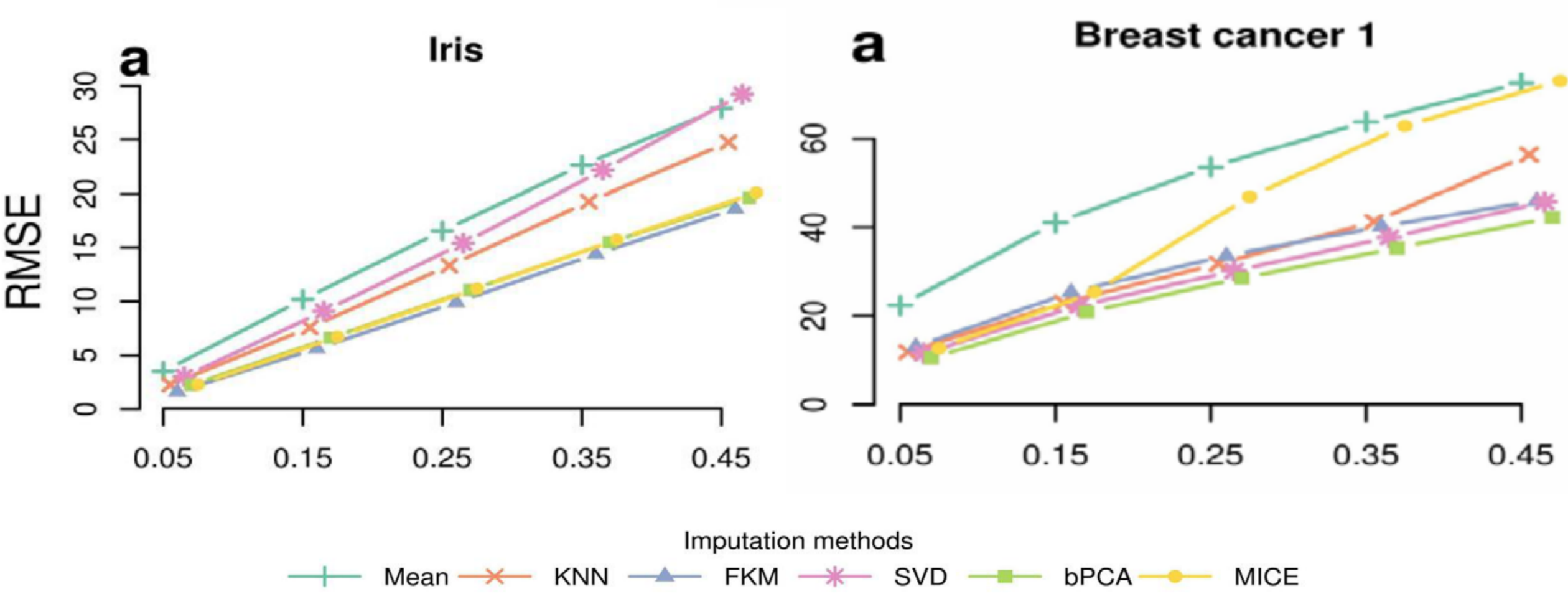
- Нужно настраивать SVD
- Может долго работать

Пропущенные значения: методы восстановления

Совет

Для признака с пропусками добавить признак-индикатор было ли значение или нет

Пропущенные значения: сравнение методов



A Comparison of Six Methods for Missing Data Imputation

Пропущенные значения: сравнение методов

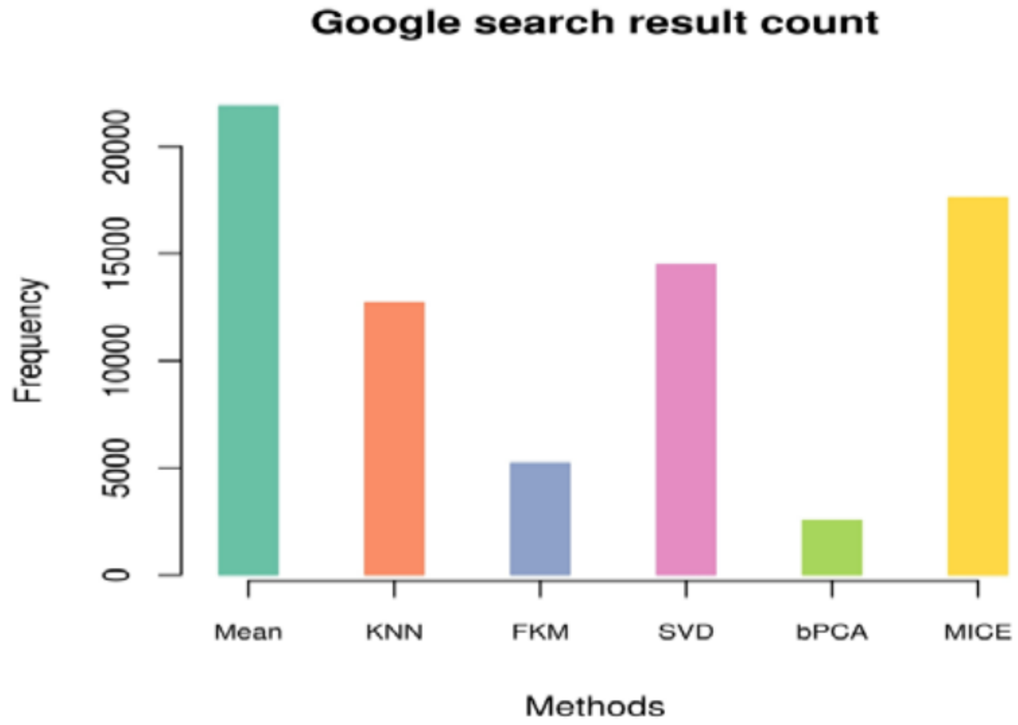


Figure 2: Google search result count. A simple Google search of each method associated to the term 'missing data' provides an idea of the respective popularity of the 6 methods.

A Comparison of Six Methods for Missing Data Imputation

Пропущенные значения: резюме

- Перед тем как восстанавливать пропуски убедитесь, что это повлияет на качество
- Метод восстановления зависит от данных
- Начинать лучше с простых методов
- Сложные методы порой дают большее качество, но их трудно внедрить в продакшн

Пропущенные значения: ссылки

- Хороший код с примерами
- KNN для восстановления пропусков
- Bayessian PCA
- Fuzzy K-Means

Спасибо