Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина)

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ЗНАЧЕНИЙ В НАБОРЕ ДАННЫХ

Выполнил: Вологдин Максим Дмитриевич, гр. 7381
Руководитель: Жукова Наталия Александровна, д.т.н., доцент
Консультант: Жангиров Тимур Рафаилович, ассистент

Санкт-Петербург, 2021

Актуальность

При проведении различных социологических, экономических и статистических исследований в полученных наборах данных могут содержаться пропущенные значения. Это может произойти по следующим причинам:

- Ошибки при записи
- Ошибки при измерении
- Невозможность сбора данных

Не все алгоритмы способны работать с неполными данными, поэтому необходимо уметь их восстанавливать наиболее точным образом

Цель и задачи

Цель: Проведение сравнительного анализа методов восстановления пропущенных значений в наборе данных

Задачи:

1. Определение списка методов для исследования
2. Проведение экспериментов по сравнению методов на различных наборах данных
3. Классификация методов и анализ полученных результатов

Методы для исследования

Для сравнительного анализа были выбраны методы, основанные на различных математических аппаратах:

1. Заполнение средним и медианой
2. Алгоритм K–средних
3. Алгоритм K–ближайших соседей
4. EM-алгоритм
5. Множественное заполнение

- Линейная регрессия
- Деревья решений
- Случайный лес
- Дополнительные деревья

6. Искусственные нейронные сети

Эксперименты по сравнению методов.

Используемые наборы данных:

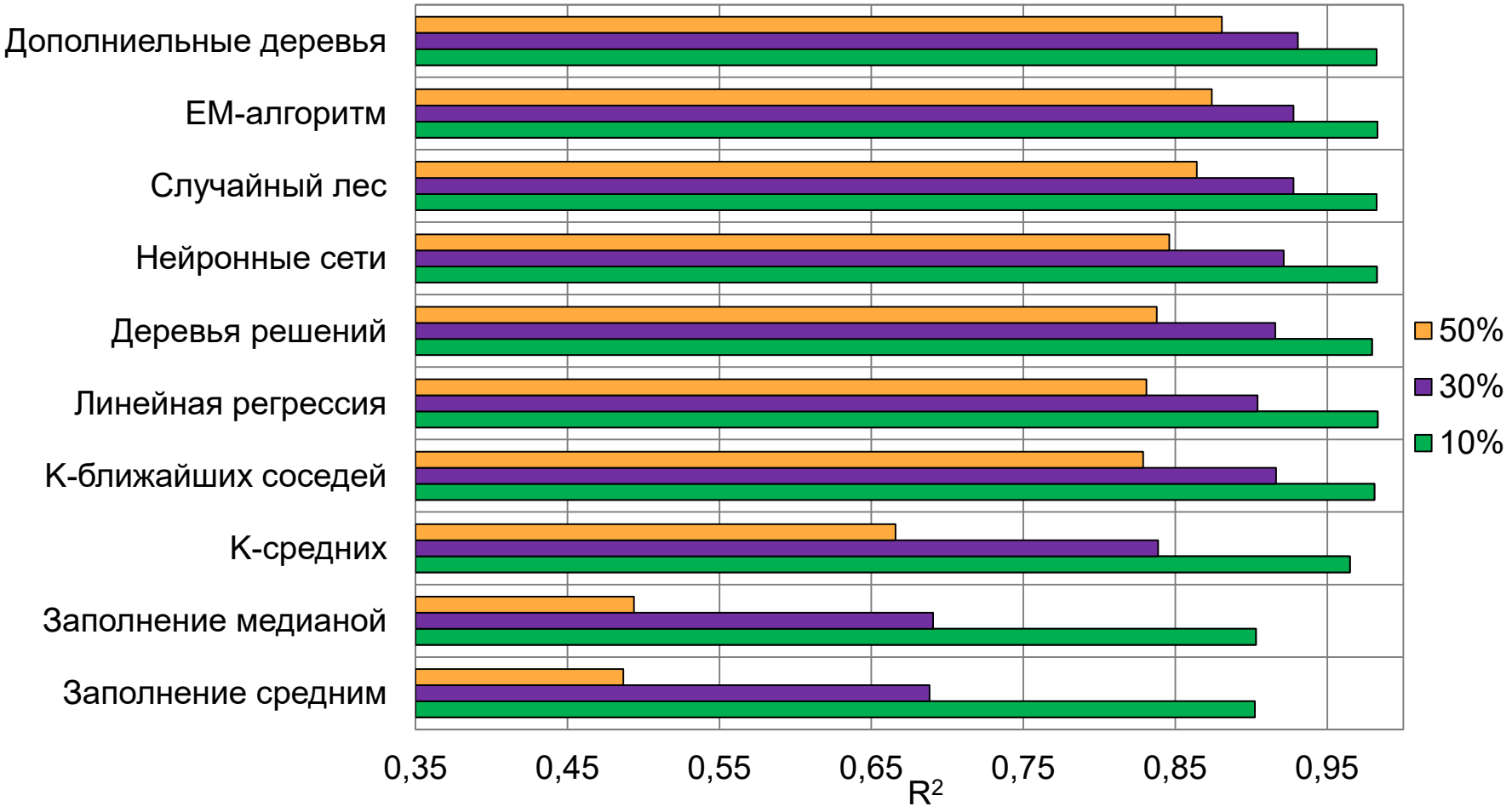
1. **Abalone Dataset** (4177 объектов и 8 признаков) – содержит данные о физических измерениях морских ушек (диаметр, вес и т.п.)
2. **Boston House Prices** (506 объектов и 13 признаков) – содержит данные о пригородах и городах Бостона (уровень преступности, процент людей низкого статуса и т.п.)
3. **California Housing Prices** (20640 объектов и 8 признаков) – содержит данные о домах в районах Калифорнии, основанные на переписи 1990 года (возраст домов, количество спален и т.п.)
4. **Breast Cancer Wisconsin (Diagnostic) Data Set** (569 объектов и 30 признаков) – содержит данные по раку груди – характеристики ядер клеток, представленных на изображении (радиус, вогнутость и т.п.)

Эксперименты проводились на 10%, 30% и 50% пропусков

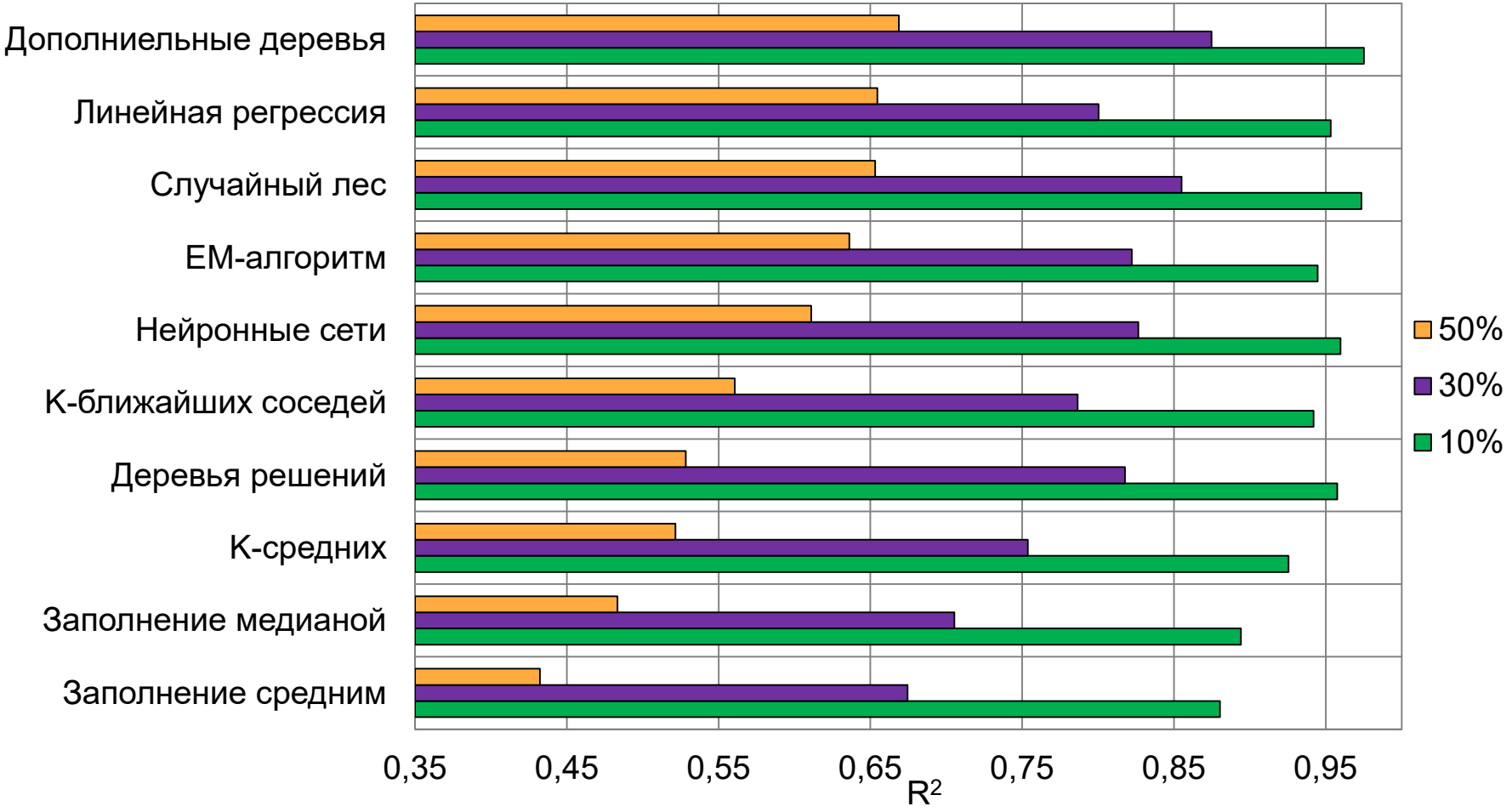
Тип данных – количественный

Метрика качества – коэффициент детерминации (далее R^2 или точность)

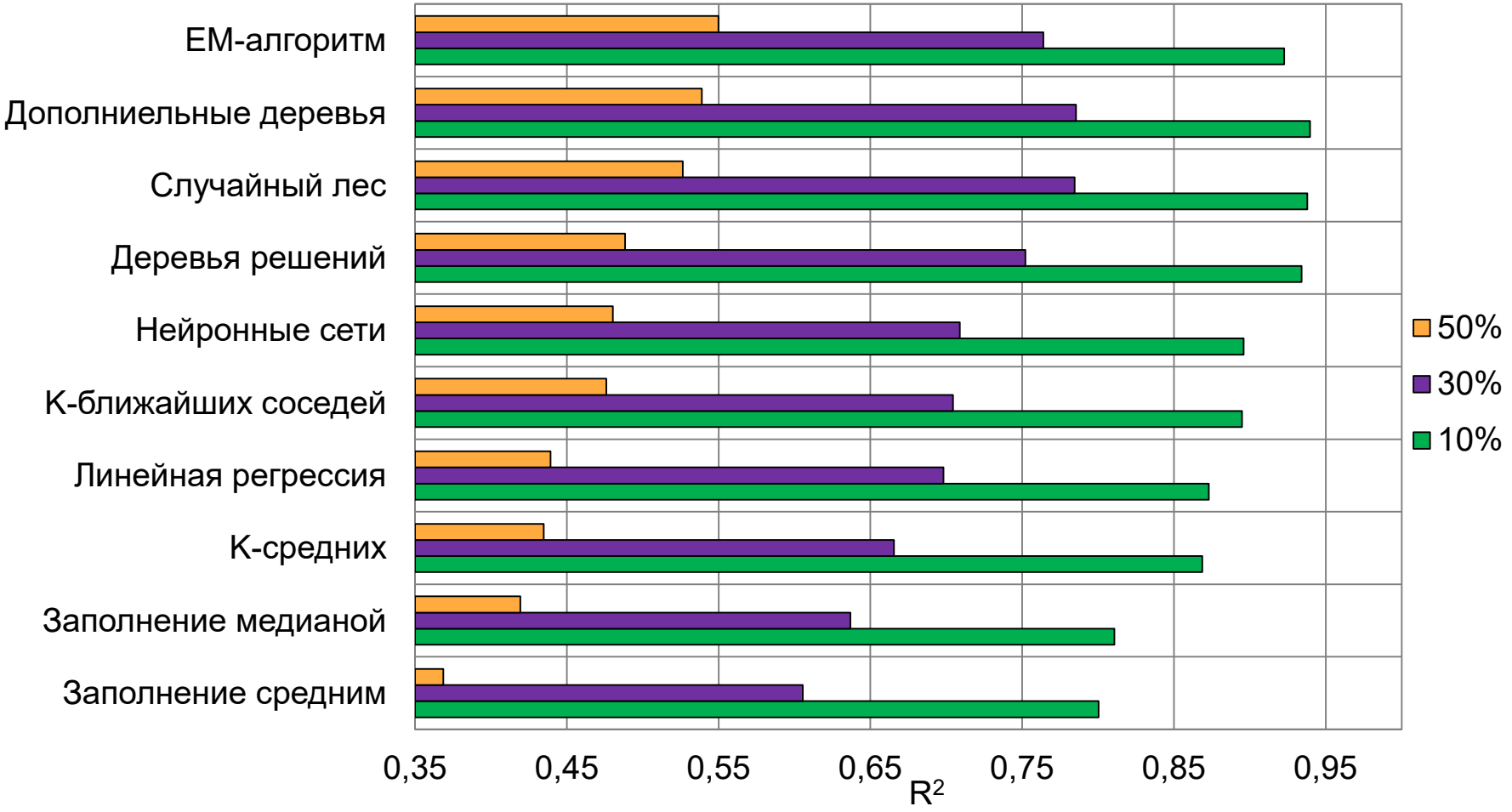
Эксперименты по сравнению методов. Набор Abalone



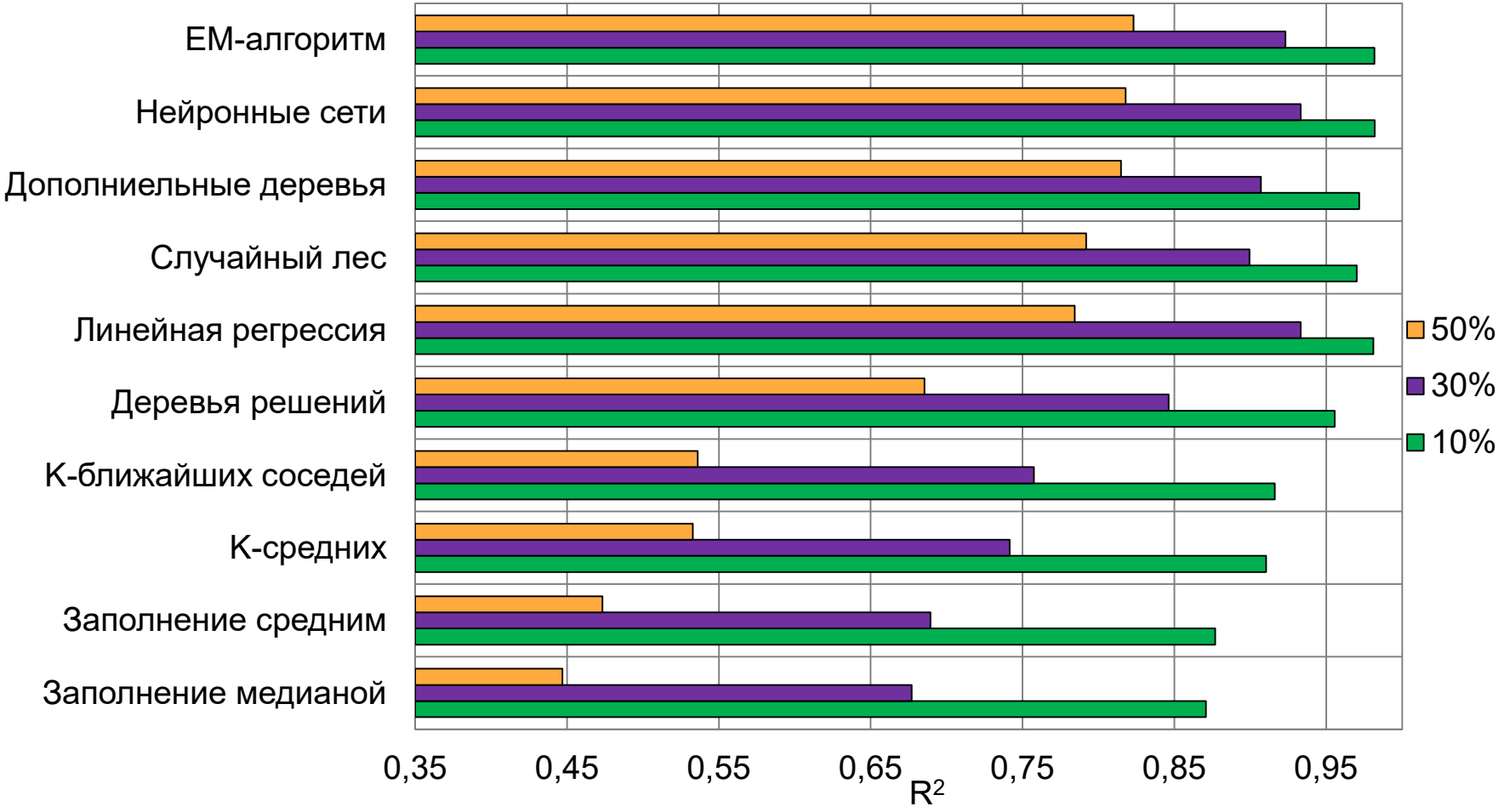
Эксперименты по сравнению методов. Набор Boston



Эксперименты по сравнению методов. Набор California



Эксперименты по сравнению методов. Набор Cancer



Классификация методов

Точность / Скорость	Низкая (нижний квартиль)	Средняя	Высокая (верхний квартиль)
Высокая (В среднем <0.5с)	Заполнение средним и медианой	Множественное заполнение Линейной регрессией	–
Средняя (В среднем 0.5-10с)	Алгоритм K-средних	Алгоритм K–ближайших соседей и Множественное заполнение Деревьями решений	EM-алгоритм
Низкая (В среднем >10с)	–	Искусственные нейронные сети	Множественное заполнение Случайным лесом и Дополнительными деревьями

Заключение

- Был выполнен обзор предметной области и выбраны методы для изучения, основанные на различных математических аппаратах
- Были проведены эксперименты по сравнению методов на 10%, 30% и 50% пропусков и по их результатам построены графики для каждого набора
- Методы были классифицированы по их точности и скорости работы в задаче восстановления пропущенных значений. Наиболее высокие результаты в этих параметрах показали EM–алгоритм и Множественное заполнение с помощью Линейной регрессии.

Дальнейшие направления исследований включают в себя рассмотрение методов, не вошедших в данную работу, а также более глубокое изучение нейронных сетей в контексте поставленной задачи.

Апробация работы

- Репозиторий проекта
  https://github.com/Makkksx/VKR2021

Запасные слайды

Коэффициент детерминации

Коэффициент детерминации, или R^2 показывает, насколько условная дисперсия полученной модели отличается от дисперсии реальных значений, то есть насколько хорошо модель описывает данные.

$$R^2\left(y, y_i\right) = 1 - \frac{\sum_{i=0}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=0}^{n}\left(y_i - \bar{y}\right)^2}$$

Где $\hat{y}_i$ – предсказанное значение, y_i – соответствующее истинное значение, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Область применения

- Полученные результаты можно использовать в различных количественных исследованиях – исследованиях, связанных с подсчетом результатов наблюдений. Количественные данные — это любые данные, представленные в числовом формате.
- Количественные исследования широко используются в психологии, экономике, демографии, социологии, маркетинге, здравоохранении, гендерных исследованиях и политологии; реже в антропологии и истории. Исследования в математических науках, таких как физика, являются количественными по определению.

Случайный лес и Дополнительные деревья

- В случайных лесах каждое дерево строится независимо друг от друга на случайной выборке из начального набора данных. Кроме того, в процессе построения деревьев выбирается оптимальная точка при разбиении каждого узла.
- Алгоритм дополнительных деревьев также построен на множестве решающих деревьев, но в отличии от случайного леса, дополнительные деревья выбирают случайную точку при разбиении узлов, что позволяет алгоритму работать быстрее.

Для построения Деревьев решений используется алгоритм CART (Classification And Regression Tree) – алгоритм бинарного дерева решений

Постановка задачи

Имеется матрица объектов-признаков $X^{n \times d}$, где n – количество объектов, d – количество признаков. Часть значений матрицы пропущена. Необходимо получить матрицу объектов-признаков без пропущенных значений, наиболее близкую к оригинальной.

Реализация методов

Все методы были реализованы на языке программирования Python 3.9 с помощью библиотек:

- Scikit-learn 0.24.1 для основной части методов (CPU)
- Datawig 0.2.0 для нейронных сетей mxnet (GPU)

Характеристики ПК, на котором проводились эксперименты: графическая карта NVIDIA GeForce GTX 1050Ti (4 Гб памяти), 12Гб оперативной памяти и процессор Intel® Core™ i5-7300HQ

Время работы методов

	Abalone, мс	Boston, мс	California, мс	Cancer, мс	Среднее, мс
Заполнение средним	0,01	0,01	0,02	0,01	0,01
Заполнение медианой	0,02	0,01	0,05	0,02	0,03
Линейная регрессия	0,17	0,11	0,50	0,38	0,29
K-средних	0,35	0,15	0,95	0,16	0,40
EM-алгоритм	1,14	0,45	4,11	0,56	1,56
Деревья решений	1,16	0,32	8,67	1,89	3,01
K-ближайших соседей	1,15	0,03	33,92	0,07	8,79
Дополнительные деревья	9,26	12,31	63,77	45,30	32,66
Случайный лес	23,83	18,73	256,56	75,69	93,70
Нейронные сети	316,60	61,45	1693,81	168,11	559,99

Типы пропущенных значений

- Пропуски называют полностью случайными (MCAR), если условная вероятность P (X_j пропущено/прочие X) не зависит ни от X_j, ни от X.
- Пропуски называют случайными (MAR), если условная вероятность P (X_j пропущено/прочие X) не зависит от X_j, но может зависеть от других X. В этом и предыдущем случае механизм пропусков несущественен и к данным применимы методы восстановления.
- Пропуски называют неслучайными (MNAR), если условная вероятность P (X_j пропущено/прочие X) зависит от X_j. В этом случае механизм пропусков существенен и для корректного анализа данных необходимо знать этот механизм.

Сложность методов

n – количество объектов, d – количество признаков, i – количество итераций

1. Заполнение средним и медианой – O(n×d)
2. Алгоритм K–средних – O(i×k×d×n), где k – количество кластеров
3. Алгоритм K–ближайших соседей – O(k×d×n), где k – количество соседей
4. EM-алгоритм – O(i×n×d)
5. Итеративное вменение

- Линейная регрессия – O(i×d×n)
- Деревья решений – O(i×d×n×log(n)),
- Случайный лес и Дополнительные деревья – O(m×i×d×n×log(n)),

где m – количество деревьев

Квартили

- Квартили — значения, которые делят таблицу данных (или ее часть) на четыре группы, содержащие приблизительно равное количество наблюдений. Общий объем делится на четыре равные части: 25%, 50%, 75% 100%.
- 25% значений меньше, чем нижняя квартиль, 75% значений меньше, чем верхняя квартиль.