Построение линии регрессии

Что такое регрессия?

- Регрессия — это метод моделирования зависимостей между переменными.
- Используется для предсказания числового значения.
- Например, прогнозирование цен, температуры, спроса, стоимости жилья.

Виды регрессий

1. Линейная
2. Логистическая
3. Полиномиальная
4. Гребневая (ридж) регрессия
5. Регрессия по методу «лассо»
6. Регрессия «эластичная сеть»

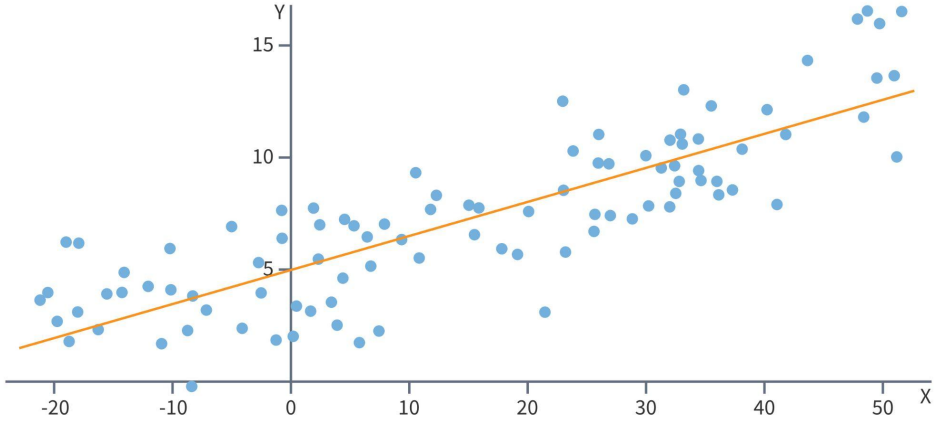
В рамках курса мы рассматриваем только первые 3 вида, так что поговорим на счет них подробнее.

Виды регрессий. Линейная регрессия.

- Линейная регрессия – это метод, используемый для моделирования зависимости одной переменной (зависимой) от другой или нескольких других (независимых) с помощью линейной функции.

$$Y = b_0 + b_1X_1 + b_2X_2 \ldots b_nX_n$$

- Y — зависимая переменная
- b_0 — точка пересечения с осью y
- b_1X_1 — коэффициент наклона 1, независимая переменная 1
- b_2X_2 — коэффициент наклона 2, независимая переменная 2
- b_nX_n — коэффициент наклона n, независимая переменная n



Виды регрессий. Логистическая регрессия.

- Логистическая регрессия – это метод, используемый для моделирования вероятности принадлежности объекта к одному из двух классов. В отличие от линейной регрессии, она предсказывает вероятность (значение от 0 до 1), а не непрерывные числа.

Sigmoid (aka logistic) function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

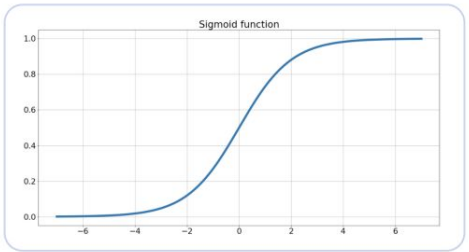
Sigmoid is odd relative to (0, 0.5) point

Symmetric property:

$$1 - \sigma(z) = \sigma(-z)$$

Derivative:

$$\sigma(z)' = \sigma(z)\big(1 - \sigma(z)\big)$$



Виды регрессий. Полиномиальная регрессия

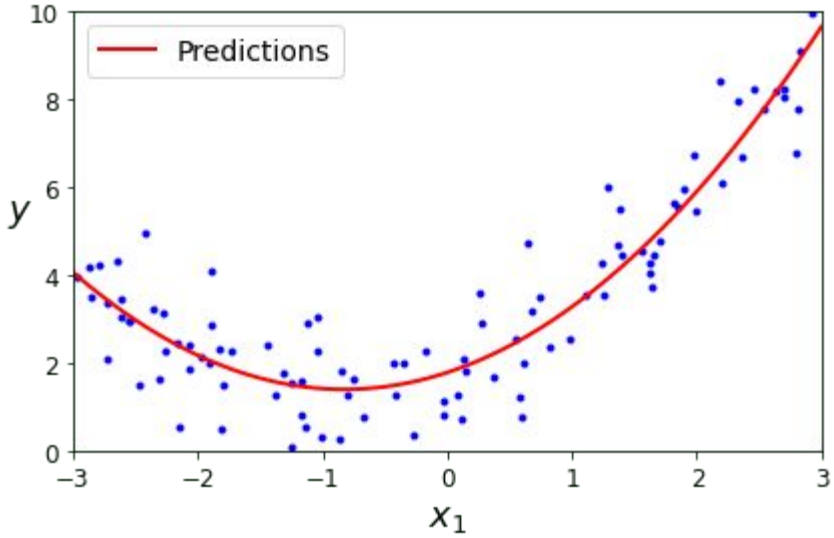
- Полиномиальная регрессия - это обобщение линейной регрессии, где зависимая переменная моделируется с помощью полинома (многочлена) вместо простой прямой линии.

$$y = a + b_1 x + b_2 x^2 + b_3 x^3 + \cdots + b_n x^n + \varepsilon$$

Где:

- y — зависимая переменная (то, что мы предсказываем);
- x — независимая переменная (фактор);
- a — свободный член (intercept);
- $b_1, b_2, \ldots, b_n$ — коэффициенты полинома;
- $x^2, x^3, \ldots, x^n$ — дополнительные нелинейные признаки;
- ε — ошибка модели.

Чем выше степень n, тем сложнее кривая.



Ошибки для регрессии

- Mean Absolute Error

  Сумма модулей различий между значениями, поделенное на кол-во элементов

- Mean Square Error

  Сумма квадратов различий между значениями, поделенное на кол-во элементов

- Sum of Squared Errors

  Сумма квадратов различий между значениями

- Mean Absolute Percentage Error

Как найти коэффициенты.

- Аналитически
- Метод наименьших квадратов (МНК)

  Минимизируем сумму квадратов разностей между предсказанными значениями и реальными данными.

- Градиентный спуск (и его оптимизации)

  Спускаемся по функции ошибки к минимуму.

Как оценить качество. R^2

Коэффициент детерминации R^2 (англ. coefficient of determination) — это метрика, показывающая, насколько хорошо линейная регрессия объясняет данные.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Где:

- $SS_{res} = \sum(y_i - \hat{y}_i)^2$ — сумма квадратов остатков (ошибок модели).
- $SS_{tot} = \sum(y_i - \bar{y})^2$ — общая сумма квадратов отклонений от среднего значения.

Что показывает R^2:

- Если $R^2 = 1$, модель идеально объясняет данные.
- Если $R^2 = 0$, модель не лучше, чем простое среднее $\bar{y}$.
- Если $R^2 < 0$, модель хуже среднего, возможно, данные вообще не связаны линейной зависимостью.

Алгоритм построения линии регрессии

1. Подготавливаем данные
2. Пытаемся выделить, какой вид регрессии к нам подходит
3. Обучаем модель, используя один из видов подсчета коэффициентов.
4. При помощи R^2 оцениваем модель, если надо, то повторяем шаг 2.
5. Используем полученную модель