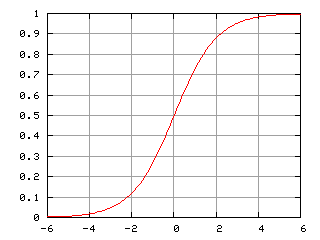
Логистическая регрессия

**Логистическая регрессия** - статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём его сравнения с логистической кривой. Эта регрессия выдаёт ответ в виде вероятности бинарного события (1 или 0).

**Л**огистическая регрессия применяется для прогнозирования вероятности возникновениянекоторого события по значениям множества признаков.

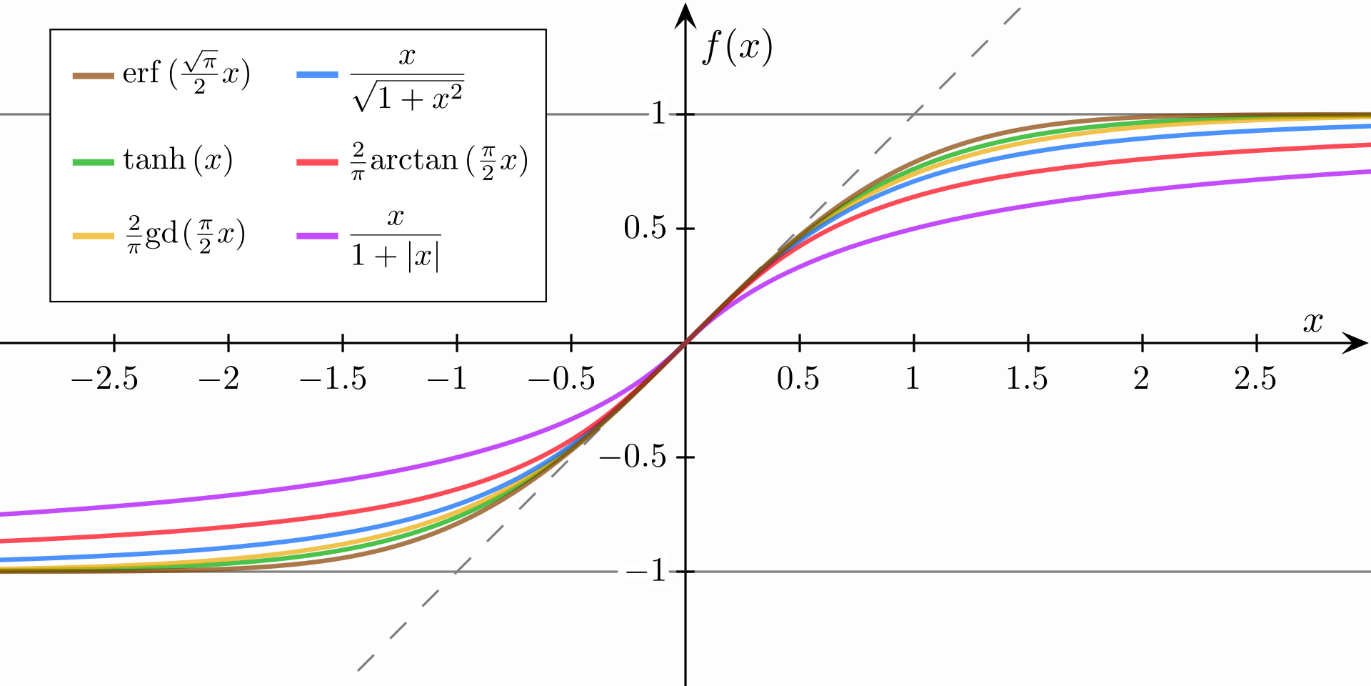
Используется в задачах классификации.

# Сигмо́ида

Сигмо́ида — это гладкая монотонная возрастающая нелинейная функция, имеющая форму буквы «S», которая часто применяется для «сглаживания» значений некоторой величины.

Логистическая сигмоида

* 2 асимтоты y = 0 и y = 1
* удобна для представления функции вероятности события

Так же бывают и другие виды сигмоид: 

# Метод наибольшего правдоподобия (MLE)

Введем функцию , которая может принимать значения 0 (событие не произошло) или 1 (событие произошло) и зависит от множества независимых переменных .

Вероятность наступления события

и векторы-столбцы значений независимых переменных и параметров (коэффициентов регрессии).

так называемая логистическая функция сигмоид

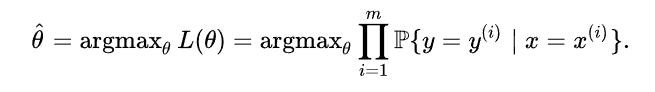
Так как принимает лишь значения 0 и 1, то вероятность принять значение 0 равна:

Обобщенно функцию распределения можно записать так

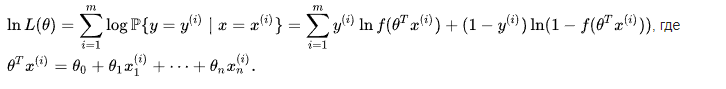
(тут используется хак, если y = 1, остается только правая часть, то есть вероятность наступления события 1, а если y = 0, то левая часть)

Теперь нужно выбрать такие параметры , которые максимизируют нашу функцию правдоподобия (это по факту ).

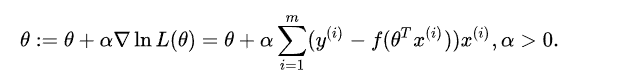
Для этого нам пригодится метод наибольшего правдоподобия.



Максимизация функции правдоподобия эквивалентна максимизации её логарифма, что позволяет заменить произведения на сумму логарифмов.



Мы получили функцию которую нужно максимизировать, для этого хорошо подходит метод градиентного спуска:



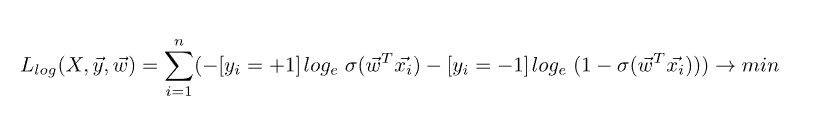
# Логистическая регрессия для меток от (-1), до 1

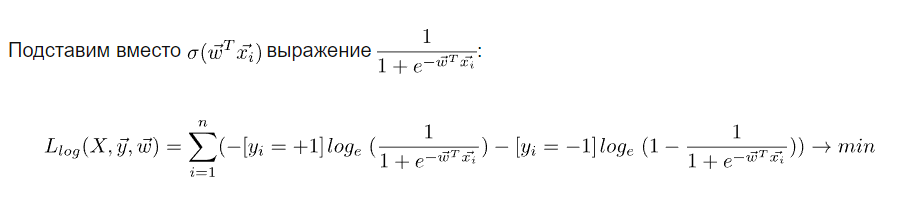
Для такого случая обобщённая формула, практически такая же как и для случая 0;1

То есть вместо степеней появляется “если”

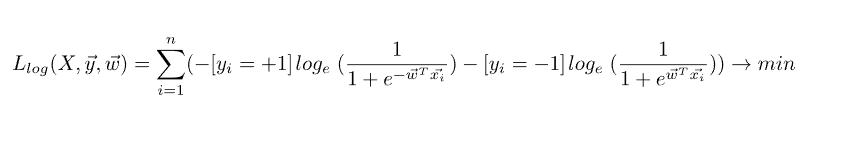
* Если y = -1, то нужно взять левую часть
* Если y = 1, то нужно взять правую часть

Теперь логарифмируем эту функцию правдоподобия:

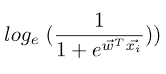




Упростим правое слагаемое под логарифмом, используя простые арифметические приемы и получим:

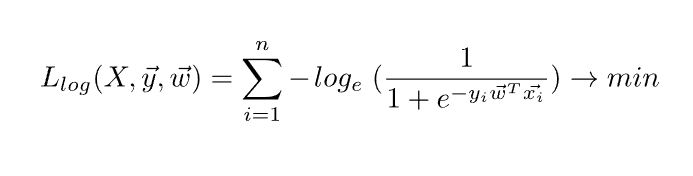


Теперь у нас получается, что:

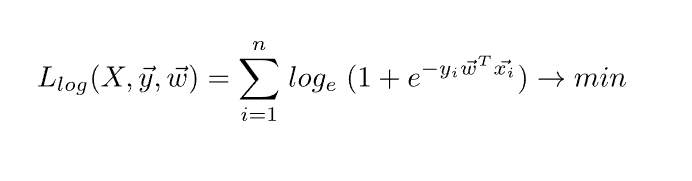
* Если y = -1, то берем слагаемое с + в степени - 
* Если y = 1, то берем слагаемое с - в степени- Изображение выглядит как текст, часы

  Автоматически созданное описание

Это приводит к следующей формуле:



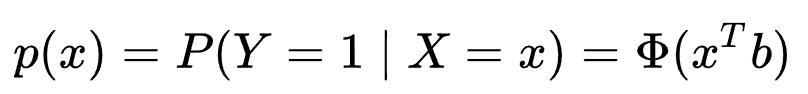
Еще можно в соответствии с правилами логарифмирования, перевернем дробь и вынесем знак "-" (минус) за логарифм, получим:



# Пробит-регрессия

Пробит-модель является частным случаем модели бинарного выбора (как и логистическая регрессия), в которой используется нормальное распределение.

В пробит-модели предполагается, что вероятность того, что определяется нормальным распределением, таким образом пробит-модель имеет вид:



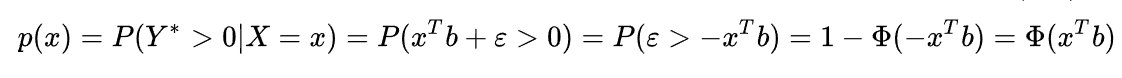
где — интегральная функция распределения (CDF) стандартного нормального распределения, — неизвестные параметры, которые требуется оценить.

Как и в общем случае модели бинарного выбора в основе модели лежит предположение о наличии некоторой скрытой (ненаблюдаемой) переменной в зависимости от значений которой наблюдаемая переменная принимает значение 0 или 1:

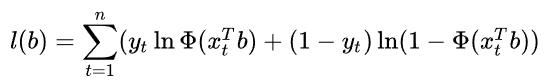
Изображение выглядит как текст

Автоматически созданное описание

Предполагается, что скрытая переменная зависит от факторов X в смысле обычной линейной регрессии где случайная ошибка в данном случае имеет стандартное нормальное распределение Тогда



Далее так же методом максимального правдоподобия подбираем коэффициенты



# Обобщённые линейные модели

Допустим что Y линейно заисит от X,то есть *Y = bTX*б, но:

* Исходные данные могли измерить с ошибкой
* Упускаются еще какие-то факторы X, которые влияют на Y, поэтому нужно учесть случайную ошибку ℇ
* нет никакой гарантии, что Y линейно зависят сразу напрямую от X, поэтому может быть так : *log Y = bTX*

Таким образом мы приходим к обобщенной постановке линейной задачи:

Y\* = bTX\* + ℇ

где X\* = f(X), Y\* = g(Y),

ℇ — случайная величина.

Несмотря на то, что f и g могут быть нелинейными функциями, и Y в результате может весьма нелинейно зависеть от Х, модель все равно остается линейной относительно параметров b. Именно поэтому она и называется линейной моделью.

Заметим важный факт. Поскольку произведение *bTX\** абсолютно детерминировано, то можно сказать, что *Y\** тоже является случайной величиной, которая имеет такую же форму распределения, что и *ℇ*. И тогда можно сделать удобный вывод, переписав модель в виде *E[Y\*] = bTX\**, что означает, что наша линейная модель предсказывает не само значение *Y\**, а его математическое ожидание.

Изображение выглядит как текст

Автоматически созданное описание

* – линейная функция
* – связующая функция

В случае логистической регрессии связующая функция является логистическая функция.

В пробит модели:

Изображение выглядит как текст

Автоматически созданное описание

Зависимая переменная (Y) также может содержать как исходные данные — и это будет простой моделью — так и их преобразование. Причем когда преобразованная зависимая переменная принадлежит к экспоненциальному семейству распределений, то речь уже идет о так называемой обобщенной линейной модели (GLM, generalized linear model), к которым в частности относятся нормальная, логистическая, Пуассоновская, экспоненциальная, биномиальная и многие другие модели. Обобщенные модели очень важны и удобны в использовании, поскольку для них доказаны и параметры сходимости, и качества получаемых оценок, и влияние функционалов разных видов. В идеале старайтесь свести вашу задачу к какой-нибудь GLM-модели.