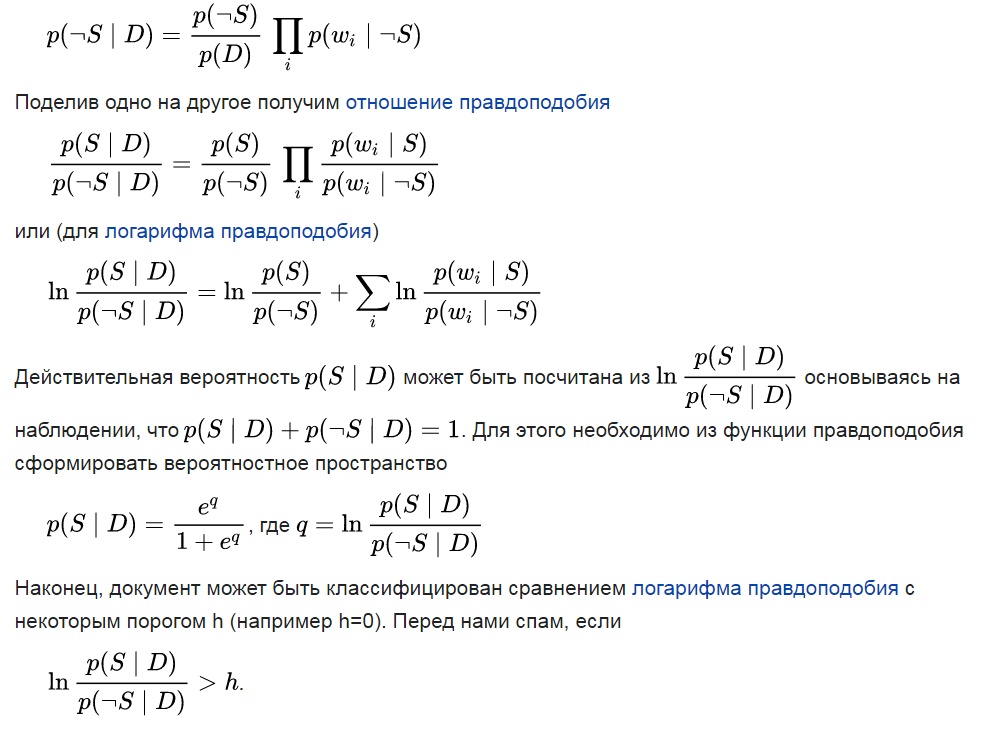
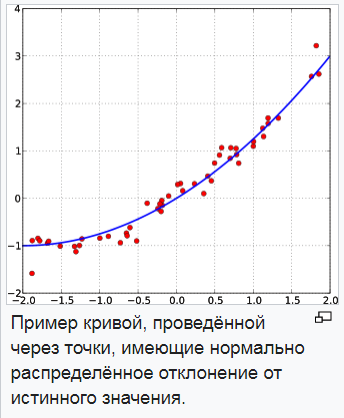


<https://ranalytics.github.io/data-mining/072-NBC.html>

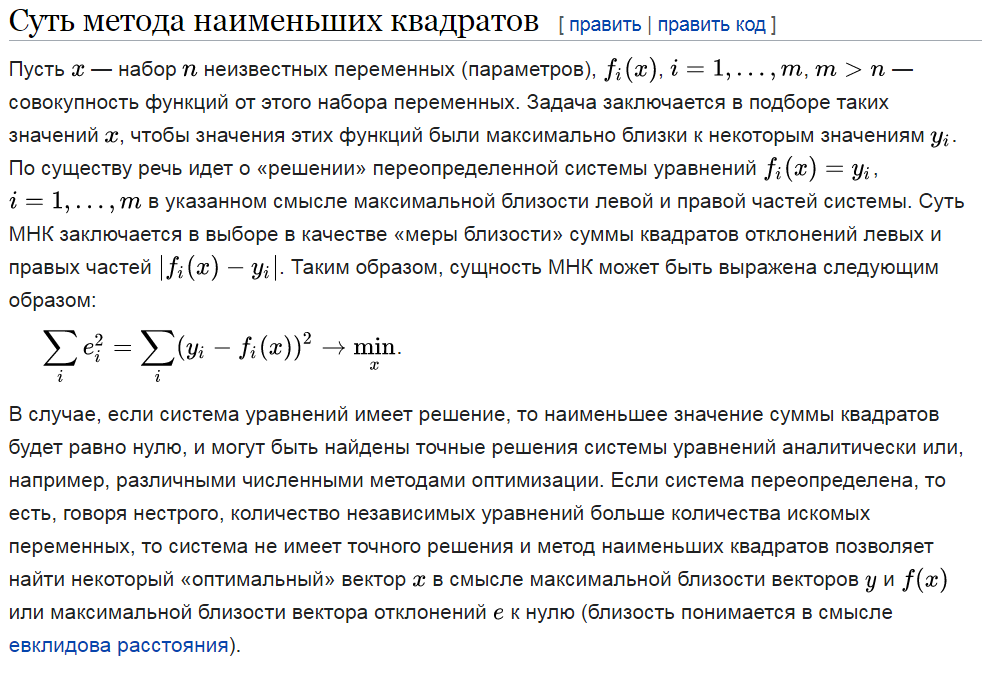


**Метод наименьших квадратов**

****

Всем, кто хоть немного изучал статистику, знакомо понятие линейной регрессии. К вариантам её реализации относятся и наименьшие квадраты. Обычно с помощью линейной регрессии решают задачи по подгонке прямой, которая проходит через множество точек. Вот как это делается с помощью метода наименьших квадратов: провести прямую, измерить расстояние от неё до каждой из точек (точки и линию соединяют вертикальными отрезками), получившуюся сумму перенести наверх. В результате та кривая, в которой сумма расстояний будет наименьшей, и есть искомая (эта линия пройдёт через точки с нормально распределённым отклонением от истинного значения).

Линейная функция обычно используется при подборе данных для машинного обучения, а метод наименьших квадратов – для сведения к минимуму погрешностей путем создания метрики ошибок.



|  |
| --- |
| Метод наименьших квадратов и его оценки Создание метода наименьших квадратов восходит к трудам Карла Фридриха Гаусса в конце XVI11 и начале XIX века в области исследований по астрономии. Математический метод был открыт в связи с необходимостью обработки неравноценных наблюдаемых данных.  В дальнейшем применил способ наименьших квадратов и развил теорию ошибок Пьер Симон Лаплас. Также существенный вклад в развитие метода внес Адриен Мари Лежандр.  Этот метод приобрел самую широкую известность благодаря фундаментальным трудам многих статистиков и математиков и его применению в экономикостатистических расчетах.  Рассмотрим метод наименьших квадратов на простом примере зависимости между двумя переменными х и *у,* причем *у* зависит от х. Если установлено, что связь между ними криволинейная и описывается параболой, т.е. полиномом второй степени, с параметрами:  РЕКЛАМА  https://studme.org/htm/img/29/2230/64.png  то задача сводится к отысканию неизвестных трех параметров.  При числе наблюдений (количестве уровней в рядах) *п,* значения величин х и *у* представлены двумя рядами данных: https://studme.org/htm/img/29/2230/65.png  Если бы все значения, полученные по данным наблюдения, лежали строго на линии, описываемой уравнением параболы, то для каждой точки было бы справедливо следующее равенство:  https://studme.org/htm/img/29/2230/66.png  Однако в действительности  https://studme.org/htm/img/29/2230/67.png  которое существует вследствие ошибок измерения и случайных неучтенных факторов. Необходимо найти такие коэффициенты регрессии, чтобы ошибка была минимальной. Можно минимизировать сумму абсолютных (по модулю) отклонений или сумму кубических отклонений или наибольшую абсолютную ошибку. Однако оптимальным подходом является минимизация квадрата отклонений  https://studme.org/htm/img/29/2230/68.png  Минимизация квадратов отклонений обладает тем свойством, что число нормальных уравнений равно числу неизвестных параметров. Минимизация суммы  https://studme.org/htm/img/29/2230/69.png  дает три уравнения для каждого из трех параметров. Для нахождения значений неизвестных параметров необходимо приравнять нулю частные производные указанной суммы по этим параметрам:  https://studme.org/htm/img/29/2230/70.png  Проведение простейших преобразований приводит к системе нормальных уравнений:  https://studme.org/htm/img/29/2230/71.png  Решение системы линейных относительно неизвестных параметров уравнений любым из способов дает значения Обычно полиномы выше третьей степени практически нс используются, то система нормальных уравнений такого полинома будет состоять соответственно из четырех уравнений.  МНК даже при сравнительно небольшом числе наблюдений приводит к получению достаточных оценок. Оценки могут быть точечными и интервальными. Точечные оценки обладают свойствами несмещенности, эффективности, состоятельности, описанными в предыдущем параграфе.  Однако любая оценка истинного значения параметра по выборочным данным может быть произведена только с определенной степенью достоверности. Степень этой достоверности определяется путем построения доверительных интервалов.  Метод наименьших квадратов может быть использован и в случаях, когда имеются данные косвенных наблюдений, являющиеся функциями многих неизвестных. МНК является основой регрессионного анализа, используемого при выполнении предпосылок, рассмотренных выше. Также условием его применения является линейность уравнений регрессии относительно параметров. Исходя из классификации видов регрессии МНК применим для линейных и нелинейных регрессий первого класса. |