Занятие 3
Линейная регрессия.

Елена Кантонистова

elena.kantonistova@yandex.ru

ВШЭ, 2021

ПЛАН ЗАНЯТИЯ

- Теоретическая часть:

- оценка обобщающей способности модели

- линейная регрессия в деталях

- метрики качества в задаче регрессии

- Решение задачи регрессии в Python

ОЦЕНКА ОБОБЩАЮЩЕЙ СПОСОБНОСТИ МОДЕЛИ

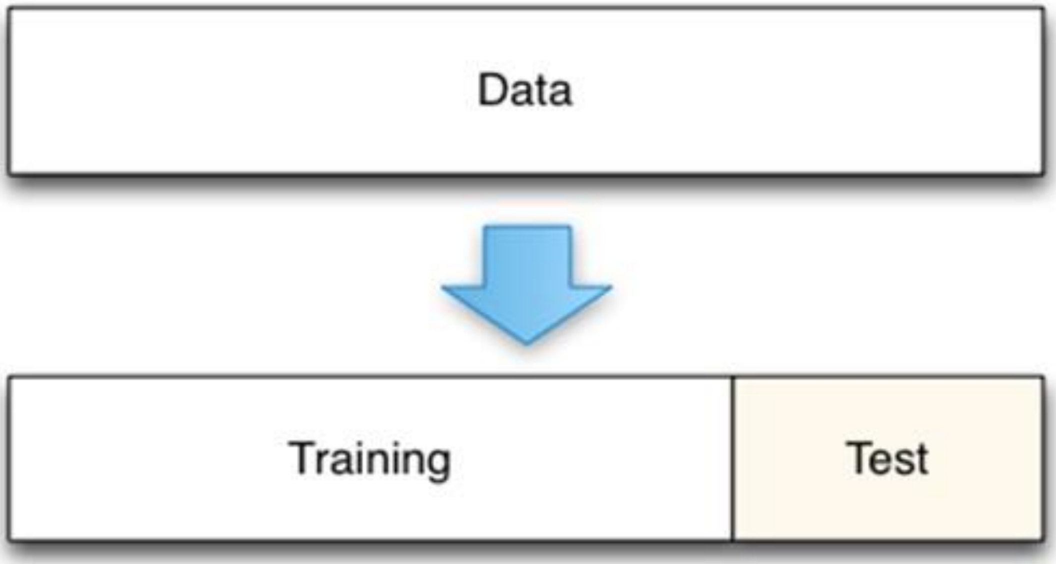
ОЦЕНКА ПРЕДСКАЗАТЕЛЬНОЙ СПОСОБНОСТИ АЛГОРИТМА

- Пусть мы решаем задачу *предсказания стоимости дома* по его признакам.

- В обучающей выборке 1000 домов.
- Мы обучаем алгоритм по имеющимся 1000 домам. *На каких объектах будем проверять качество алгоритма?*

ОЦЕНКА ПРЕДСКАЗАТЕЛЬНОЙ СПОСОБНОСТИ АЛГОРИТМА

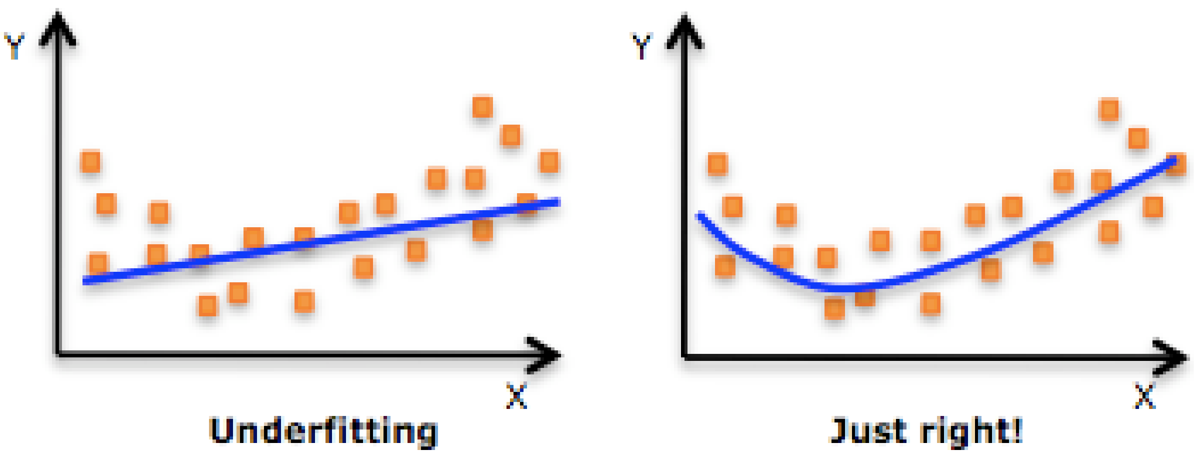
- Перед началом обучения отложим часть обучающих объектов и не будем использовать их для построения модели (отложенная выборка).



ОТЛОЖЕННАЯ ВЫБОРКА

- Перед началом обучения отложим часть обучающих объектов и не будем использовать их для построения модели (отложенная выборка).
- Тогда можно измерить качество построенной модели на отложенной выборке и оценить ее предсказательную силу.

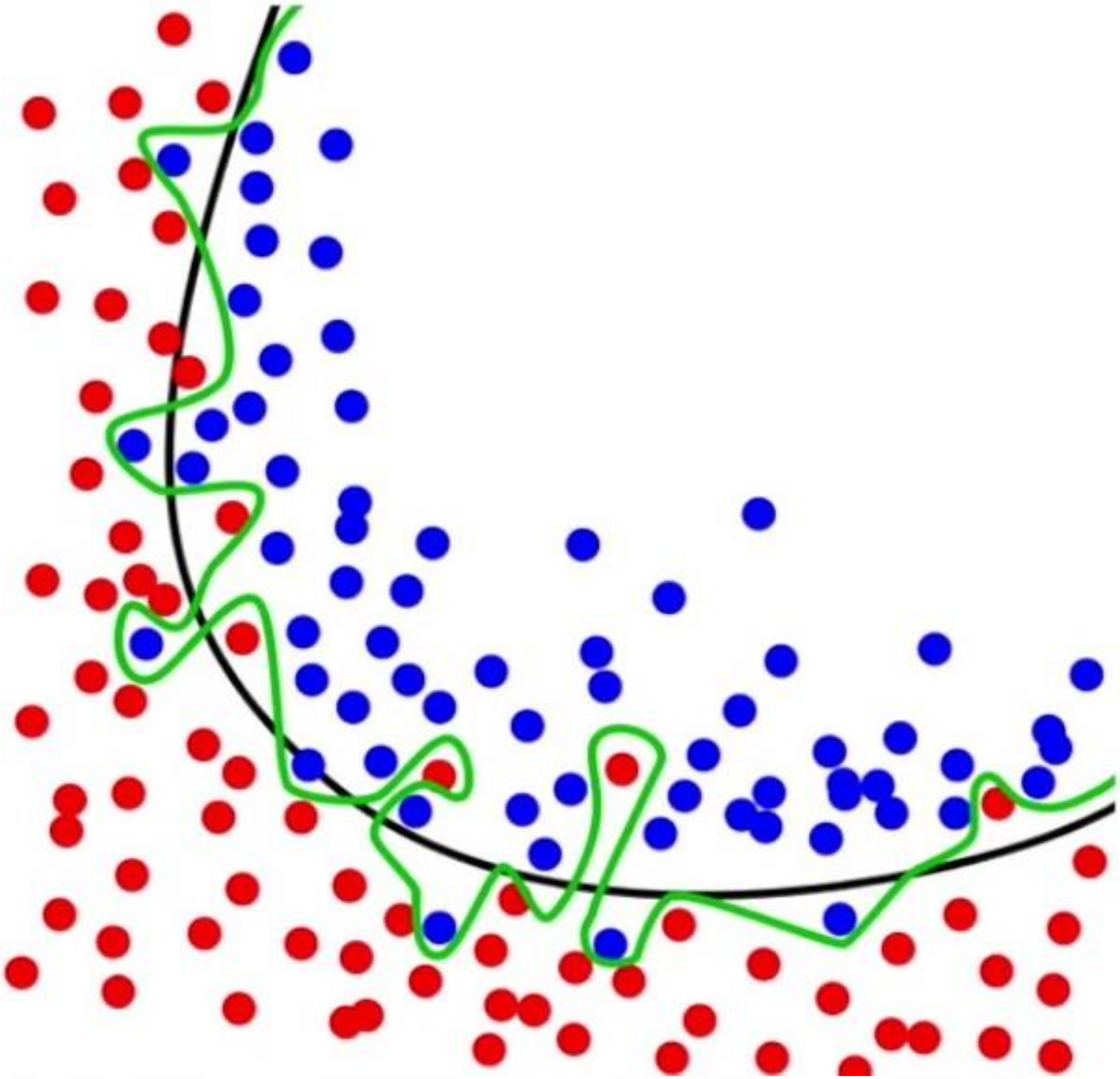
ПЕРЕОБУЧЕНИЕ И НЕДООБУЧЕНИЕ



ИЗ-ЗА ЧЕГО ВОЗНИКАЕТ ПЕРЕОБУЧЕНИЕ

- Избыточная сложность модели (большое количество весов). В этом случае лишние степени свободы в модели “тратятся” на чрезмерно точную подгонку под обучающую выборку.

ПРИМЕР ПЕРЕОБУЧЕНИЯ В ЗАДАЧЕ КЛАССИФИКАЦИИ

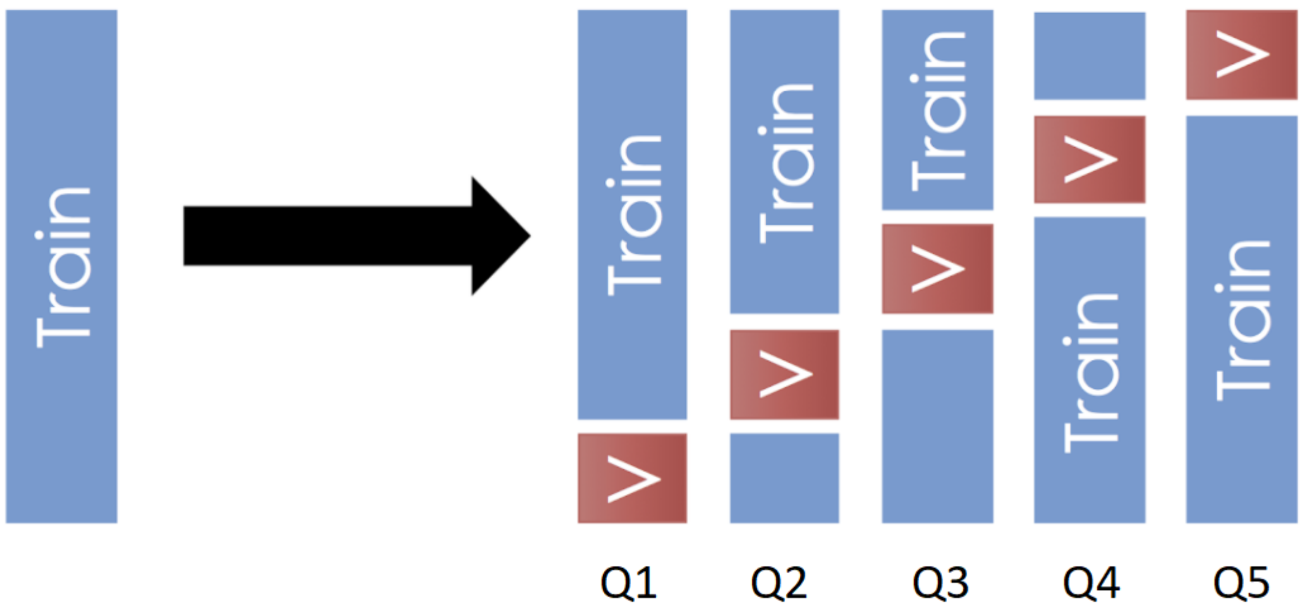


ПРИЗНАК ПЕРЕОБУЧЕНИЯ

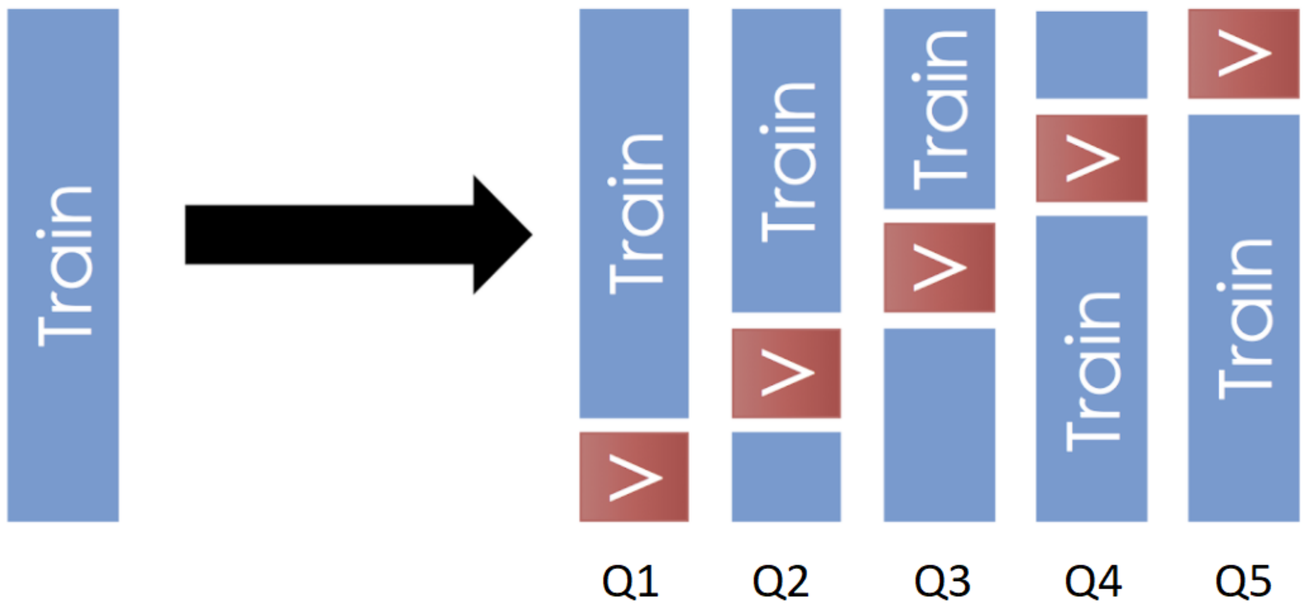
- *Если качество на отложенной выборке сильно ниже качества на обучающих данных, то происходит переобучение*

КРОСС-ВАЛИДАЦИЯ

- Разбиваем данные на k частей (фолдов)
- Обучаем алгоритм на всех частях, кроме одной (тестовой) и оцениваем качество. Так делаем k раз, меняя тестовую часть.
- Усредняем полученные результаты



КРОСС-ВАЛИДАЦИЯ



$$CV = \frac{1}{k}\sum_{i=1}^{k} Q_i$$

ЛИНЕЙНАЯ РЕГРЕССИЯ В ДЕТАЛЯХ

ЛИНЕЙНАЯ РЕГРЕССИЯ

Пример (напоминание):

Предположим, что мы хотим предсказать *стоимость дома* y по его *площади* (x_1) и *количеству комнат* (x_2).

Линейная модель для предсказания стоимости:

$$\boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{w_0} + \boldsymbol{w_1}\boldsymbol{x_1} + \boldsymbol{w_2}\boldsymbol{x_2},$$

где w_0, w_1, w_2 -

параметры модели (*веса*).

Линейная регрессия

Пример (напоминание):

Предположим, что мы хотим предсказать *стоимость дома* y

по его *площади* (x_1) и *количеству комнат* (x_2).

Линейная модель для предсказания стоимости:

$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2},$$

где w_0, w_1, w_2 -

параметры модели (*веса*).

Общий вид (линейная регрессия):

$$\boxed{\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n},}$$

где $x_1, \dots, x_n$ - признаки объекта x.

О ПРИМЕНИМОСТИ ЛИНЕЙНОЙ МОДЕЛИ

Пример:

Предположим, что мы хотим предсказать *стоимость дома* y по его *площади* (x_1) и *количеству комнат* (x_2), *району* (x_3) *и удаленности от МКАД* (x_4).

$$\boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{w_0} + \boldsymbol{w_1 x_1} + \boldsymbol{w_2 x_2} + \boldsymbol{w_3 x_3} + \boldsymbol{w_4 x_4}.$$

О ПРИМЕНИМОСТИ ЛИНЕЙНОЙ МОДЕЛИ

Пример:

Предположим, что мы хотим предсказать *стоимость дома* y по его *площади* (x_1) и *количеству комнат* (x_2), *району* (x_3) *и удаленности от МКАД* (x_4).

$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4}.$$

Проблема №1: район (x_3) – это не число, а название района. Например, Мамыри, Дудкино, Барвиха… Что с этим делать?

О ПРИМЕНИМОСТИ ЛИНЕЙНОЙ МОДЕЛИ

Пример:

Предположим, что мы хотим предсказать *стоимость дома* y по его *площади* (x_1) и *количеству комнат* (x_2), *району* (x_3) *и удаленности от МКАД* (x_4).

$$a(x) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4.$$

Проблема №1: район (x_3) – это не число, а название района. Например, Мамыри, Дудкино, Барвиха… Что с этим делать?

Решение – one-hot encoding (OHE): создаем новые числовые столбцы, каждый из которых является индикатором района.

ONE-HOT ENCODING

Район
Дудкино
Барвиха
Мамыри
…
Барвиха

Мамыри	Дудкино	Барвиха
0	1	0
0	0	1
1	0	0
…	…	…
0	0	1

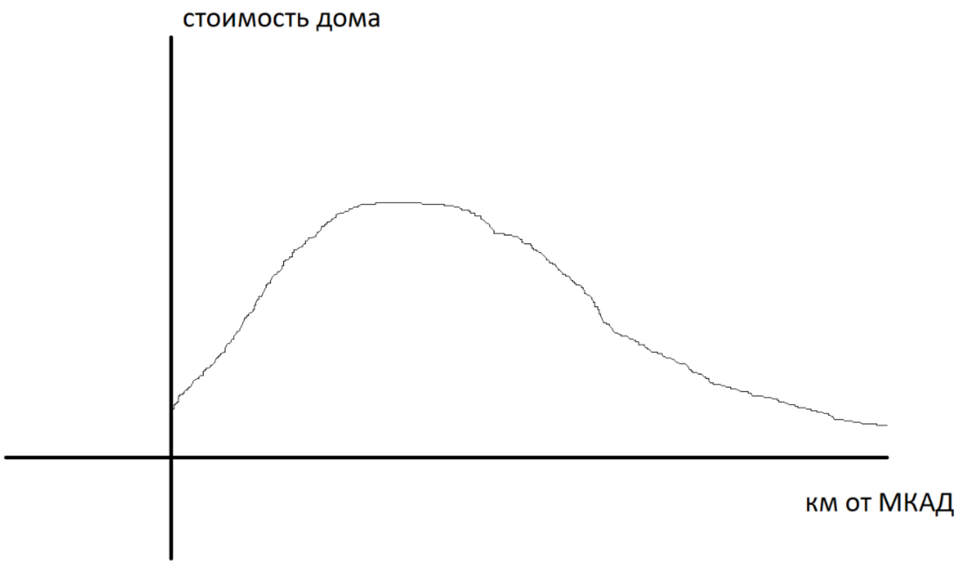
$$\boldsymbol{a(x)} =$$
$$= \boldsymbol{w_0} + \boldsymbol{w_1 x_1} + \boldsymbol{w_2 x_2} + \boldsymbol{w_{31} x}_{\text{Мамыри}} + \boldsymbol{w_{32} x}_{\text{Дудкино}} + \boldsymbol{w_{33} x}_{\text{Барвиха}} + \boldsymbol{w_4\, x_4}.$$

О применимости линейной модели

Пример:

Предположим, что мы хотим предсказать *стоимость дома* y по его *площади* (x_1) и *количеству комнат* (x_2), *району* (x_3) *и удаленности от МКАД* (x_4).

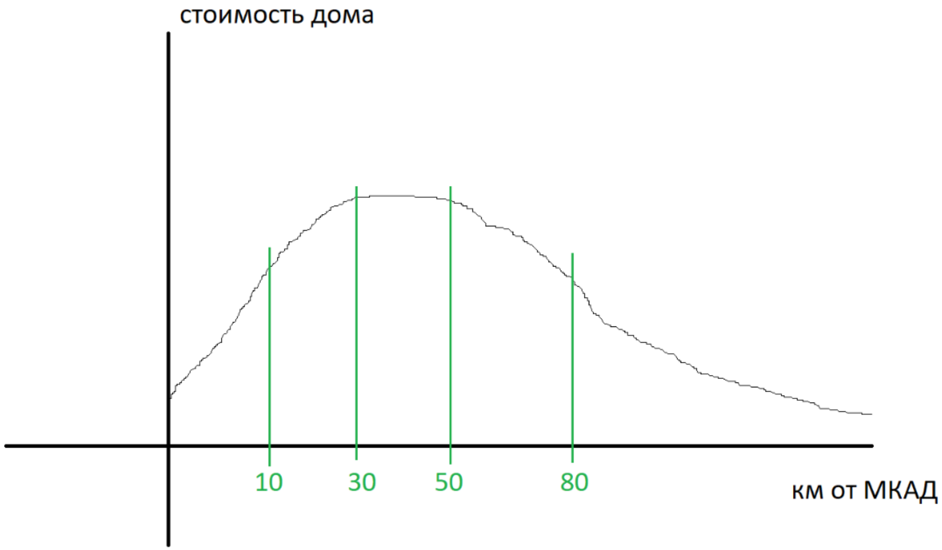
Проблема №2: удаленность от МКАД (x_4) не монотонно влияет на стоимость дома.



О применимости линейной модели

<u>Проблема №2:</u> удаленность от МКАД (x_4) не монотонно влияет на стоимость дома.

<u>Решение</u> – бинаризация (разбиение на бины).



Новые признаки:

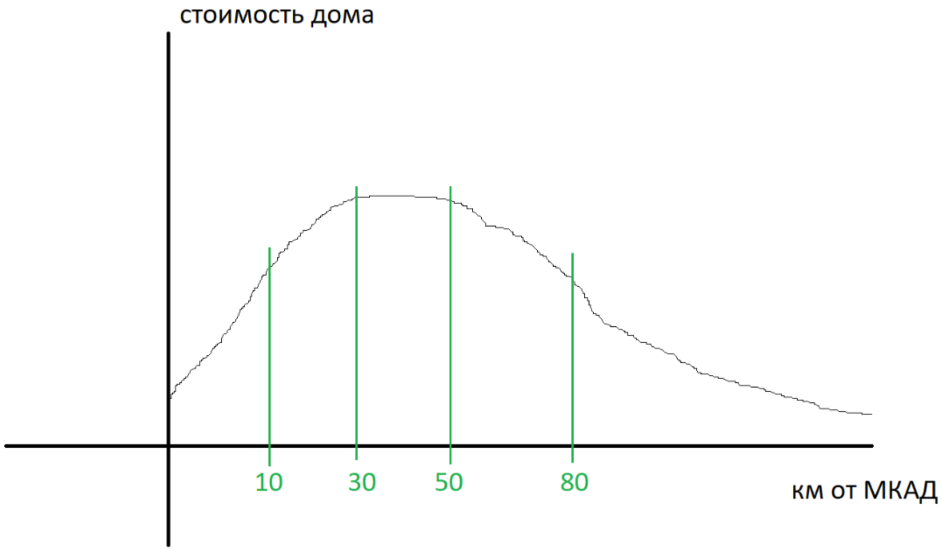
- $x_{[0;10)}$ - равен 1, если дом находится в пределах 10 км от МКАД, и 0 иначе
- $x_{[10;30)}$ - равен 1, если дом находится в пределах от 10 км до 30 км МКАД, и 0 иначе. И т.д.

О применимости линейной модели

<u>Проблема №2:</u> удаленность от МКАД (x_4) не монотонно влияет на стоимость дома.

<u>Решение</u> – бинаризация (разбиение на бины).

Новые признаки:



- $x_{[0;10)}$ - равен 1, если дом находится в пределах 10 км от МКАД, и 0 иначе
- $x_{[10;30)}$ - равен 1, если дом находится в пределах от 10 км до 30 км МКАД, и 0 иначе. И т.д.

$$\boldsymbol{a(x) = \\ = w_0 + w_1x_1 + w_2x_2 + \cdots + w_{41}x_{[0;10)} + w_{42}x_{[10;30)} + w_{43}x_{[30;50)} \\ + w_{44}x_{\geq 50}}$$

МЕТРИКИ КАЧЕСТВА В ЗАДАЧЕ РЕГРЕССИИ

СРЕДНЕКВАДРАТИЧНОЕ ОТКЛОНЕНИЕ: MSE (MEAN SQUARED ERROR)

Среднеквадратичное отклонение:

$$MSE(a, X) = \frac{1}{l}\sum_{i=1}^{l}\boldsymbol{(a(x_i) - y_i)^2}$$

Плюсы:

- Позволяет сравнивать модели
- Подходит для контроля качества во время обучения

СРЕДНЕКВАДРАТИЧНОЕ ОТКЛОНЕНИЕ: MSE

Среднеквадратичное отклонение:

$$MSE(a, X) = \frac{1}{l}\sum_{i=1}^{l}(\boldsymbol{a}(\boldsymbol{x_i}) - \boldsymbol{y_i})^2$$

Плюсы:

- Позволяет сравнивать модели
- Подходит для контроля качества во время обучения

Минусы:

- Плохо интерпретируется, т.к. не сохраняет единицы измерения (если целевая переменная – кг, то MSE измеряется в кг в квадрате)
- Тяжело понять, насколько хорошо данная модель решает задачу, так как MSE не ограничена сверху.

RMSE (ROOT MEAN SQUARED ERROR)

Корень из среднеквадратичной ошибки:

$$RMSE(a, X) = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2}$$

Плюсы:

- Все плюсы MSE
- **Сохраняет единицы измерения (в отличие от MSE)**

Минусы:

- Тяжело понять, насколько хорошо данная модель решает задачу, так как RMSE не ограничена сверху.

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ (R^2)

Коэффициент детерминации:

$$R^2(a, X) = 1 - \frac{\sum_{i=1}^{l}(a(x_i) - y_i)^2}{\boldsymbol{\sum_{i=1}^{l}(y_i - \overline{y})^2}},$$

где $\overline{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$.

Коэффициент детерминации (R^2)

Коэффициент детерминации:

$$R^2(a, X) = 1 - \frac{\sum_{i=1}^{l}(a(x_i) - y_i)^2}{\sum_{i=1}^{l}(y_i - \overline{y})^2},$$

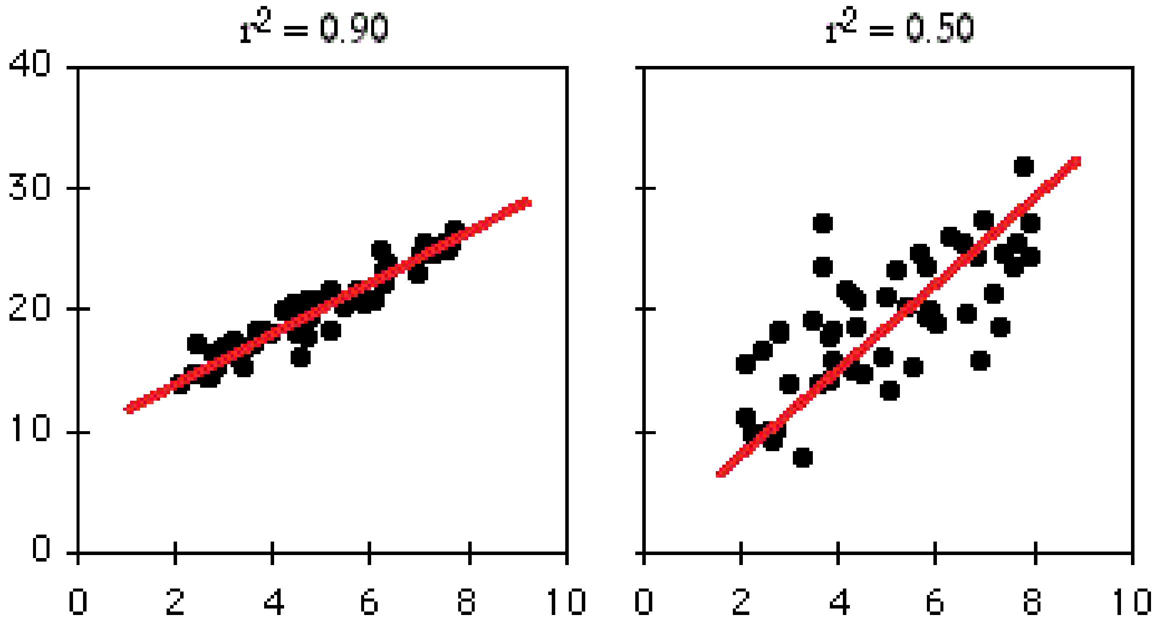
где $\overline{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$.

Коэффициент детерминации ***это доля дисперсии целевой переменной, объясняемая моделью***.

- Чем ближе R^2 к 1, тем лучше модель объясняет данные
- Чем ближе R^2 к 0, тем ближе модель к константному предсказанию
- Отрицательный R^2 говорит о том, что модель плохо решает задачу

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ (R^2)

$$\boldsymbol{R^2 \leq 1}$$



MAE (MEAN ABSOLUTE ERROR)

Средняя абсолютная ошибка:

$$MAE(a, X) = \frac{1}{l} \sum_{i=1}^{l} \boldsymbol{|a(x_i) - y_i|}$$

MAE (MEAN ABSOLUTE ERROR)

Средняя абсолютная ошибка:

$$MAE(a, X) = \frac{1}{l} \sum_{i=1}^{l} |\boldsymbol{a}(\boldsymbol{x}_i) - \boldsymbol{y}_i|$$

Плюсы:

- Менее чувствителен к выбросам, чем MSE

Минусы:

- MAE - не дифференцируемый функционал