Современные методы анализа данных и машинного обучения

Тема 6. Лекция 7

Классическое машинное обучение. Обучение с учителем. Регрессия

Юрий Саночкин

ysanochkin@hse.ru

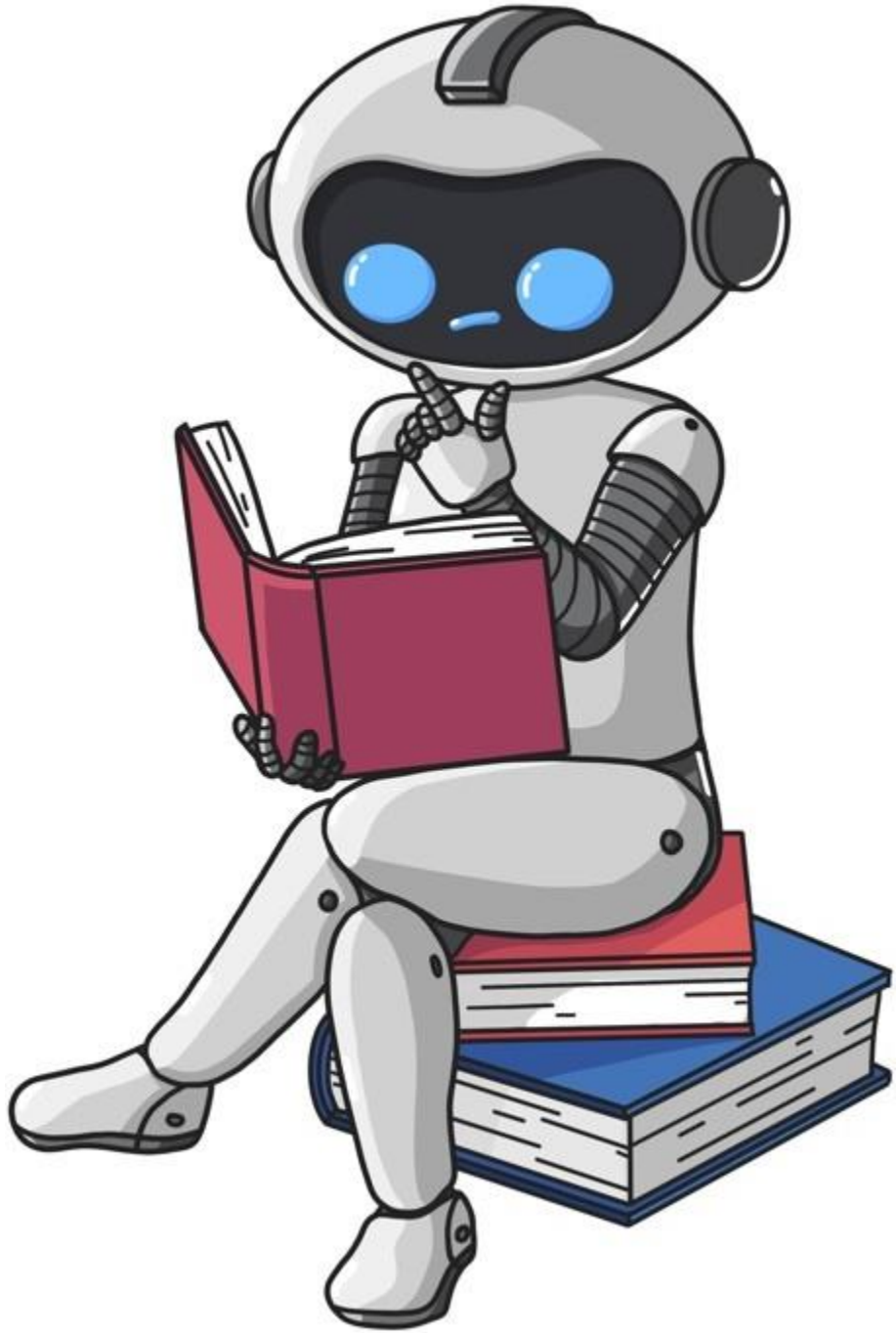
НИУ ВШЭ, 2024

Машинное обучение

- Давайте вспомним, в чём заключается концепция классического машинного обучения!

Машинное обучение

Наука о поиске
закономерностей в данных
с помощью компьютера и
математики



Машинное обучение

- На какие два больших блока можно разделить задачи классического машинного обучения?

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

Ассоциация

«Найди какие вещи я часто ношу вместе»

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Классическое Обучение

Данные заранее категоризированы или численные

С учителем

Предсказать категорию

Классификация

«Разложи носки по цвету»

Предсказать значение

Регрессия

«Разложи галстуки по длине»

Данные никак не размечены

Без учителя

Разделить по схожести

Кластеризация

«Разложи похожие вещи по кучкам»

Найти зависимости

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Выявить последовательности

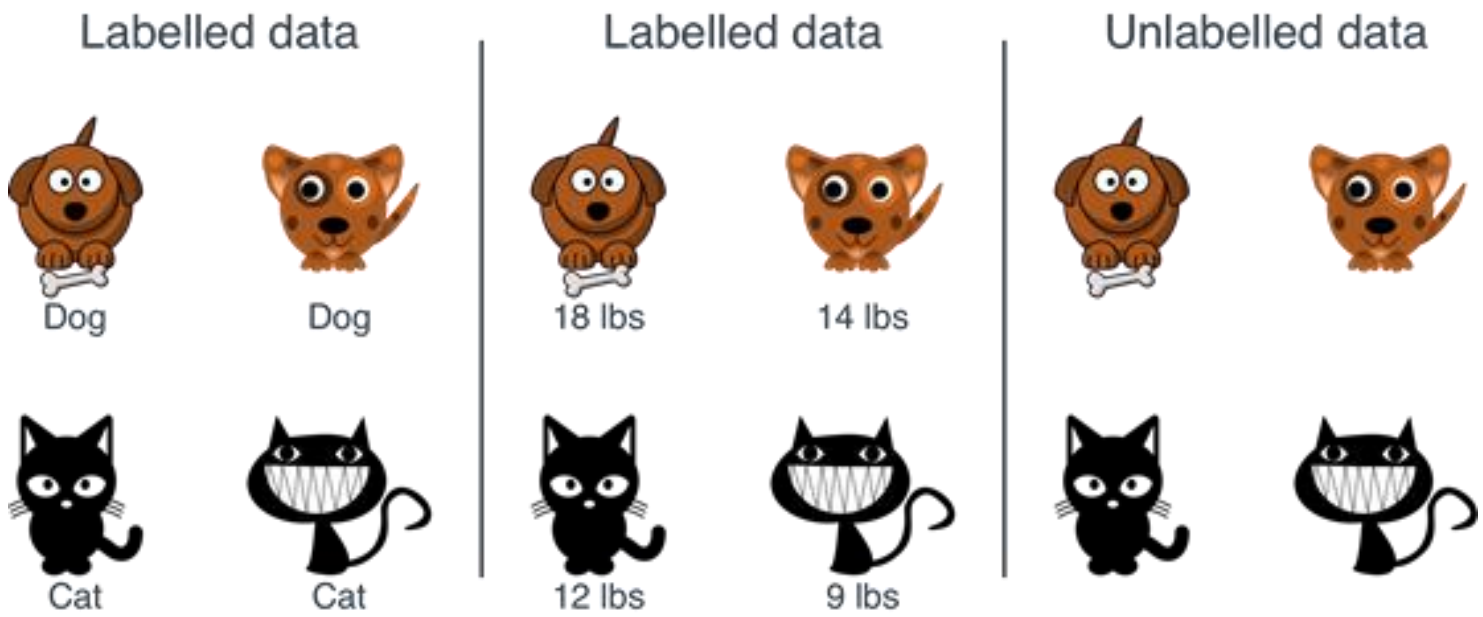
Ассоциация

«Найди какие вещи я часто ношу вместе»

Машинное обучение

- Данные блоки задач машинного обучения неразрывно связаны с понятием размеченных/неразмеченных данных

Размеченные (labelled) vs неразмеченные (unlabelled) данные



Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»

В рамках текущей темы мы поговорим о задачах supervised learning

Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

Ассоциация

«Найди какие вещи я часто ношу вместе»

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

Ассоциация

«Найди какие вещи я часто ношу вместе»

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

А если еще конкретнее – сегодня речь пойдет о регрессии!

Обучение с учителем

- Еще раз повторим основные обозначения!

Обучение с учителем

- Еще раз повторим основные обозначения!
- X — множество всех объектов в пространстве признаков
- Y — область значений целевой переменной

Обучение с учителем

- Еще раз повторим основные обозначения!
- X — множество всех объектов в пространстве признаков
- Y — область значений целевой переменной

- Что представляет собой машинное обучение в этих терминах?
- Фактически это про поиск неизвестной зависимости:
- f: $X \rightarrow Y$ — неизвестная закономерность, функция
- Может даже иметь стохастическую природу!

Обучение с учителем

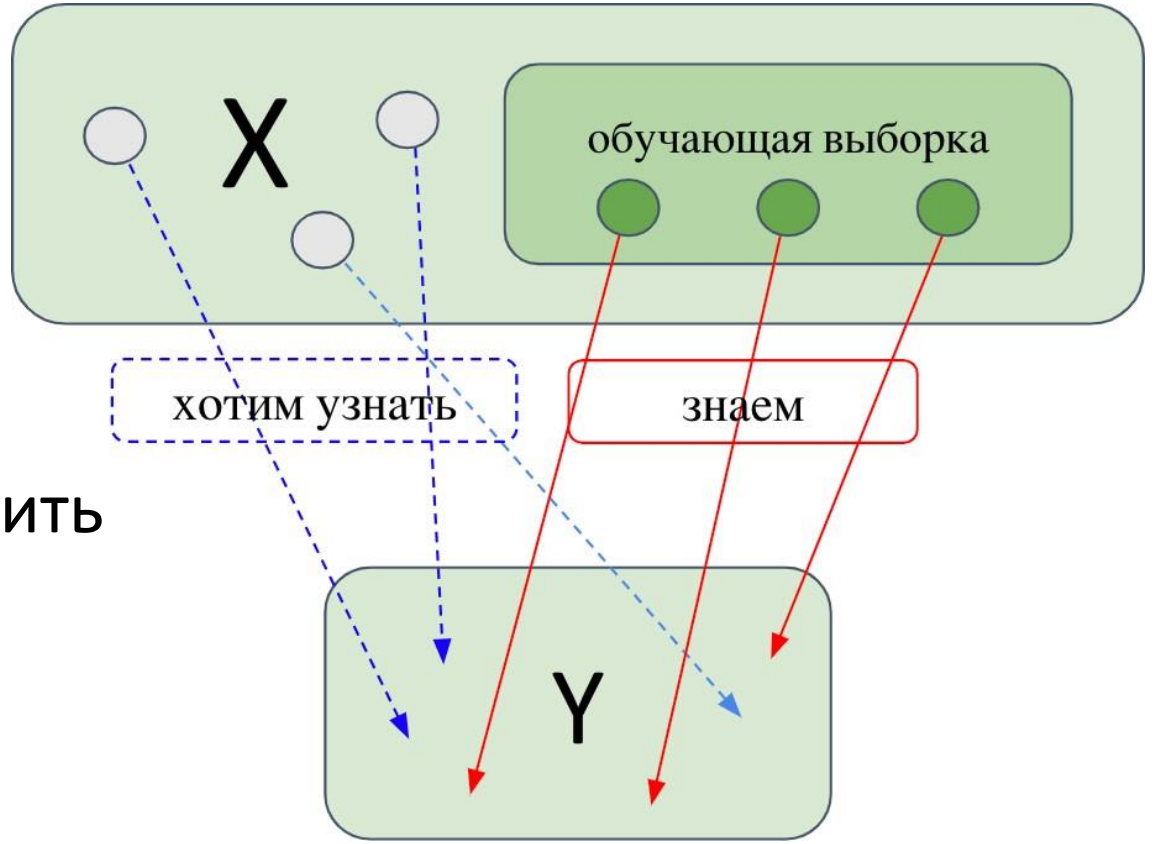
- Как мы это осуществляем?

- Дано: Обучающая выборка вида $\{(X_i, y_i)\}_{i=1}^{n}$
- Цель: Максимально точно определить и приблизить f.

Обучение с учителем

- Как мы это осуществляем?

- Дано: Обучающая выборка вида $\{(X_i, y_i)\}_{i=1}^{n}$
- Цель: Максимально точно определить и приблизить f.



Задача регрессии

- Что такое задача регрессии?
- Говоря по простому — задача, где мы хотим предсказать некоторое численное (вещественное) значение

Задача регрессии

- Что такое задача регрессии?
- Говоря по простому — задача, где мы хотим предсказать некоторое численное (вещественное) значение
- Примеры задач регрессии:
  - Предсказание стоимости жилья для риэлторской компании
  - Предсказание времени доставки
  - Предсказание спроса на такси в конкретном районе в конкретный час завтрашнего дня
  - И так далее

Метрики качества и функционал ошибки

Метрики качества и функционал ошибки

- Допустим, мы кое-что более-менее поняли про постановку задачи машинного обучения (и даже смогли обучить какую-то простенькую модель – что, кстати, правда: на семинаре попробовали же!)
- Но как понять, хорошая у нас получилась модель или нет?

Метрики качества и функционал ошибки

- Допустим, мы кое-что более-менее поняли про постановку задачи машинного обучения (и даже смогли обучить какую-то простенькую модель – что, кстати, правда: на семинаре попробовали же!)
- Но как понять, хорошая у нас получилась модель или нет?
- Для этого существует такое понятие как метрика качества и функционал (функция) ошибки

Метрики качества и функционал ошибки

- Допустим, мы кое-что более-менее поняли про постановку задачи машинного обучения (и даже смогли обучить какую-то простенькую модель – что, кстати, правда: на семинаре попробовали же!)
- Но как понять, хорошая у нас получилась модель или нет?
- Для этого существует такое понятие как метрика качества и функционал (функция) ошибки
- Кстати говоря, а в чем разница между ними? :)

Метрики качества и функционал ошибки

- Метрики качества используются для непосредственной оценки качества обученного алгоритма, с учетом наших бизнес-потребностей
- Проще говоря: смотрим на полученный моделью результат и сравниваем его с правильными ответами

Метрики качества и функционал ошибки

- Метрики качества используются для непосредственной оценки качества обученного алгоритма, с учетом наших бизнес-потребностей
- Проще говоря: смотрим на полученный моделью результат и сравниваем его с правильными ответами
- Функционал ошибки же оценивает некую математическую функцию, которую в процессе обучения пытается минимизировать модель

Метрики качества и функционал ошибки

- Метрики качества используются для непосредственной оценки качества обученного алгоритма, с учетом наших бизнес-потребностей
- Проще говоря: смотрим на полученный моделью результат и сравниваем его с правильными ответами
- Функционал ошибки же оценивает некую математическую функцию, которую в процессе обучения пытается минимизировать модель

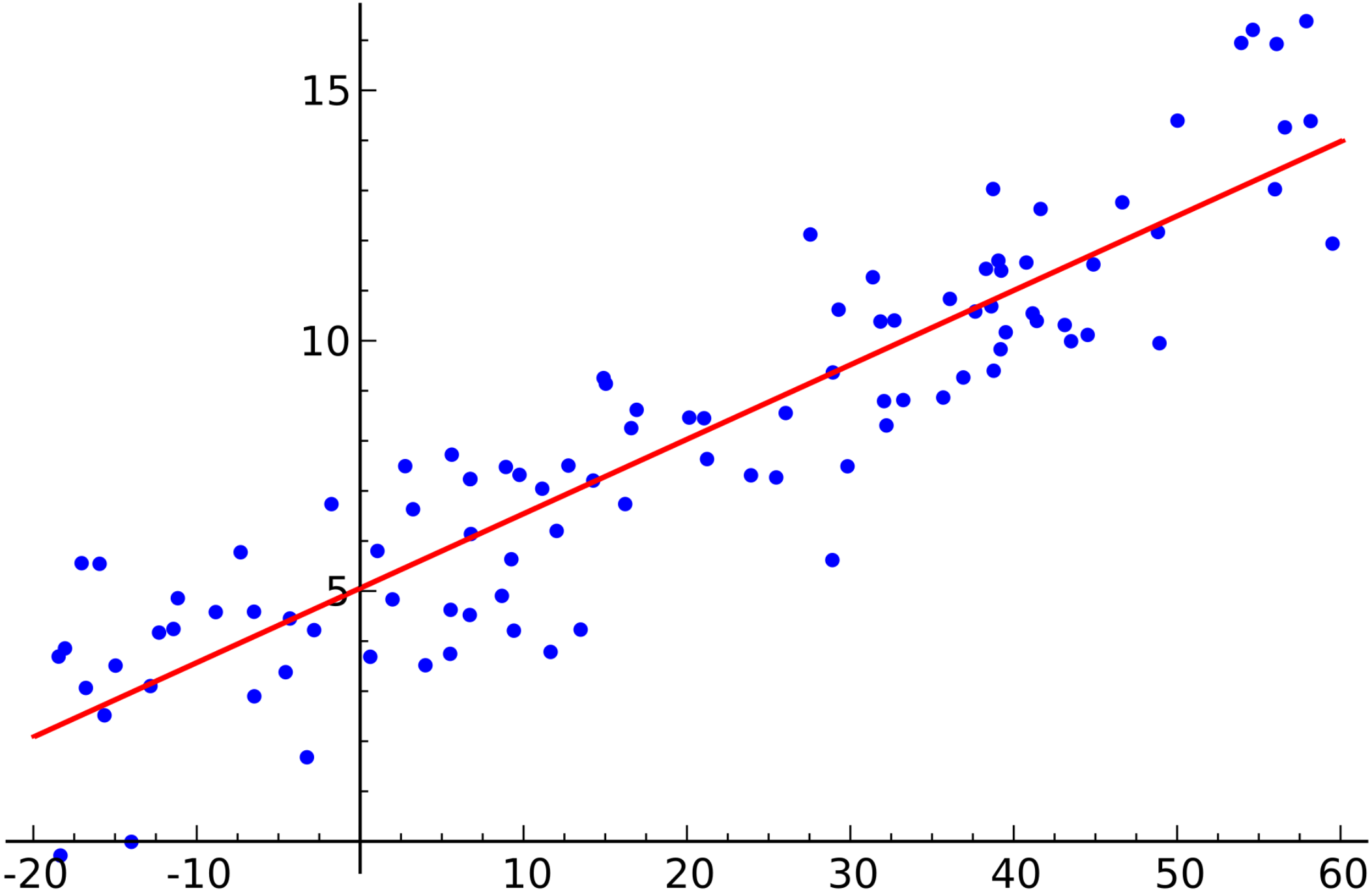
Казалось бы – при чем тут градиентный спуск и все наши разговоры про математический анализ?...

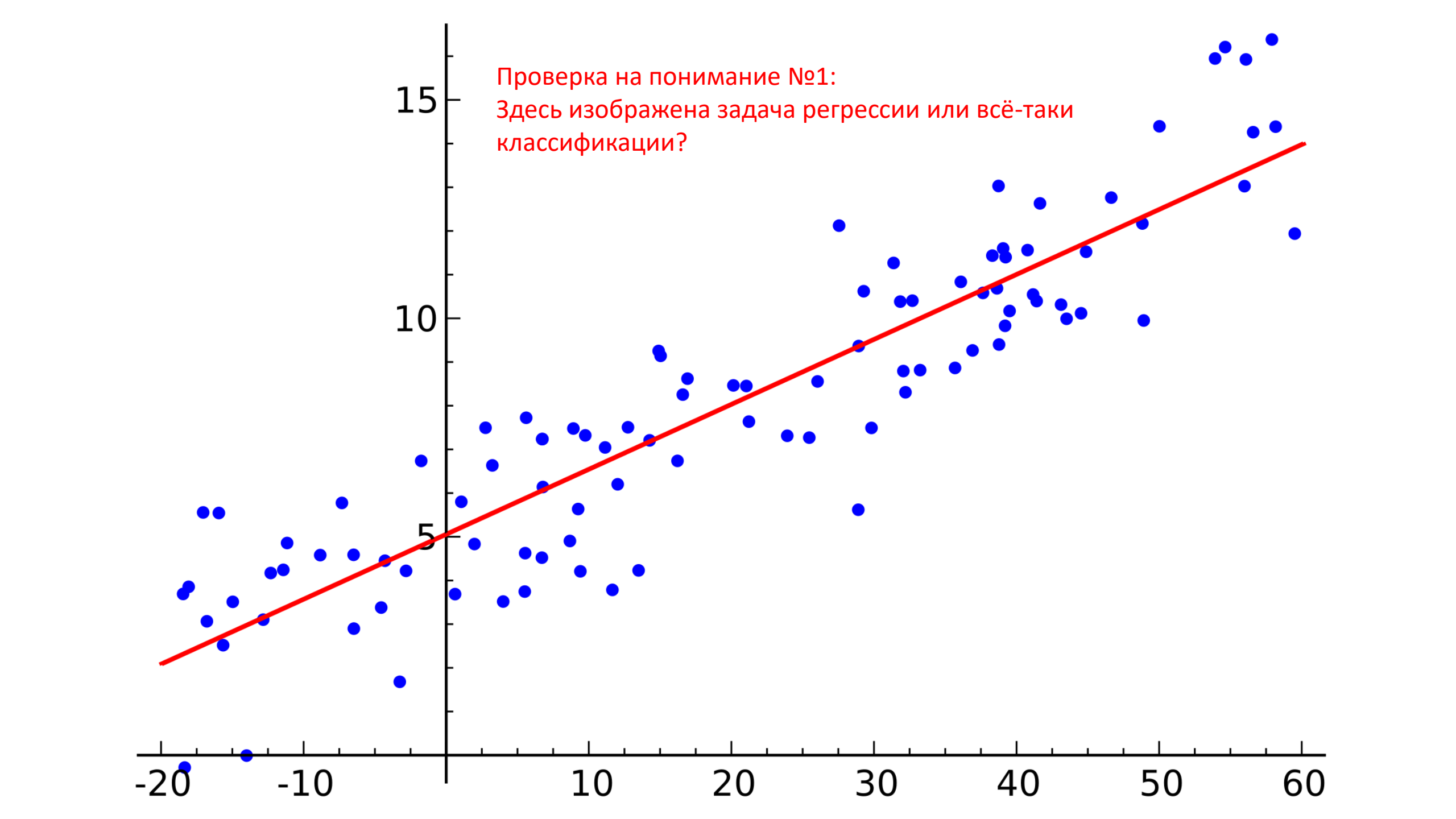
Метрики качества и функционал ошибки

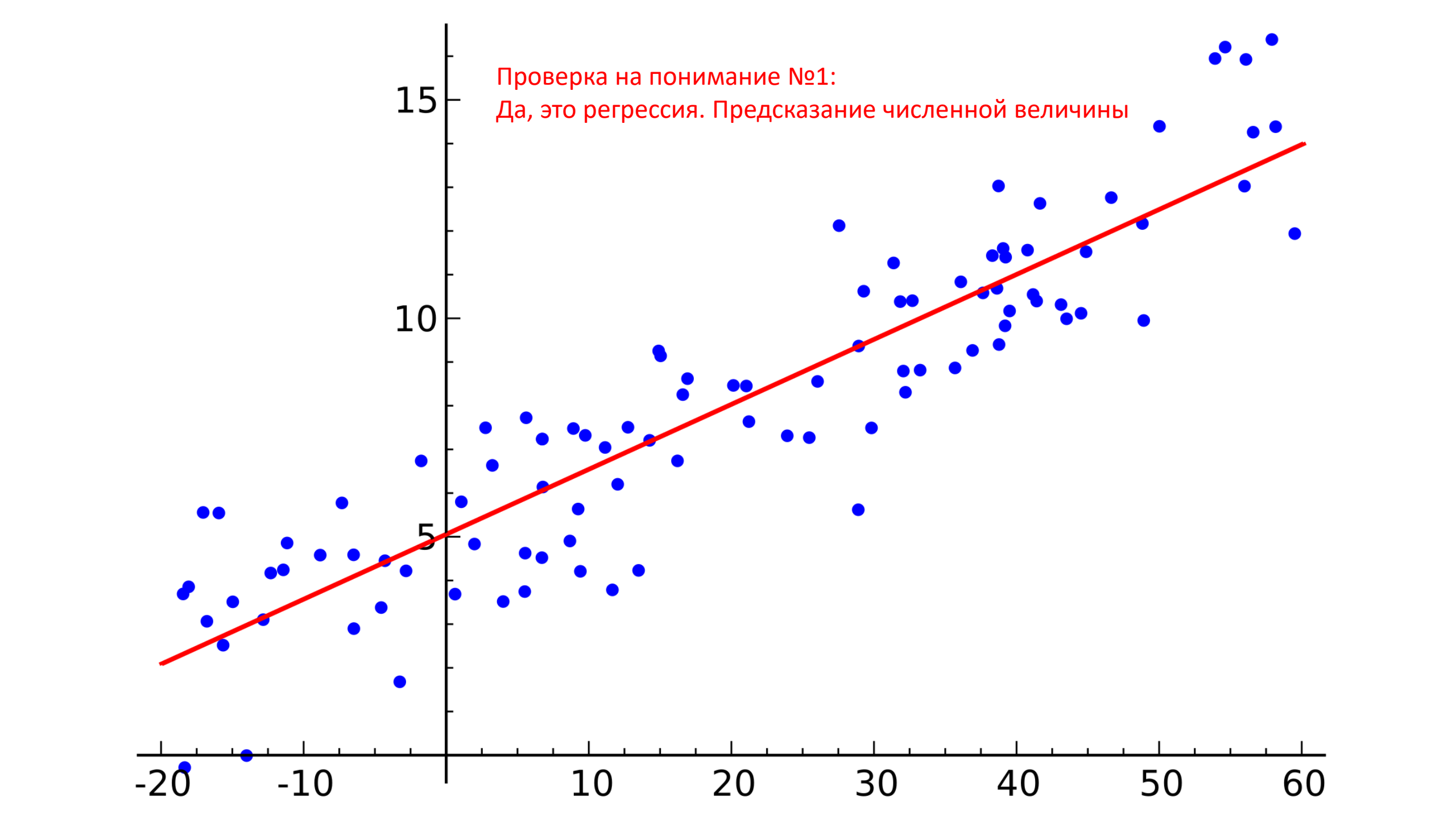
- Впрочем, хорошая новость заключается в том, что для задачи регрессии эти два понятия это очень часто одно и то же! :)
- Чего, однако, нельзя сказать о классификации…

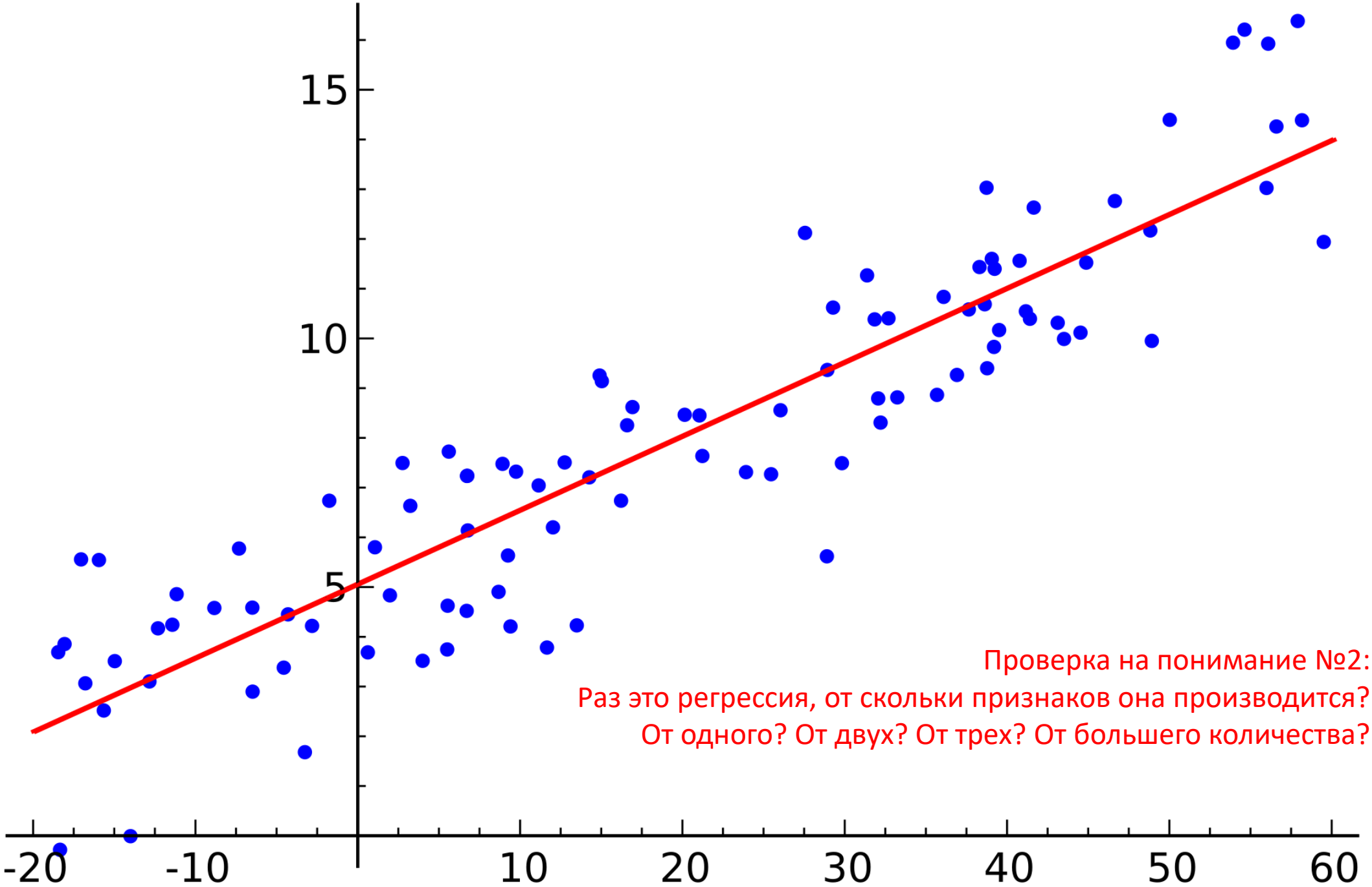
Метрики качества и функционал ошибки

- Впрочем, хорошая новость заключается в том, что для задачи регрессии эти два понятия это очень часто одно и то же! :)
- Чего, однако, нельзя сказать о классификации…
  - Кстати, как вы думаете, почему?

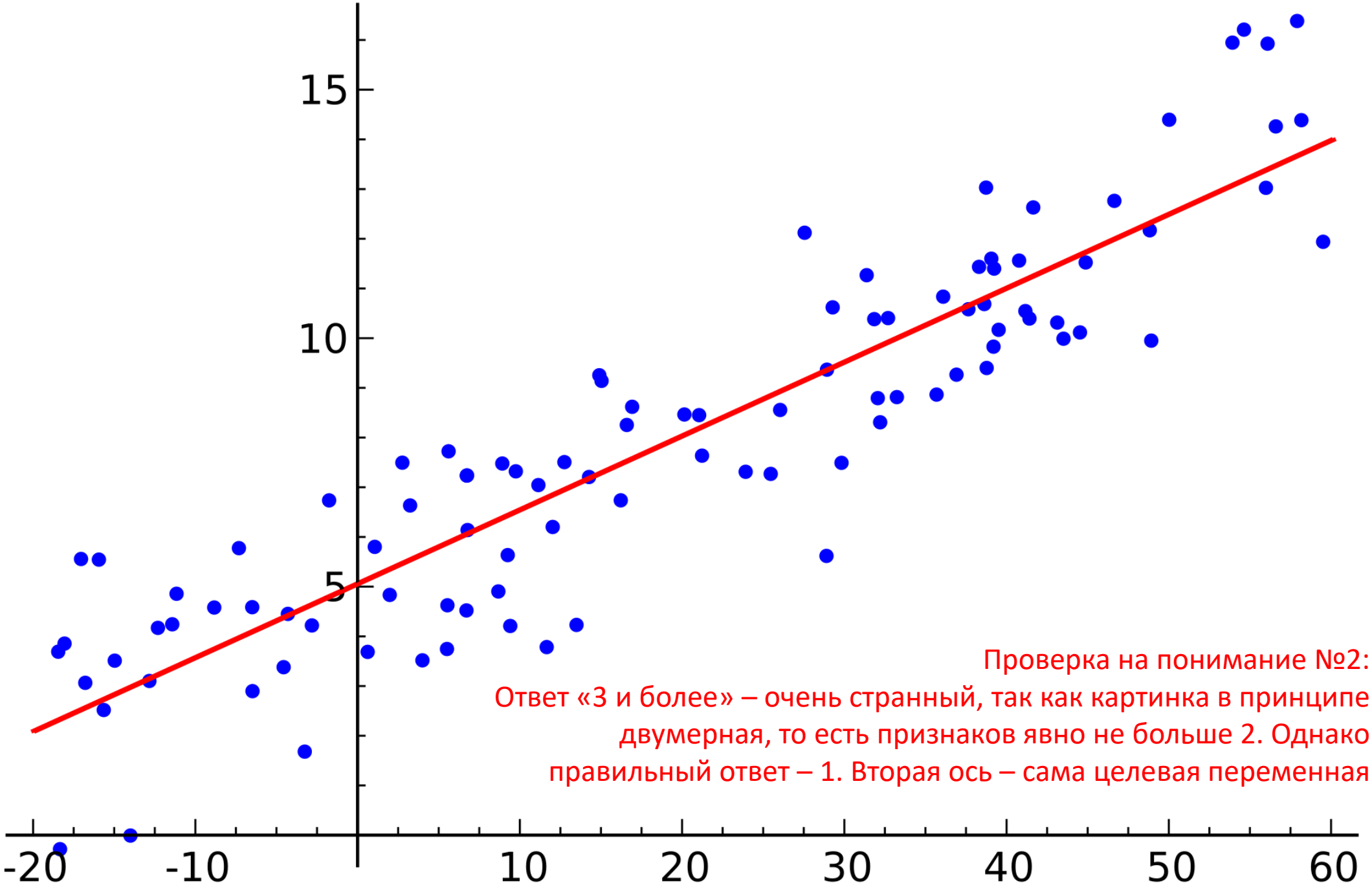

15
10
5
-20
-10
10
20
30
40
50
60


Проверка на понимание №1:
Здесь изображена задача регрессии или всё-таки классификации?
15
10
5
-20
-10
10
20
30
40
50
60


Проверка на понимание №1:
Да, это регрессия. Предсказание численной величины
15
10
5
-20
-10
10
20
30
40
50
60



Проверка на понимание №2:
Раз это регрессия, от скольки признаков она производится?
От одного? От двух? От трех? От большего количества?



Проверка на понимание №2:
Ответ «3 и более» – очень странный, так как картинка в принципе двумерная, то есть признаков явно не больше 2. Однако правильный ответ – 1. Вторая ось – сама целевая переменная

Метрики качества и функционал ошибки

- Какие метрики качества для регрессии вы знаете?

Метрики качества и функционал ошибки

- Какие метрики качества для регрессии вы знаете?

- Mean Squared Error (MSE):

$$\frac{1}{n}\sum_i^n(\widetilde{y}_i - y_i)^2$$

- Mean Absolute Error (MAE):

$$\frac{1}{n}\sum_i^n|\widetilde{y}_i - y_i|$$

- Max Error:

$$\max_i|\widetilde{y}_i - y_i|$$

Метрики качества и функционал ошибки

- Какие метрики качества для регрессии вы знаете?

- Mean Squared Error (MSE):

$$\frac{1}{n}\sum_i^n(\widetilde{y_i} - y_i)^2$$

- Mean Absolute Error (MAE):

$$\frac{1}{n}\sum_i^n|\widetilde{y_i} - y_i|$$

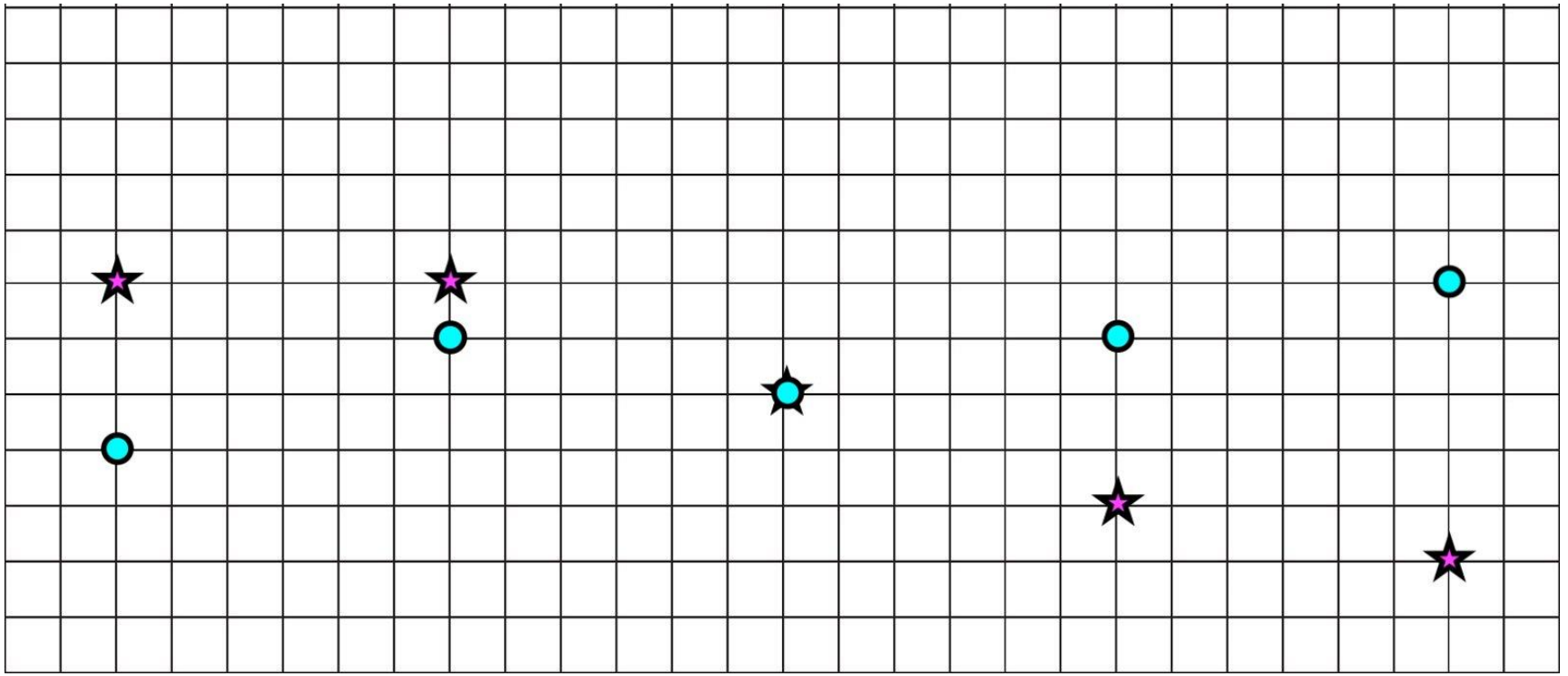
- Max Error:

$$\max_i|\widetilde{y_i} - y_i|$$

Прокомментируйте обозначения: что такое $\widetilde{y_i}$, y_i, что показывает n

Метрики качества и функционал ошибки

- Давайте мы порисуем эти метрики и потренируемся в самом простом случае!



Метрики качества и функционал ошибки

- Впрочем, у всех предыдущих метрик есть важный недостаток!
- Какой?

Метрики качества и функционал ошибки

- Впрочем, у всех предыдущих метрик есть важный недостаток!
- Какой?
- Они не позволяют оценить качество в абсолютном выражении, поскольку зависят от единиц измерения.
- Иными словами, вы сможете только сравнить две разные модели друг с другом по качеству, но не сказать, хорошие ли это в целом модели или же нет.

Метрики качества и функционал ошибки

- Впрочем, у всех предыдущих метрик есть важный недостаток!
- Какой?
- Они не позволяют оценить качество в абсолютном выражении, поскольку зависят от единиц измерения.
- Иными словами, вы сможете только сравнить две разные модели друг с другом по качеству, но не сказать, хорошие ли это в целом модели или же нет.
- Какие метрики решают эту проблему?

Метрики качества и функционал ошибки

- Coefficient of determination (R^2):

$$1 - \frac{\sum_i^n (\widetilde{y_i} - y_i)^2}{\sum_i^n (\overline{y_i} - y_i)^2} \in (-\infty, 1]$$

- Mean Absolute Percentage Error (MAPE):

$$100 \cdot \frac{1}{n} \sum_i^n \left| \frac{y_i - \widetilde{y_i}}{y_i} \right| \in [0, +\infty)$$

Метрики качества и функционал ошибки

- Coefficient of determination (R^2):

$$1 - \frac{\sum_i^n (\widetilde{y_i} - y_i)^2}{\sum_i^n (\overline{y_i} - y_i)^2} \in (-\infty, 1]$$

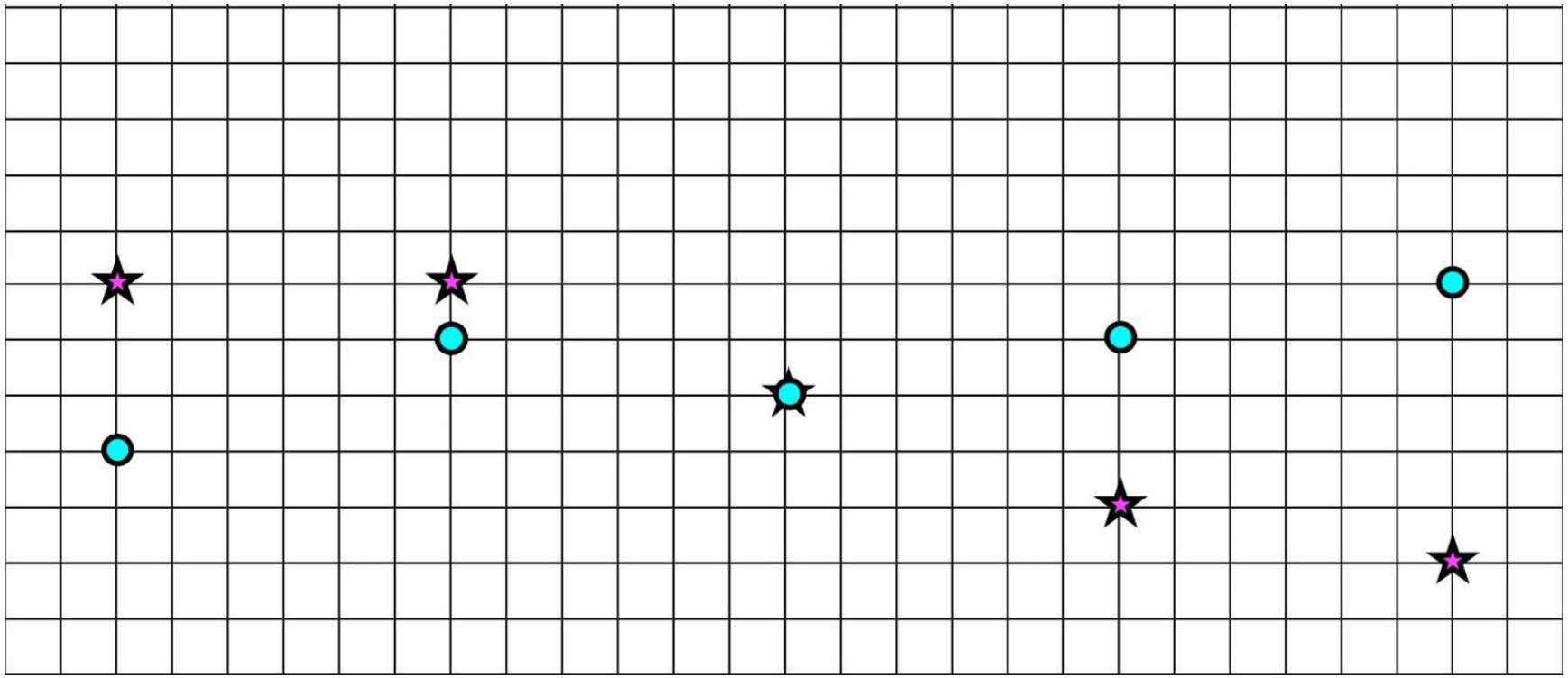
- Mean Absolute Percentage Error (MAPE):

$$100 \cdot \frac{1}{n} \sum_i^n \left| \frac{y_i - \widetilde{y_i}}{y_i} \right| \in [0, +\infty)$$

Какие значения данных метрик будут у наилучшей модели?
Чем эти метрики лучше предыдущих?

Метрики качества и функционал ошибки

- Потренируемся и с ними!



Алгоритм KNN

Алгоритм KNN

- Давайте теперь вернемся к некогда рассматривавшемуся нами алгоритму – алгоритму К ближайших соседей.
- Теперь мы уже знаем всю необходимую математику, чтобы обсудить его полноценно!

Алгоритм KNN

- Давайте теперь вернемся к некогда рассматривавшемуся нами алгоритму – алгоритму К ближайших соседей.
- Теперь мы уже знаем всю необходимую математику, чтобы обсудить его полноценно!
- Вспомним основную идею.

Алгоритм KNN

- На вход подается вектор — признаковое описание какого-то объекта

Алгоритм KNN

- На вход подается вектор — признаковое описание какого-то объекта
- Находится K ближайших к нему векторов, для которых ответ известен

Алгоритм KNN

- На вход подается вектор — признаковое описание какого-то объекта
- Находится К ближайших к нему векторов, для которых ответ известен

Именно в этом месте у нас была ранее основная загвоздка! Очень скоро вернемся к этому!

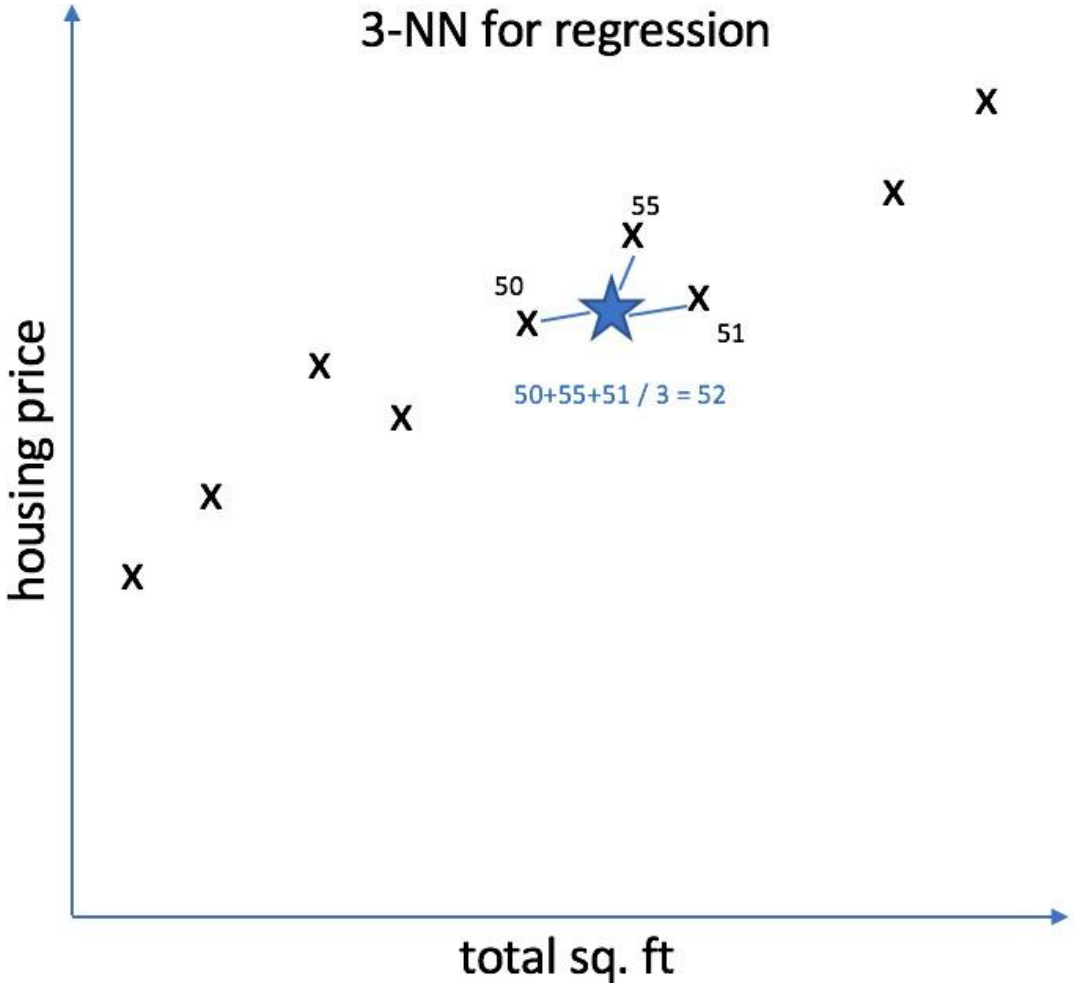
Алгоритм KNN

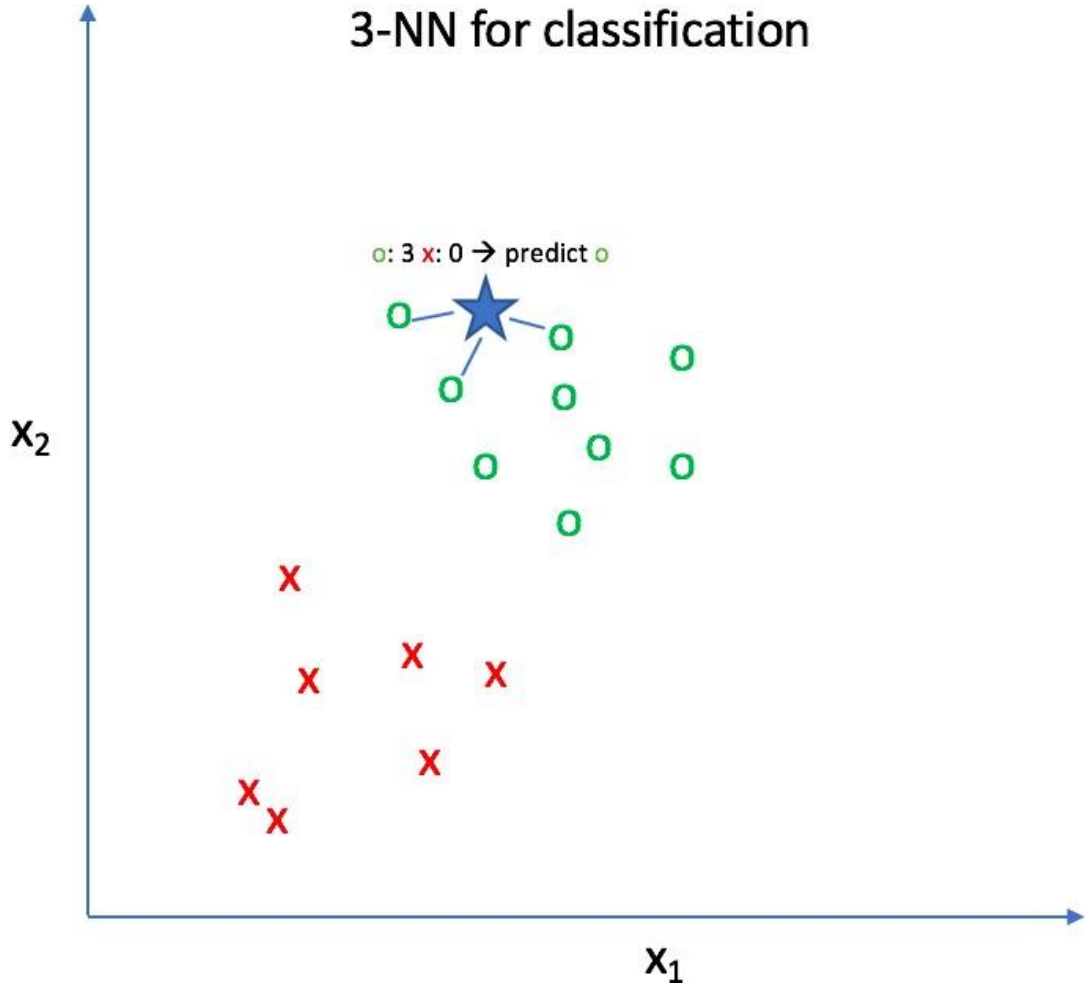
- На вход подается вектор — признаковое описание какого-то объекта
- Находится K ближайших к нему векторов, для которых ответ известен
- Ответ для нового объекта выбирается с помощью:
  - Усреднения, в случае регрессии
  - Голосования, в случае классификации

Алгоритм KNN

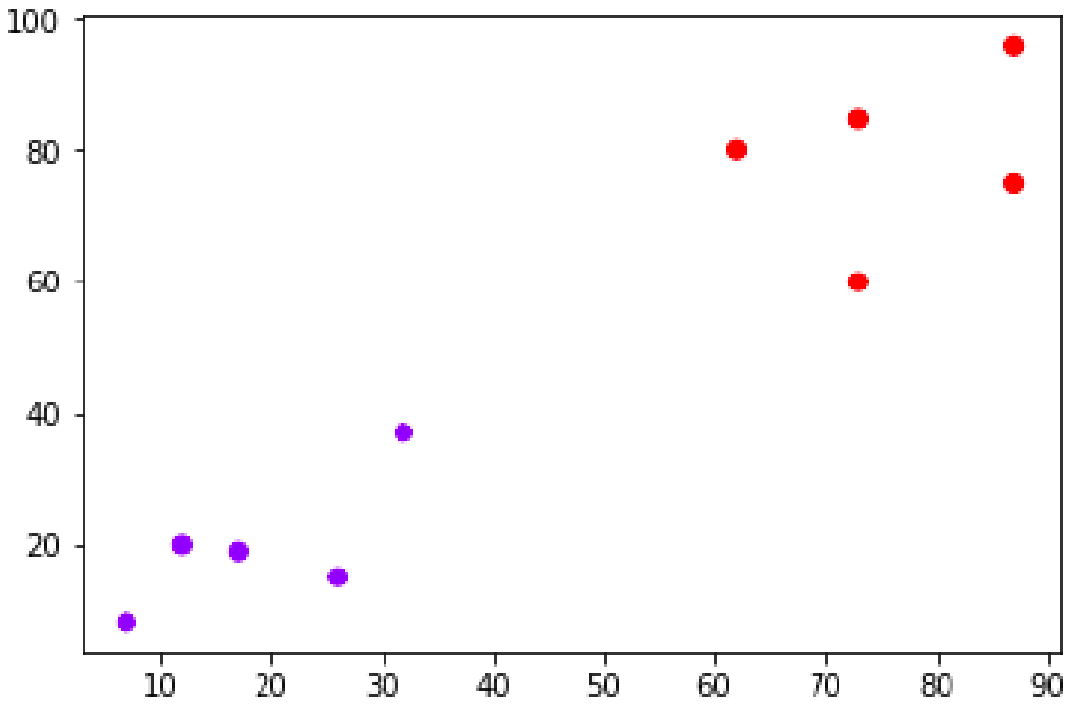
- На вход подается вектор — признаковое описание какого-то объекта
- Находится K ближайших к нему векторов, для которых ответ известен
- Ответ для нового объекта выбирается с помощью:
  - Усреднения, в случае регрессии
  - Голосования, в случае классификации
- Возможно также усреднение/голосование с весами и многие другие модификации стандартного алгоритма

Алгоритм KNN

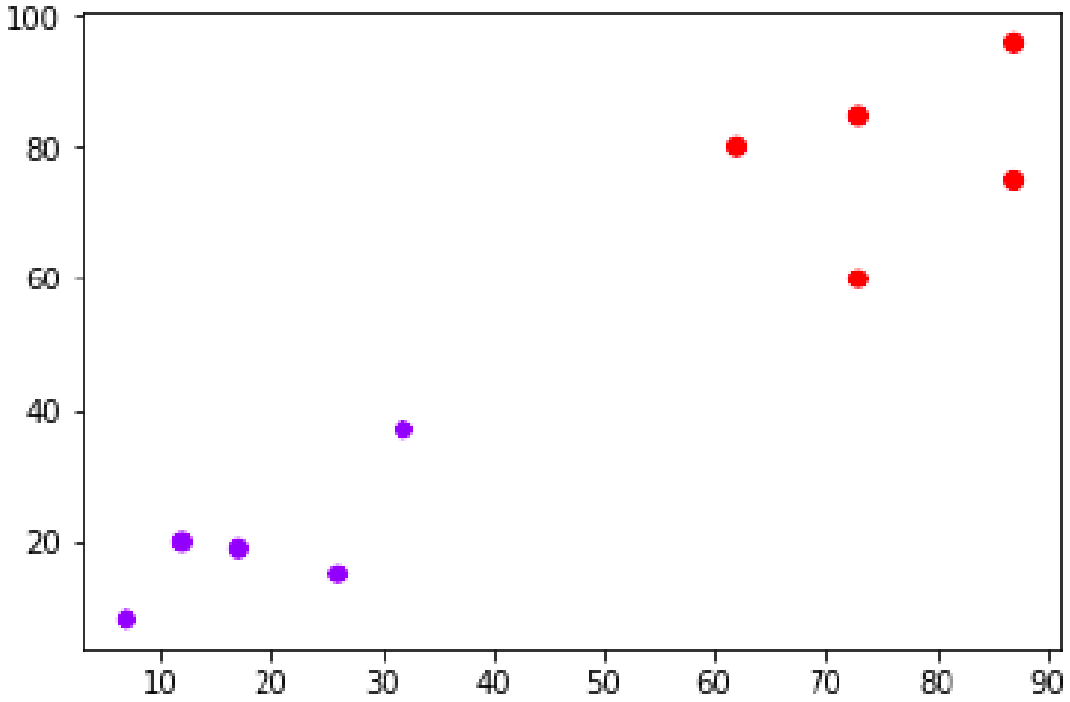


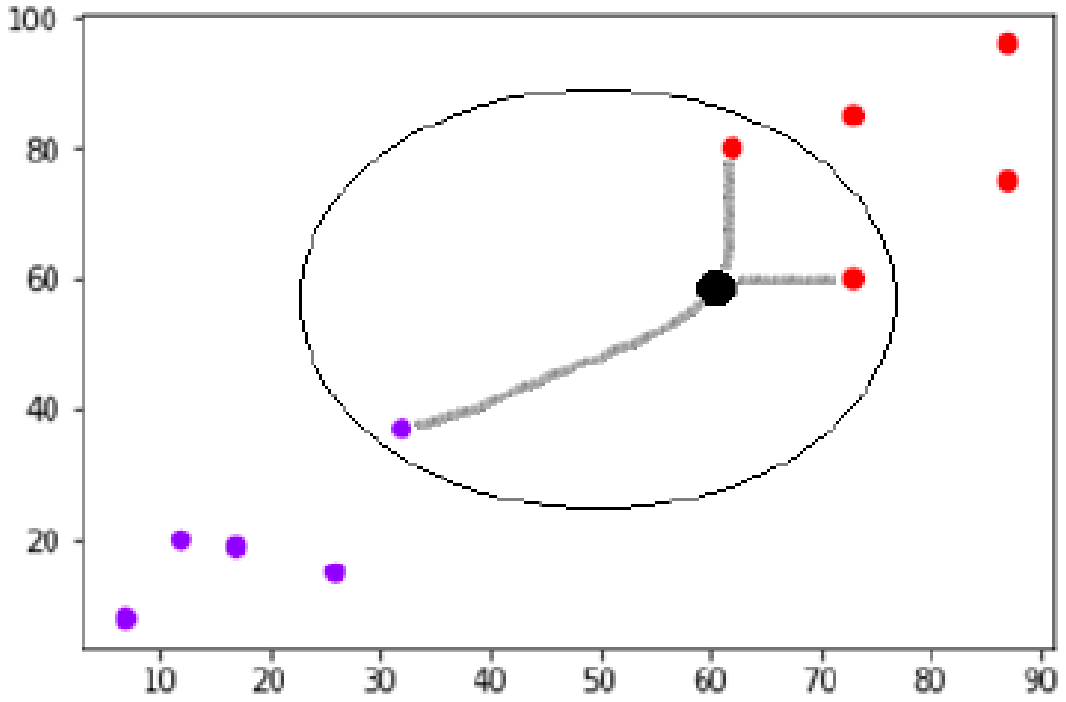


Алгоритм KNN

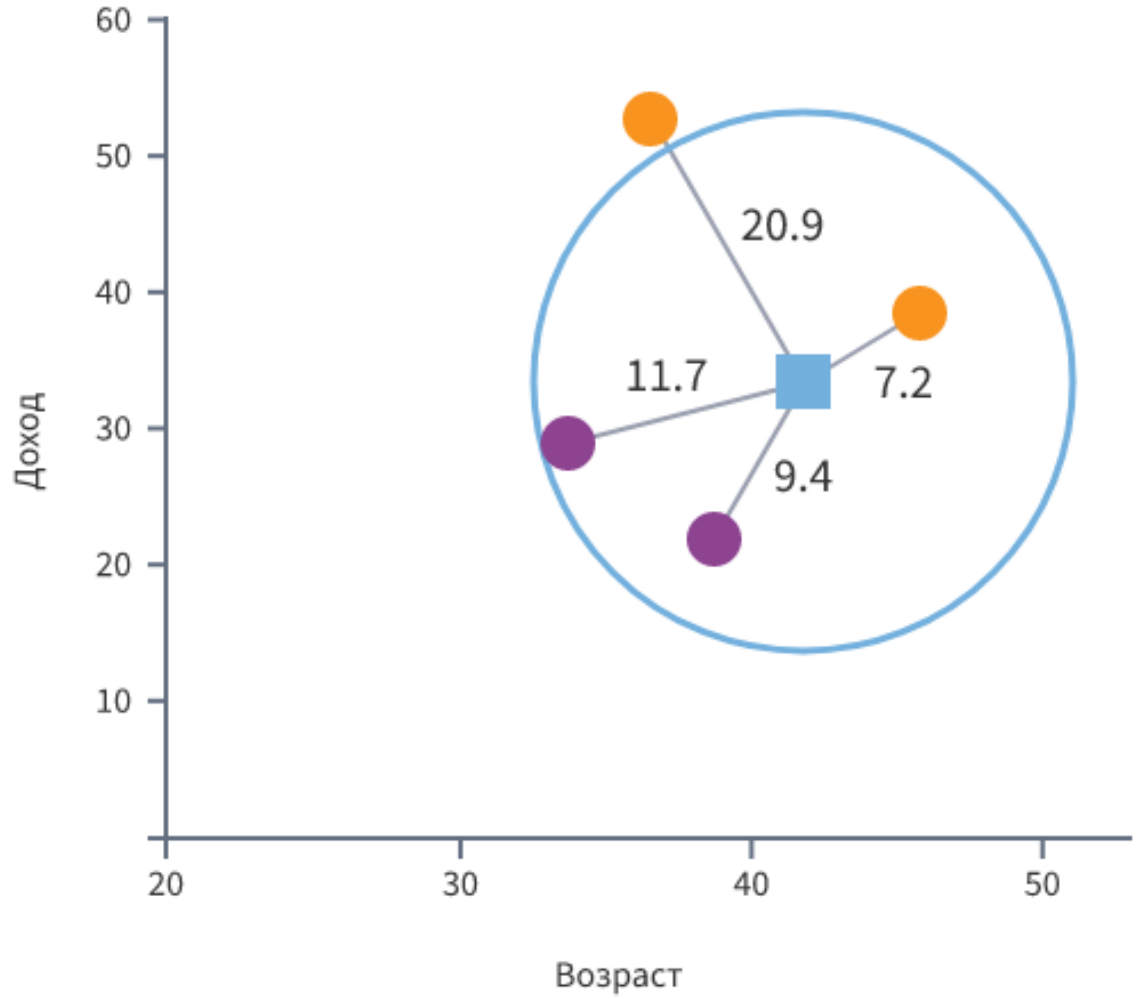


Алгоритм KNN

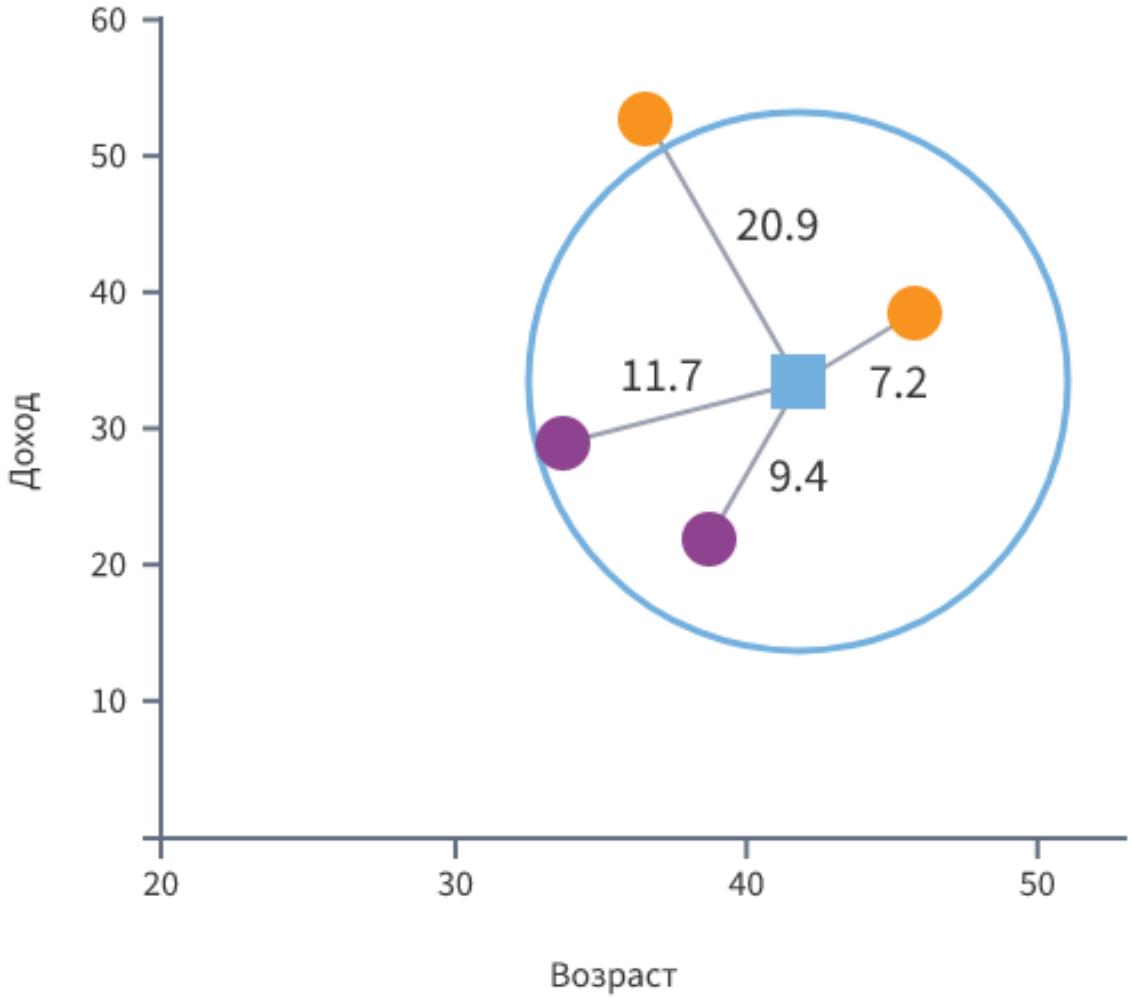




Алгоритм KNN

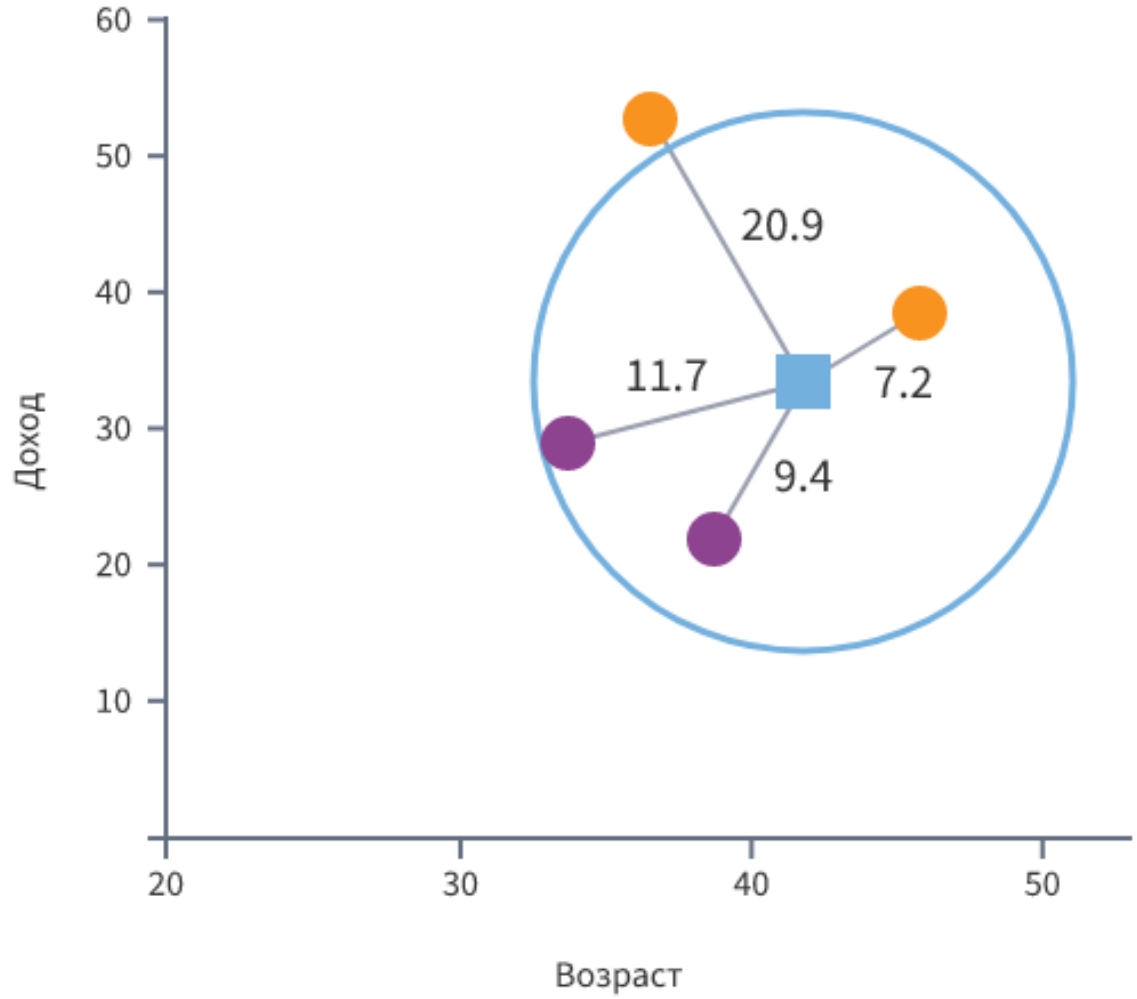


Алгоритм KNN



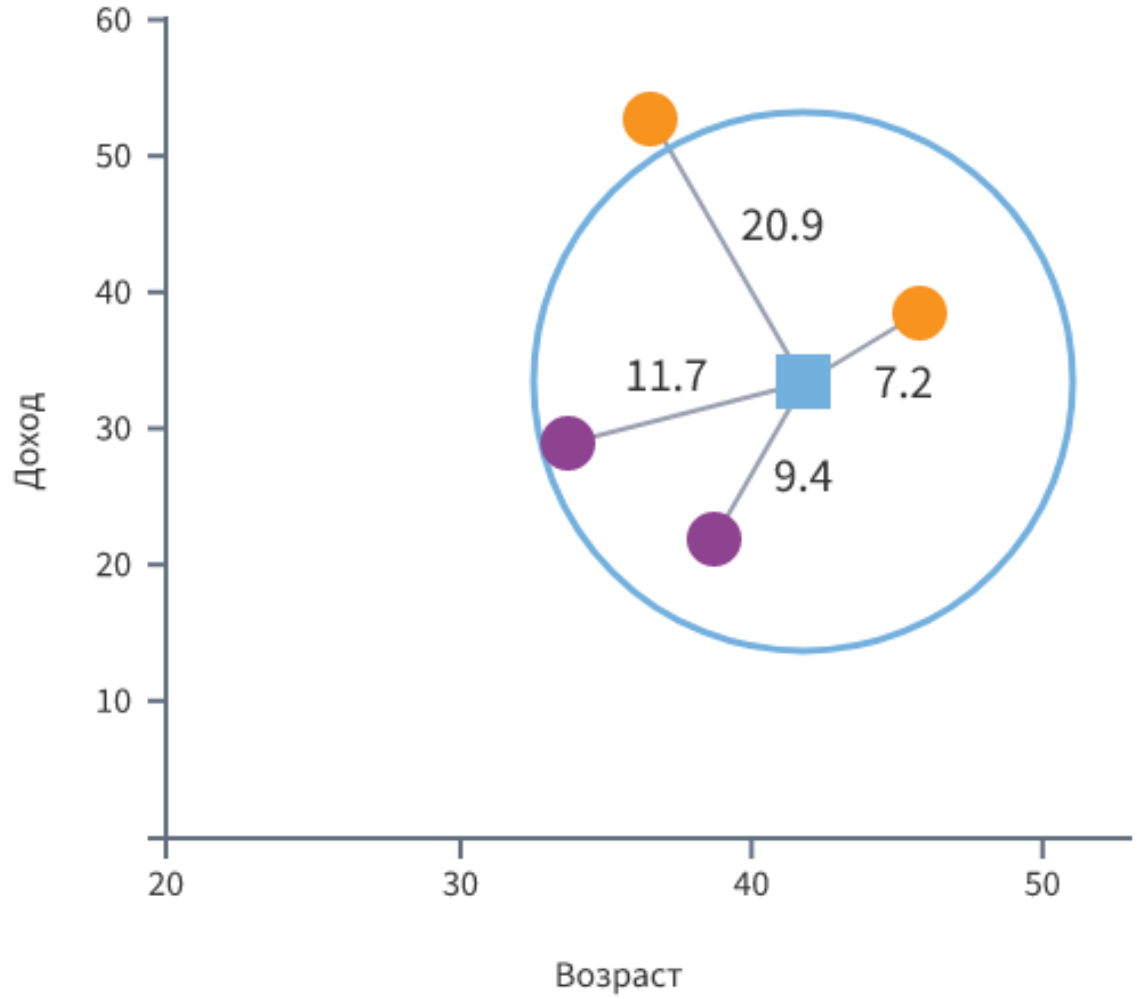
Проверка на понимание №1:
Здесь изображена задача регрессии или всё-таки классификации?

Алгоритм KNN



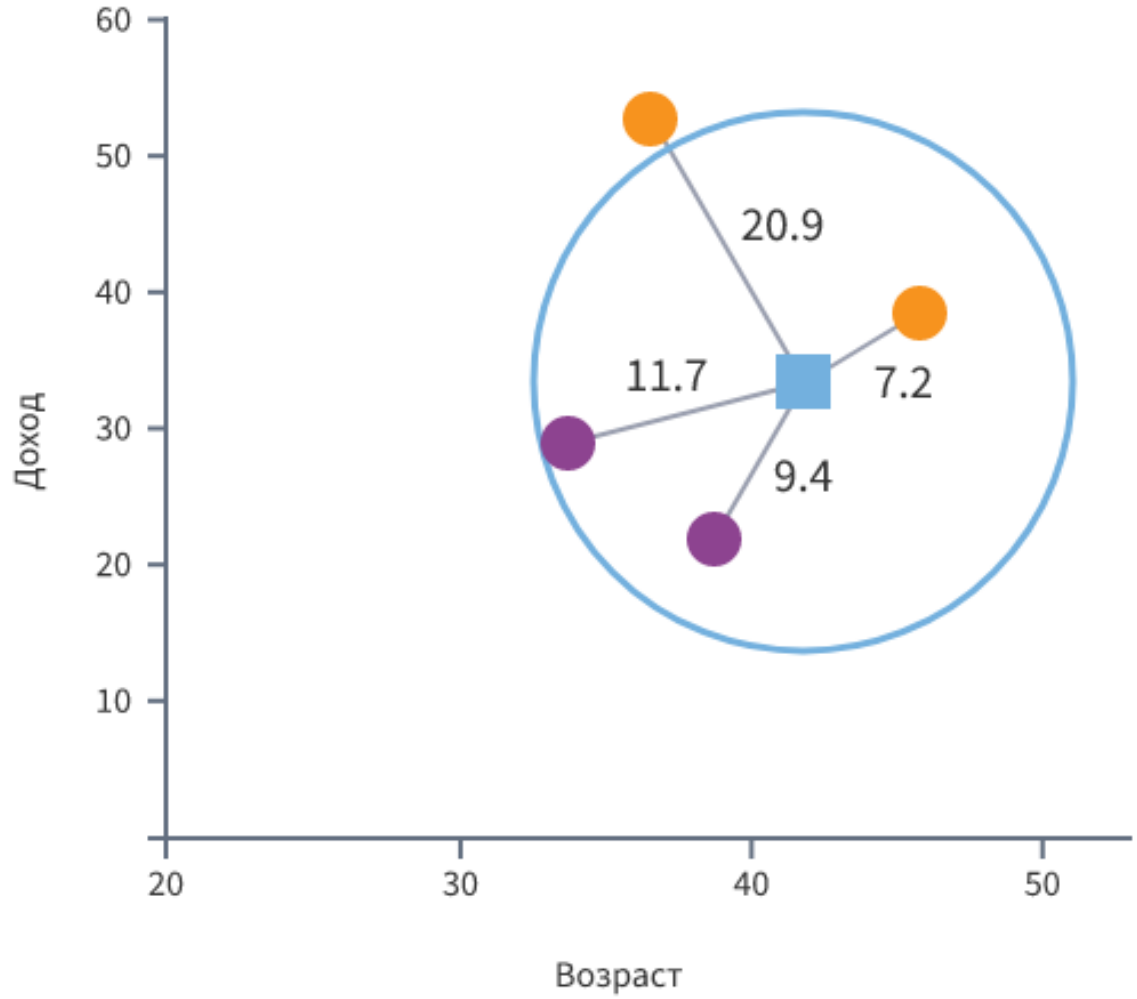
Проверка на понимание №1:
И та, и другая задача!

Алгоритм KNN



Проверка на понимание №2:
Раз это регрессия, от скольки признаков она производится? От одного? От двух? От трех? От большего количества?

Алгоритм KNN



Проверка на понимание №2:
Здесь уже от двух, а не от одного!
Целевая переменная (её значение) указана рядом с точками просто в виде лейбла (цвета или числа)

Метрические пространства

- Ну а теперь, как мы и собирались, вернемся к главному вопросу: а как же всё-таки искать расстояние между «похожими» точками (векторами)?

Метрические пространства

- Ну а теперь, как мы и собирались, вернемся к главному вопросу: а как же всё-таки искать расстояние между «похожими» точками (векторами)?
- Для этого нам потребуется ввести важное понятие – метрическое пространство!

Метрические пространства

- Ну а теперь, как мы и собирались, вернемся к главному вопросу: а как же всё-таки искать расстояние между «похожими» точками (векторами)?
- Для этого нам потребуется ввести важное понятие – метрическое пространство!
- Метрическое пространство – это такое пространство, на котором задана некая метрика (функция), позволяющая между любыми двумя точками этого пространства посчитать расстояние. Данная метрика называется метрикой расстояния (или функцией расстояния).

Метрические пространства

- Приведите примеры каких-нибудь метрических пространств

Метрические пространства

- Приведите примеры каких-нибудь метрических пространств
- Самый простой и близкий нам пример – это карта. Например, в Яндекс.Картах.

Метрические пространства

- Приведите примеры каких-нибудь метрических пространств
- Самый простой и близкий нам пример – это карта. Например, в Яндекс.Картах.
- Вопросы к размышлению:
  - Как на картах посчитать расстояние между начальной точкой маршрута и конечной точкой маршрута?
  - Как это сделать на листке бумаги между двумя точками?
  - А как это сделать в нашем с вами трехмерном пространстве?


Middle Earth
the Sundering Seas
Forodwaith
Northern Waste
Forochel
Arnor
Angmar
Grey Mountains
Withered Heath
Ettenmoors
Misty Mountains
Iron Hills
Erebor
Rhovanion
Mirkwood
Rhûn
Forlindon
Blue Mountains
Harlindon
Eriador
the Shire
Minhiriath
Dunland
Enedwaith
the Brown Lands
Rohan
Anórien
White Mountains
Gondor
Ithilien
Ash Mountains
Mount Doom
Gorgoroth
Mordor
Nurn
Mountains of Shadow
Mountains of the East
Bay of Belfalas
Harondor
Khând
Near Harad
Haradwaith
Umbar

Метрики расстояний

- Стандартная, привычная для нас метрика расстояния, доказываемая через теорему Пифагора и расширенная на многомерный случай
- Называется Евклидова метрика

Метрики расстояний

- Стандартная, привычная для нас метрика расстояния, доказываемая через теорему Пифагора и расширенная на многомерный случай
- Называется Евклидова метрика

$$\rho(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Метрики расстояний

- Стандартная, привычная для нас метрика расстояния, доказываемая через теорему Пифагора и расширенная на многомерный случай
- Называется Евклидова метрика

$$\rho(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Вопрос на понимание: что в этой формуле представляет собой i?

Метрики расстояний

- Метрика, которая нередко оказывается эффективнее Евклидовой, засчет структуры, близкой ко многим предметным областям (например, тем же картам)
- Манхэттенская метрика.. :)

Метрики расстояний

- Метрика, которая нередко оказывается эффективнее Евклидовой, засчет структуры, близкой ко многим предметным областям (например, тем же картам)
- Манхэттенская метрика

$$\rho(x, y) = \sum_i |x_i - y_i|$$

Метрики расстояний

- Метрика, которая нередко оказывается эффективнее Евклидовой, засчет структуры, близкой ко многим предметным областям (например, тем же картам)
- Манхэттенская метрика

$$\rho(x, y) = \sum_i |x_i - y_i|$$

В чем преимущество такой метрики и причем здесь Манхэттен? :)

Метрики расстояний

- Общий случай – метрика Миньковского

Метрики расстояний

- Общий случай – метрика Миньковского

$$\rho(x, y) = \sqrt[p]{\sum_i |x_i - y_i|^p}$$

Алгоритм KNN

- А теперь вернемся к постановке задачи KNN.
- Пусть у нас есть данные, которые записаны привычным для нас образом, в виде матрицы признаков размера $M \times N$, где M – число объектов, а N – число признаков.

Алгоритм KNN

- А теперь вернемся к постановке задачи KNN.
- Пусть у нас есть данные, которые записаны привычным для нас образом, в виде матрицы признаков размера $M \times N$, где M – число объектов, а N – число признаков.
- Сделаем предположение, что все N признаков находятся в одном метрическом пространстве. Если это так, то каждый из M объектов является просто точкой в N-мерном метрическом пространстве, а это значит, что мы можем посчитать расстояние между любыми двумя точками.

Алгоритм KNN

- Сделаем предположение, что все N признаков находятся в одном метрическом пространстве. Если это так, то каждый из M объектов является просто точкой в N-мерном метрическом пространстве, а это значит, что мы можем посчитать расстояние между любыми двумя точками.
- Тогда, если мы выберем какую-нибудь точку, мы сможем сказать какая из $M - 1$ оставшихся точек будет являться самой ближайшей к рассматриваемой – такую точку назовем ближайшим соседом.
- Аналогичным образом можно определить второго по близости соседа и т.д.

Алгоритм KNN

- Ну а далее, поскольку мы приняли, что наше пространство метрическое, мы можем утверждать, что точки, расстояние между которыми небольшое, будут похожи между собой, а те, у которых расстояние будет большим, будут являться непохожими.

Алгоритм KNN

- Ну а далее, поскольку мы приняли, что наше пространство метрическое, мы можем утверждать, что точки, расстояние между которыми небольшое, будут похожи между собой, а те, у которых расстояние будет большим, будут являться непохожими.
- А значит алгоритм KNN является обоснованным и применимым!
- Ура!

Алгоритм KNN

- Важная ремарка (которую мы уже косвенно проговаривали).
- При использовании алгоритма KNN, «обучения» как такового не происходит. Единственное, что, по сути, вам нужно сделать – это «запомнить» для точек их расположение в пространстве, а также их таргеты.

Алгоритм KNN

- Важная ремарка (которую мы уже косвенно проговаривали).
- При использовании алгоритма KNN, «обучения» как такового не происходит. Единственное, что, по сути, вам нужно сделать – это «запомнить» для точек их расположение в пространстве, а также их таргеты.
- В дальнейшем предсказание строится просто путём усреднения таргета по ближайшим «хранящимся в памяти» точкам для обучающей выборки.

Алгоритм KNN

- Важная ремарка (которую мы уже косвенно проговаривали).
- При использовании алгоритма KNN, «обучения» как такового не происходит. Единственное, что, по сути, вам нужно сделать – это «запомнить» для точек их расположение в пространстве, а также их таргеты.
- В дальнейшем предсказание строится просто путём усреднения таргета по ближайшим «хранящимся в памяти» точкам для обучающей выборки.
- Именно это делает алгоритм KNN – непараметрическим!

Алгоритм KNN

- Важная ремарка (которую мы уже косвенно проговаривали).
- При использовании алгоритма KNN, «обучения» как такового не происходит. Единственное, что, по сути, вам нужно сделать – это «запомнить» для точек их расположение в пространстве, а также их таргеты.
- В дальнейшем предсказание строится просто путём усреднения таргета по ближайшим «хранящимся в памяти» точкам для обучающей выборки.
- Именно это делает алгоритм KNN – непараметрическим!
  - А вот гиперпараметры у алгоритма очень даже есть: важнейшим является K – количество соседей, по которому будет вестись расчет.

Алгоритм KNN

Логично вытекающее и при этом важное уточнение: а чем вообще отличаются параметры и гиперпараметры алгоритма?

- Важная ремарка (которую мы уже косвенно проговаривали).
- При использовании алгоритма KNN, «обучения» как такового не происходит. Единственное, что, по сути, вам нужно сделать – это «запомнить» для точек их расположение в пространстве, а также их таргеты.
- В дальнейшем предсказание строится просто путём усреднения таргета по ближайшим «хранящимся в памяти» точкам для обучающей выборки.
- Именно это делает алгоритм KNN – непараметрическим!
  - А вот гиперпараметры у алгоритма очень даже есть: важнейшим является K – количество соседей, по которому будет вестись расчет.