

# Регрессия и использование библиотеки `Scikit-learn` в задачах обучения с учителем

## Урок 3

На этой лекции вы узнаете:

- Что такое обучение с учителем.
- Что такое линейная регрессия
- Как считать метрики качества модели линейной регрессии.



## Булгакова Татьяна

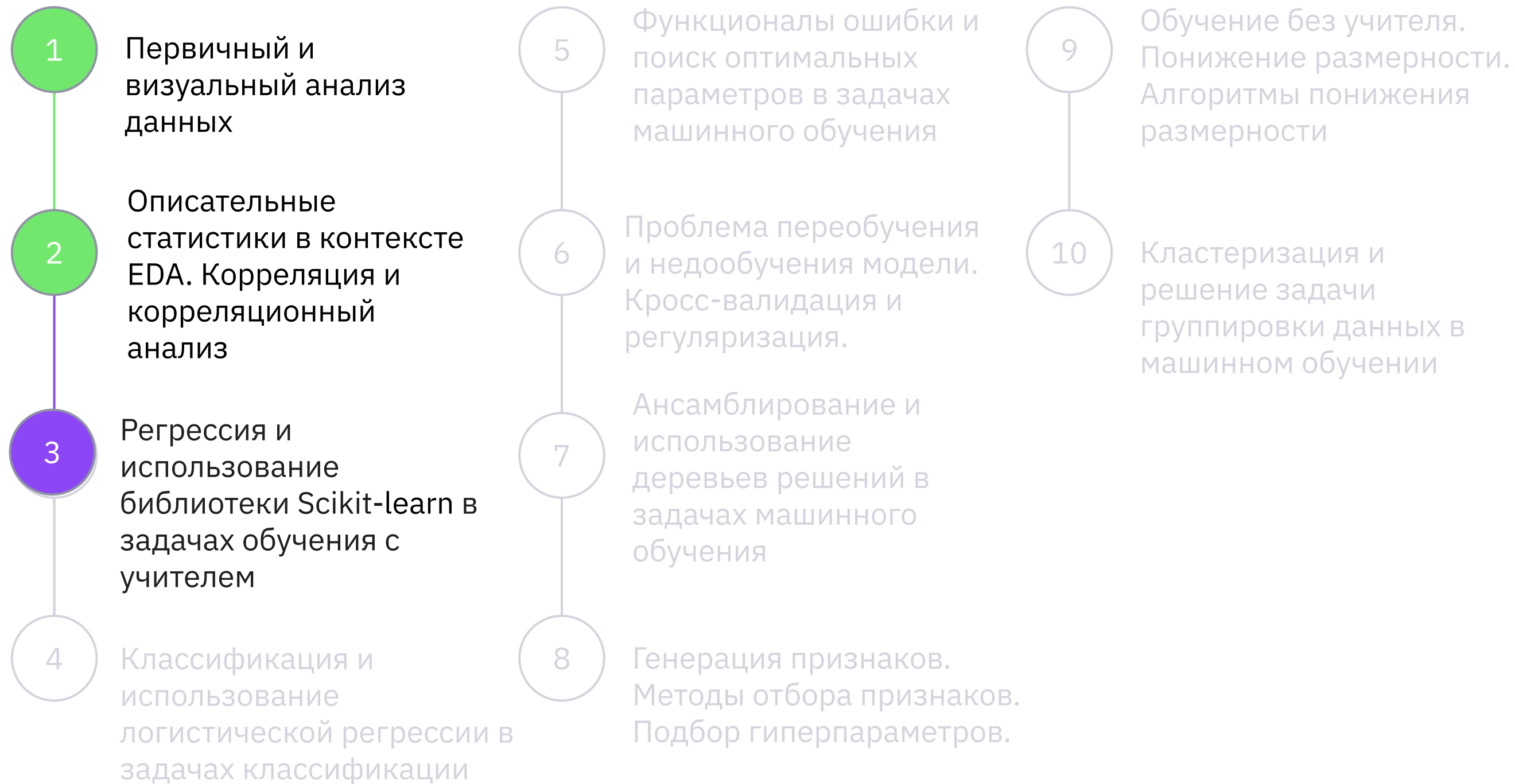
Преподаватель в GeekBrains, Нетология, Skillfactory

С 2010 года занимаюсь DataScience и NN. Фрилансер

- Участвовала в разработке программы по настройке оборудования для исследования пространственного слуха китообразных НИИ ИПЭЭ РАН
- Участвую в разработке рекомендательных систем по настройке нейростимуляторов для медицинских центров
- Работаю над курсом по нейронным сетям






# План курса





## Что будет на уроке сегодня

-  Что такое обучение с учителем.
-  Что такое линейная регрессия
-  Метрики задачи регрессии





# Задачи машинного обучения





## Задачи машинного обучения имеют много общего.



**Во-первых,** их решения можно описать как функции, которые отображают объекты или примеры (samples) в предсказания (target).



**Во-вторых,** эти задачи вряд ли имеют единственное идеальное решение.



**В-третьих,** Функция, которая сопоставляет объекты с предсказаниями, называется моделью, а набор доступных примеров - обучающей выборкой или набором данных.



## Виды машинного обучения

**Большинство методов машинного обучения относят к обучению с учителем (supervised learning) и без учителя (unsupervised learning). «Учитель» здесь – это не конкретный человек, а сам факт вмешательства в процесс обработки данных.**

### **С учителем**

В идеальном случае у нас есть исходные данные, т. е. правильные ответы для системы.

### **Без учителя**

Здесь готовых ответов нет, но менее интересно не становится, даже наоборот. Представьте, что мы знаем рост и вес большой группы людей, в соответствии с чем нужно пошить одежду трёх видов. Это кластеризация (строгого и единственно верного деления тут нет).



**Задача обучения с учителем** - В зависимости от содержания обучающего множества задачи контролируемого обучения могут быть следующих типов:

### **Задача регрессии**

предсказания  
вещественного  
значения: примерами  
задач регрессии  
является предсказание  
продолжительности  
поездки на  
каршеринге, спрос на  
конкретный товар в  
конкретный день

### **Задача классификации**

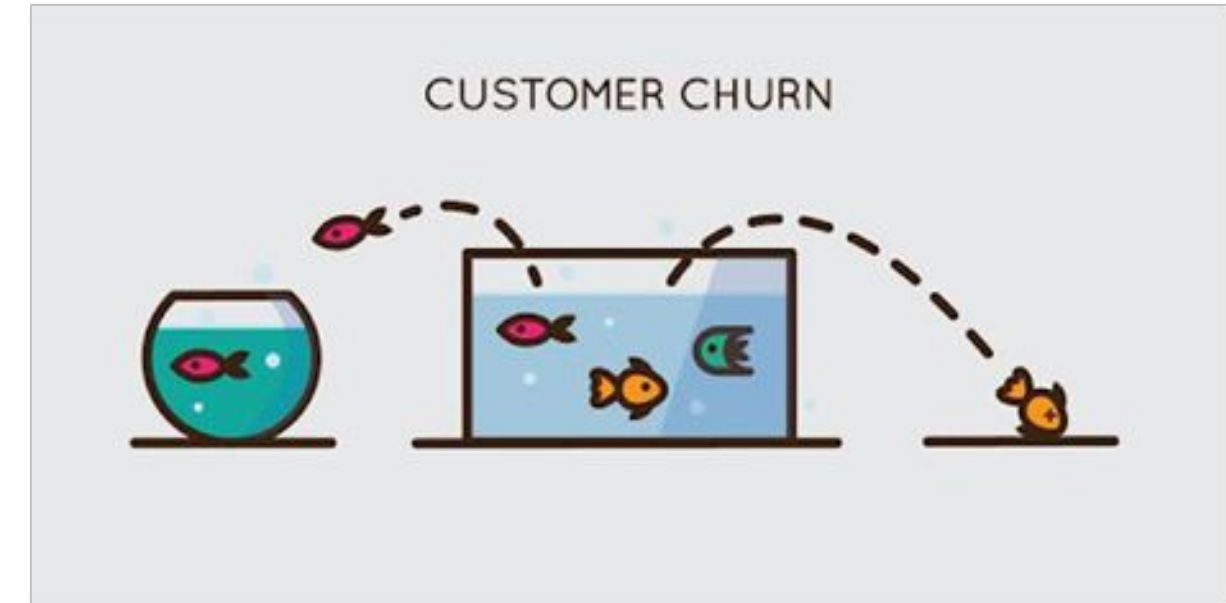
– предсказания  
категориального  
ответа (метки класса)  
с конечным  
количеством  
вариантов

### **Ранжирование**

Основным примером  
является задача  
ранжирования в  
поисковой системе.  
Здесь для любого  
заданного запроса все  
возможные документы  
должны быть  
ранжированы в  
соответствии с их  
релевантностью  
запросу.

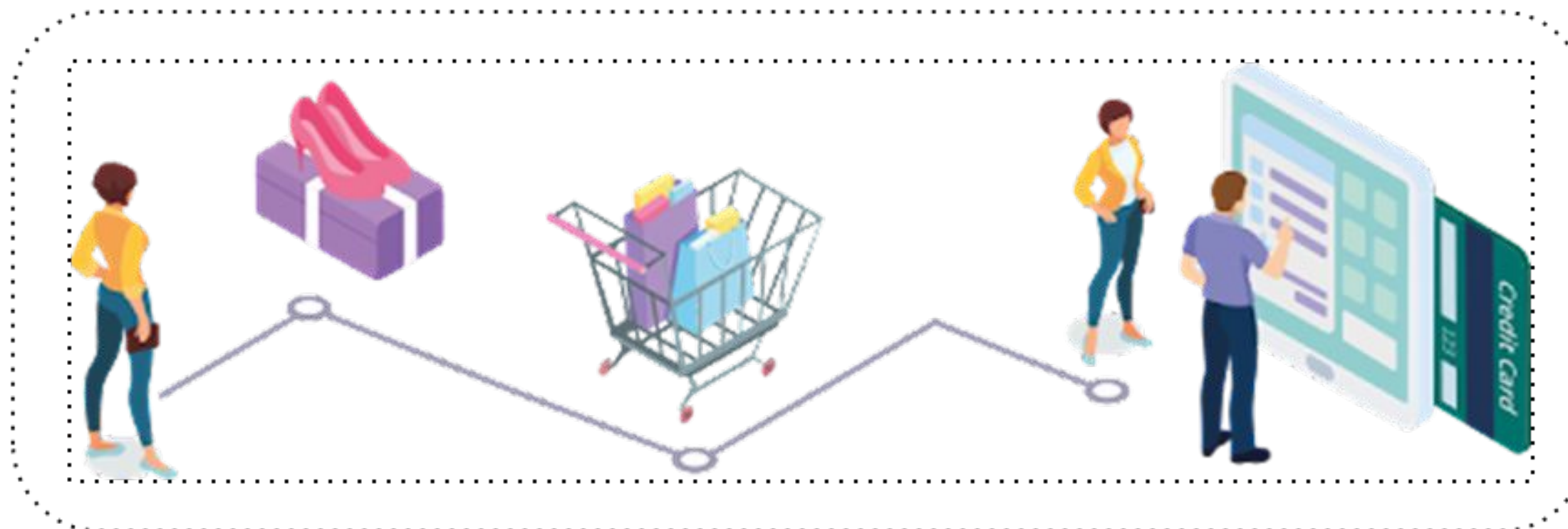


# Классификация





# Регрессия





# Регрессия

## **Задание:**

Определите тип задачи: Предсказание курса евро к доллару на следующий день.

**Ответ:** Это задача регрессии. Модель предсказывает вещественное число



## Критерии качества.

**Числовые** - например, рост, доход и т.д. Можно выделить вещественные и целочисленные атрибуты.

**Категориальные** атрибуты принимают значения из дискретного набора. Например, профессия человека или день недели.

**Бинарные** атрибуты принимают два значения: 0 и 1 или "да" и "нет". Его можно рассматривать и как числовой, и как категориальный атрибут.

**Категориальный** атрибут иногда также называют порядковым атрибутом. Этот атрибут принимает значения из упорядоченного дискретного множества. Например, класс опасности химического вещества (1-4) или продолжительность обучения студента в магистратуре являются порядковыми атрибутами.





## **Критерии качества.**

Создание информативных описаний признаков имеет решающее значение для дальнейшего анализа. Однако следует также обратить внимание на качество полученных данных.

- ☐ **Пропуски** (отсутствующие значения).
- ☐ **Выбросы** - это объекты, которые значительно отличается от других.
- ☐ **Ошибка разметки.**
- ☐ **Дрейф данных.** Данные могут меняться с течением времени.





## Обучение алгоритма машинного обучения

Модель можно представить как **функцию с параметрами**, где  $\theta$  - параметры алгоритма,  $\varepsilon$  - неустраняемая ошибка

$$y = f(\theta) + \varepsilon$$

**Параметры алгоритма** можно разделить на обучаемые (просто параметры) и необучаемые (гиперпараметры). Параметры модели задают семейство функций

$$f(g, w)$$



## Линейные модели

Линейные модели предполагают, что определяемый критерий линейно зависит от признаков описывающих объект или процесс

Линейная модель является прозрачной и понятной для аналитика. По полученным коэффициентам регрессии можно судить о том, как тот или иной фактор влияет на результат и сделать на этой основе дополнительные полезные выводы

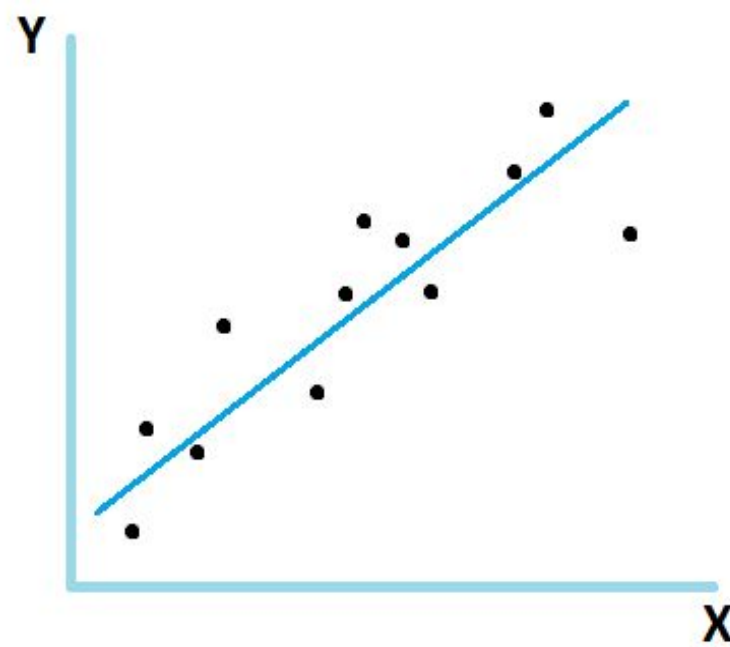
Большое количество реальных процессов в экономике и бизнесе можно с достаточной точностью описать линейными моделями

Для линейной регрессии известны типичные проблемы (например, мультиколлинеарность) и их решения. Разработаны и реализованы тесты оценки статистической значимости получаемых моделей



# Линейная регрессия

Регрессионная модель - это функция, которая принимает на вход значения атрибутов данного объекта и выдает на выходе ожидаемое значение целевой переменной.



$y$  - целевая переменная  
 $\mathbf{w}$  - вектор весов модели  
 $\mathbf{X}$  - матрица наблюдений  
 $\mathbf{e}$  - ошибка модели

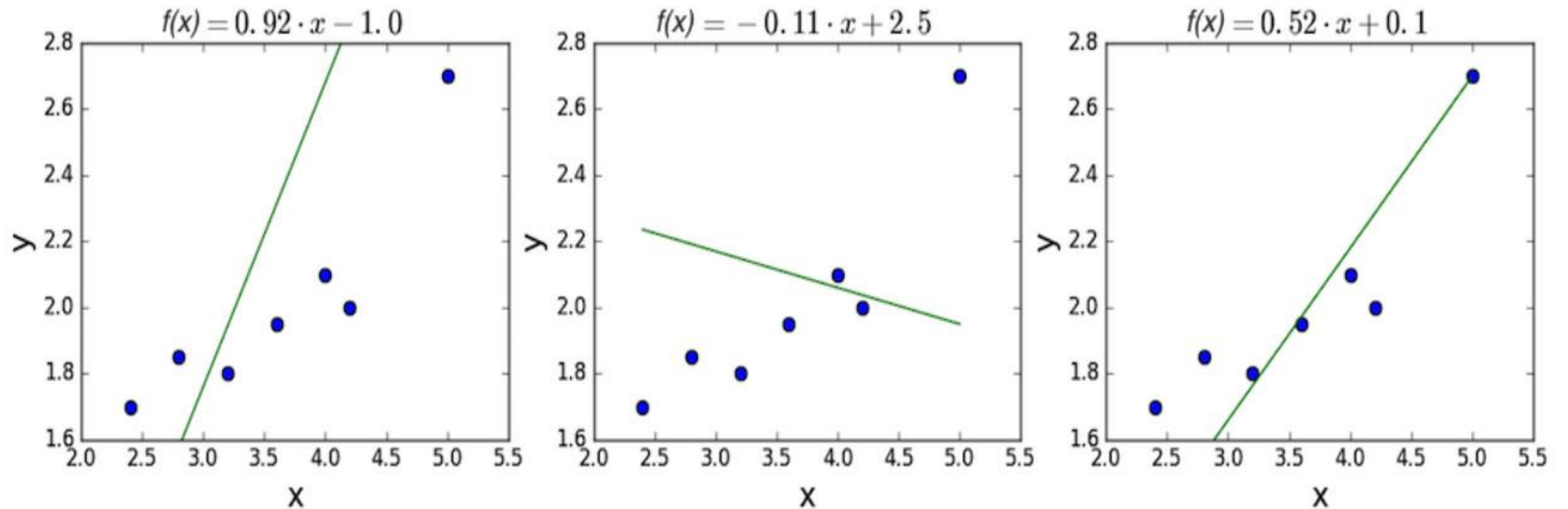
$$y = w_1x_1 + \dots + w_Dx_D + w_0$$

$$y_i = \sum_{j=1}^m w_j X_{ij} + e_i$$



# Линейная регрессия

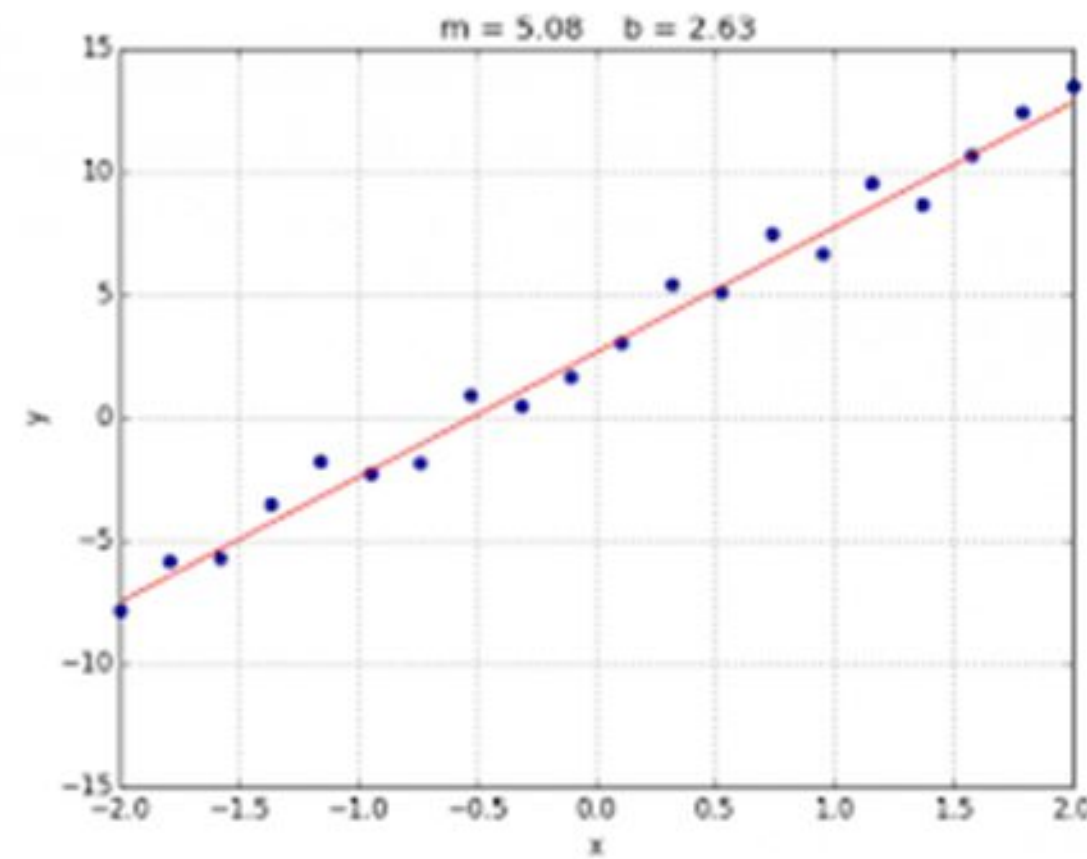
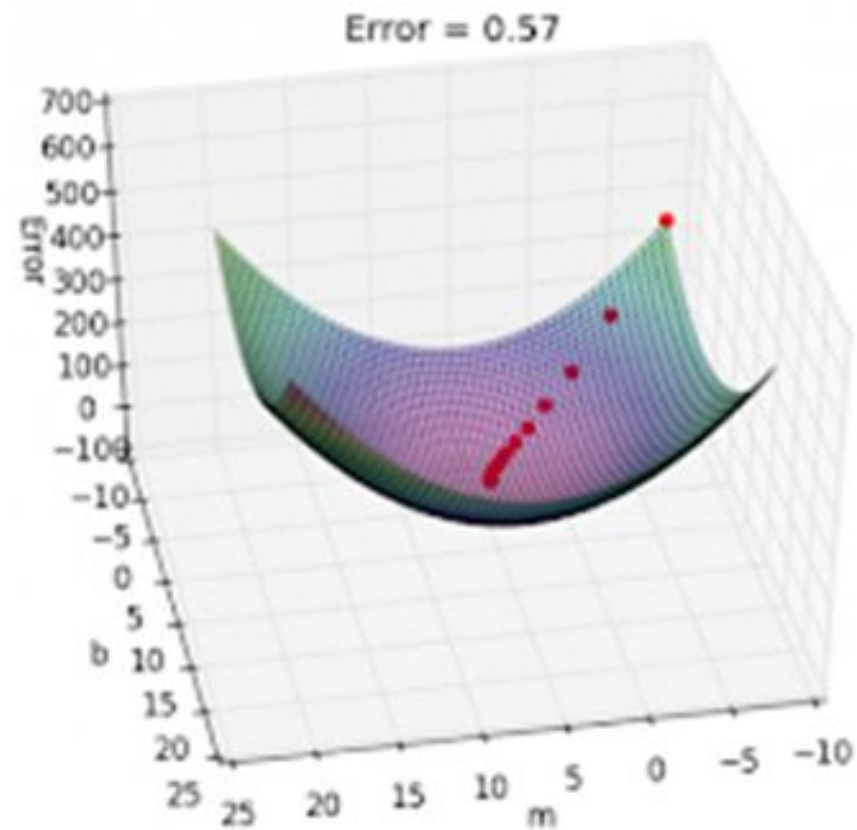
Цель линейной регрессии - найти линию, которая лучше всего подходит к этим точкам.





# Линейная регрессия

Цель линейной регрессии - найти линию, которая лучше всего подходит к этим точкам.



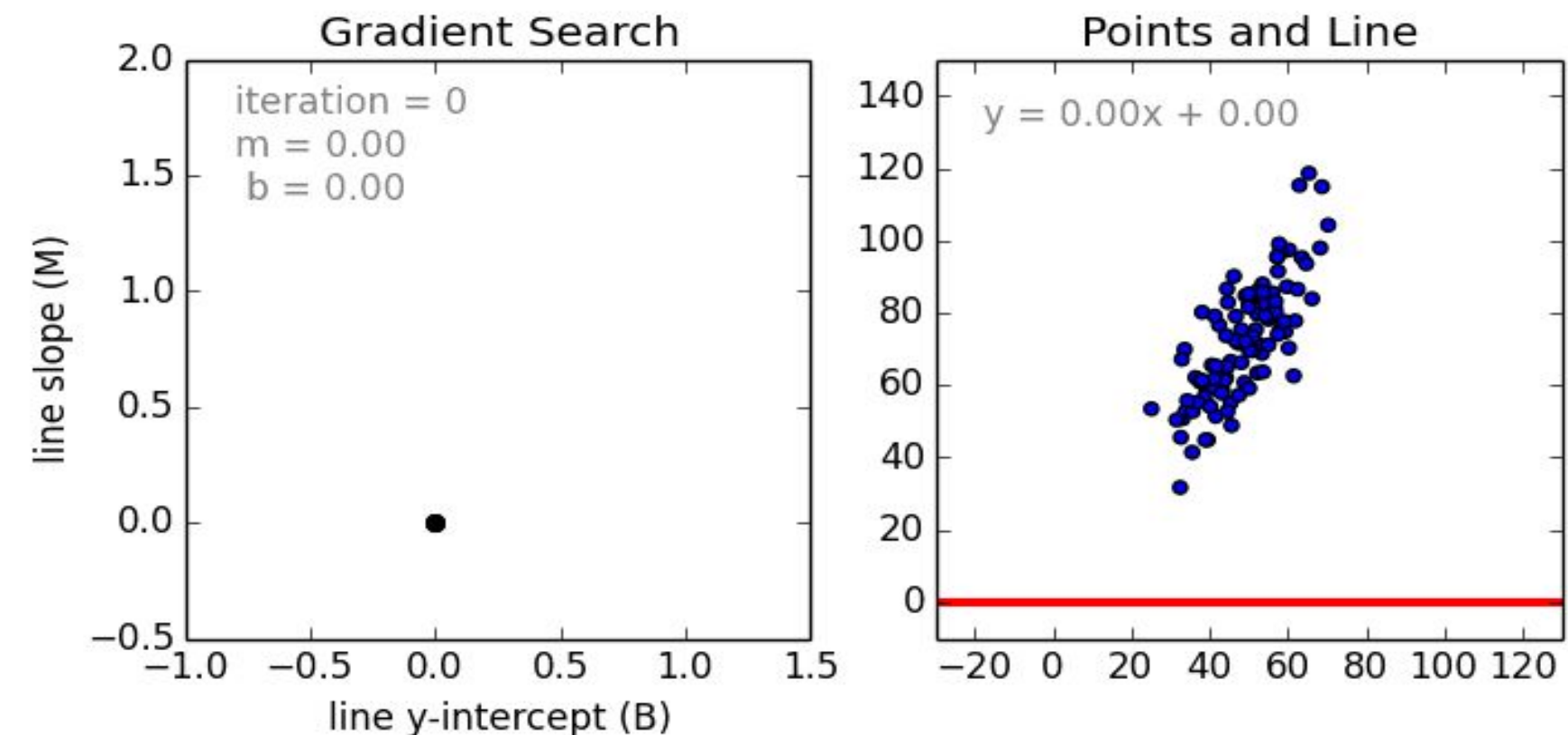




# Линейная регрессия

**Парная (простая) линейная регрессия** — это модель, позволяющая моделировать взаимосвязь между значениями одной входной независимой и одной выходной зависимой переменными с помощью линейной модели, например, прямой.

Более распространенной моделью является **множественная линейная регрессия**, которая предполагает установление линейной зависимости между множеством входных независимых и одной выходной зависимой переменных.





# Линейная регрессия

Несколько важных пунктов о линейной регрессии:

- ❑ Она легко моделируется и является особенно полезной при создании не очень сложной зависимости, а также при небольшом количестве данных.
- ❑ Обозначения интуитивно-понятны.
- ❑ Чувствительна к выбросам.



## Полиномиальная регрессия

Для создания такой модели, которая подойдет для нелинейно разделяемых данных, можно использовать полиномиальную регрессию.

В данном методе проводится кривая линия, зависящая от точек плоскости. В полиномиальной регрессии степень некоторых независимых переменных превышает 1.

Например, получится что-то подобное:

$$Y = a_1 * X_1 + (a_2)^2 * X_2 + (a_3)^4 * X_3 \dots a_n * X_n + b$$



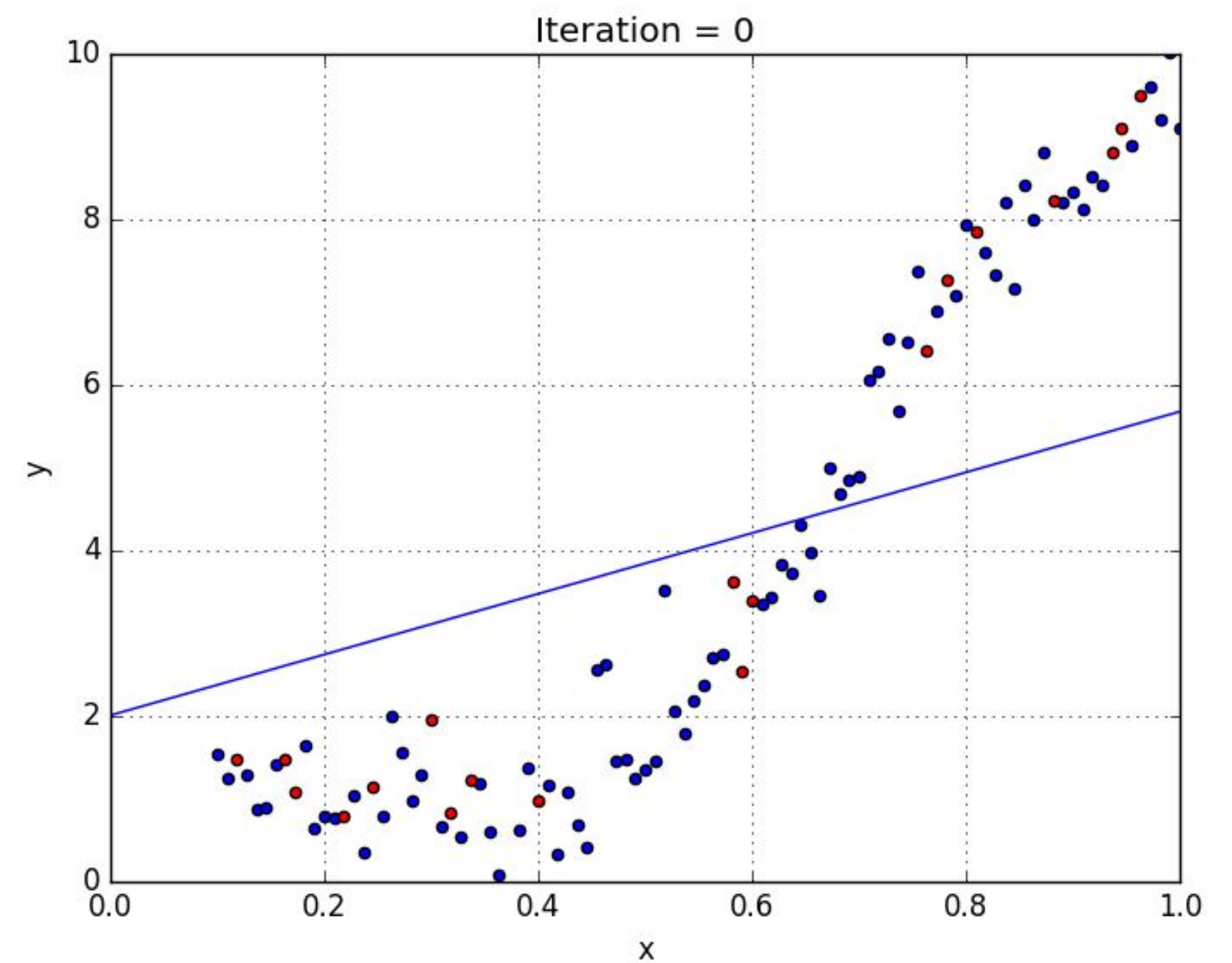
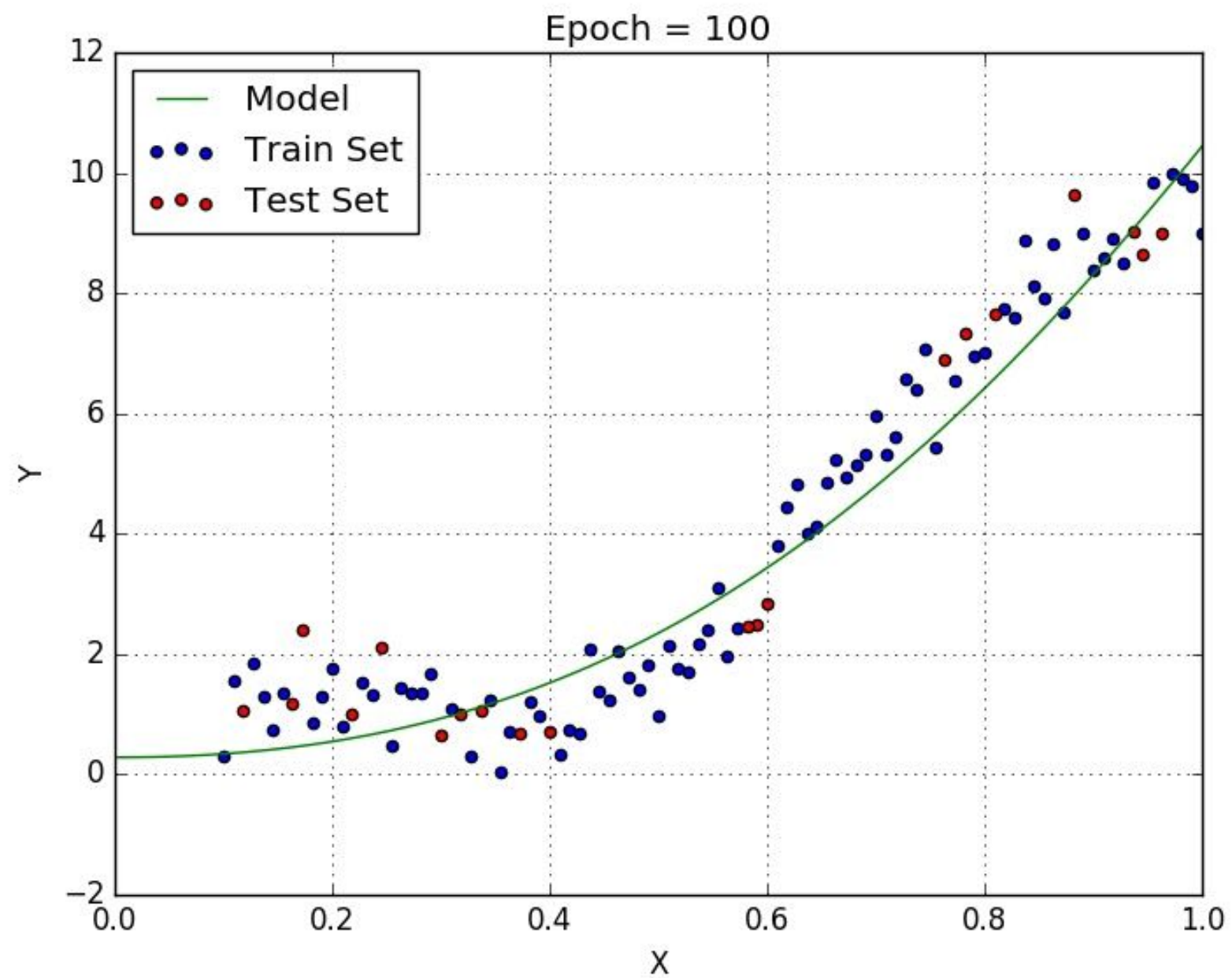
## Полиномиальная регрессия

Несколько важных пунктов о полиномиальной регрессии:

- Моделирует нелинейно разделенные данные (чего не может линейная регрессия). Она более гибкая и может моделировать сложные взаимосвязи.
- Полный контроль над моделированием переменных объекта (выбор степени).
- Необходимо внимательно создавать модель.
- Необходимо обладать некоторыми знаниями о данных, для выбора наиболее подходящей степени
- При неправильном выборе степени, данная модель может быть перенасыщена.



# Сравнение Полиномиальная регрессия и Линейная регрессия



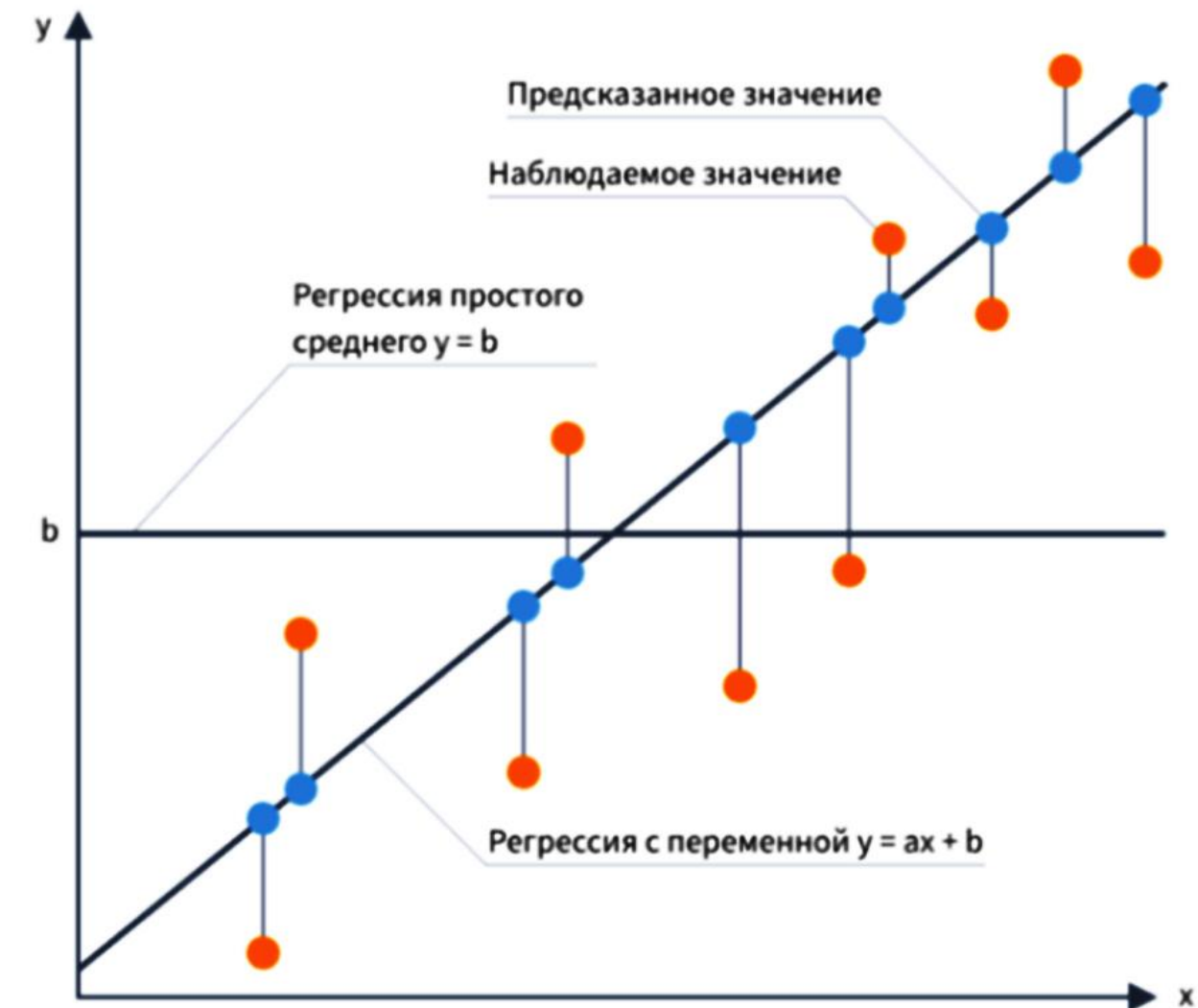




# Метрики регрессии

Задача регрессии – предсказания вещественного значения.

Наклонная линия представляет собой линию регрессии с переменной, имеющей точки, соответствующие прогнозируемому значению выходной переменной (кружки синего цвета). Оранжевые кружки представляют фактические (наблюдаемые) значения  $y$





## MSE - Среднеквадратичная ошибка (Mean Squared Error)

Используется, когда вы хотите подчеркнуть большие ошибки и хотите выбрать модель с точно меньшим количеством больших ошибок. Большие значения ошибок подчеркиваются из-за квадратичной зависимости.

$$MSE = \frac{1}{n} \sum_{i=0}^n (y_i - \tilde{y}_i)^2$$

Недостатки использования MSE является то, что если один или несколько случаев ошибок (возможно, включая выбросы) приводят к большим ошибкам, то их возведение в квадрат приводит к ошибочному выводу о низкой эффективности модели в целом. С другой стороны, если модель дает небольшие ошибки во многих случаях, это может иметь обратный эффект, т.е. недооценить слабость модели.



## RMSE - Корень из среднеквадратичной ошибки (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^n (y_i - \tilde{y}_i)^2}$$

Влияние каждой ошибки на RMSE пропорционально размеру квадрата ошибки. Поэтому большие ошибки оказывают непропорционально большое влияние на RMSE. В результате RMSE можно считать чувствительным к выбросам.



## MAE - Средняя абсолютная ошибка (Mean Absolute Error)

$$MAE = \frac{1}{n} \sum_{i=0}^n |y_i - \tilde{y}_i|$$

То есть, MAE рассчитывается как среднее абсолютных значений разницы между наблюдаемыми и предсказанными значениями; в отличие от MSE и RMSE, это линейная оценка, поэтому все ошибки в среднем имеют одинаковый вес.



## MAPE - Средняя абсолютная процентная ошибка (Mean Absolute Percentage Error)

$$MAPE = \frac{100}{n} \sum_{i=0}^n \frac{|y_i - \tilde{y}_i|}{\tilde{y}}$$

Эта ошибка не имеет размерности и очень легко интерпретируется. Она может быть выражена в виде дроби или процента.





## R-квадрат

Коэффициента детерминации, который указывает на долю дисперсии зависимой переменной, объясненную регрессионной моделью. Наиболее распространенные формулы, используемые для расчета коэффициента детерминации, следующие:

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

- Основное преимущество коэффициента детерминации перед мерами, основанными на ошибках, заключается в том, что он инвариантен по отношению к масштабу данных
- Кроме того, он всегда находится в диапазоне от  $-\infty$  до 1. Значения, близкие к 1, указывают на то, что модель хорошо согласуется с данными.



## Скорректированный R-квадрат

С коэффициентом детерминации связаны две проблемы.

Первая заключается в том, что не все переменные, добавленные в модель, делают точность модели значимой и всегда увеличивают ее сложность.

Вторая проблема заключается в том, что коэффициент детерминации нельзя использовать для сравнения моделей с разным количеством переменных.

Для преодоления этих проблем используются альтернативные показатели, одним из которых является скорректированный коэффициент детерминации .

$$R^2_{adj} = 1 - \frac{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n-k}}{\frac{\sum_{i=1}^n (\bar{y} - y_i)^2}{n-k}}$$



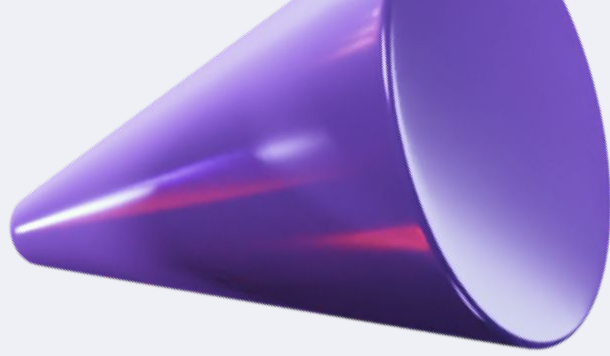
## Вывод:

- ❑ Обучение с учителем является одним из наиболее распространенных методов машинного обучения. В процессе обучения модели, аналитик использует данные, которые уже имеют определенную разметку для того, чтобы научиться прогнозировать новые значения.
- ❑ Одной из задач обучения с учителем является задача регрессии, которая заключается в прогнозировании непрерывной переменной на основе заданных входных данных.
- ❑ Линейная регрессия является одной из наиболее популярных и простых моделей регрессии, которую мы рассмотрим в данной лекции.
- ❑ Эта тема является важной для обучения, поскольку многие реальные задачи данных связаны с прогнозированием непрерывных переменных, таких как доход или цена на недвижимость.
- ❑ Поэтому, понимание линейной регрессии является необходимым для успеха в области машинного обучения и анализе данных.



## Итоги урока:

- ❑ Линейная регрессия - это статистический метод, используемый для определения отношения между независимой и зависимой переменными.
- ❑ Линейная регрессия может быть однофакторной (когда есть только одна независимая переменная) и множественной (когда несколько переменных влияют на зависимую переменную)..
- ❑ Цель линейной регрессии - найти уравнение линии, которая наилучшим образом соответствует наблюдаемым данным.
- ❑ Линейная регрессия также может быть расширена до нелинейных моделей, применением полиномиальной регрессии или других нелинейных методов
- ❑ Линейная регрессия - это мощный и универсальный инструмент для анализа данных, который может быть применен в широком спектре задач, требующих анализа зависимостей между переменными.



**Спасибо за внимание**

