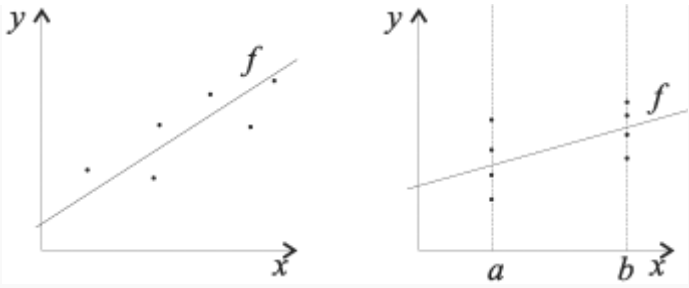
Регрессионный анализ — метод моделирования измеряемых данных и исследования их свойств. Данные состоят из пар значений зависимой переменной (переменной отклика) и независимой переменной (объясняющей переменной). Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной. Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные. Критерием качества приближения (целевой функцией) обычно является среднеквадратичная ошибка: сумма квадратов разности значений модели и зависимой переменной для всех значений независимой переменной в качестве аргумента. Регрессионный анализ — раздел математической статистики и машинного обучения. Предполагается, что зависимая переменная есть сумма значений некоторой модели и случайной величины. Относительно характера распределения этой величины делаются предположения, называемые гипотезой порождения данных. Для подтверждения или опровержения этой гипотезы выполняются статистические тесты, называемые анализом остатков. При этом предполагается, что независимая переменная не содержит ошибок. Регрессионный анализ используется для прогноза, анализа временных рядов, тестирования гипотез и выявления скрытых взаимосвязей в данных.

Определение регрессионного анализа



Выборка может быть не функцией, а отношением. Например, данные для построения регрессии могут быть такими: $\{(0,0),(0,1),(0,2),(1,1),(1,2),(1,3)\}$. В такой выборке одному значению переменной x соответствует несколько значений переменной y.

Регрессия — зависимость математического ожидания (например, среднего значения) случайной величины от одной или нескольких других случайных величин (свободных переменных), то есть $E(y|x) = f(x)$. Регрессионным анализом называется поиск такой функции f, которая описывает эту зависимость. Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих.

$$y = f(x) + \nu,$$

где f — функция регрессионной зависимости, а ν — аддитивная случайная величина с нулевым матожиданием. Предположение о характере распределения этой величины называется гипотезой порождения данных. Обычно предполагается, что величина ν имеет гауссово распределение с нулевым средним и дисперсией σ_ν^2.

Задача нахождения регрессионной модели нескольких свободных переменных ставится следующим образом. Задана выборка — множество $\{x_1,\dots,x_N | x \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1,\dots,y_N | y \in \mathbb{R}\}$соответствующих им значений зависимой переменной. Эти множества обозначаются как D, множество исходных данных $\{(x,y)_i\}$. Задана регрессионная модель — параметрическое семейство функций $f(w,x)$ зависящая от параметров $w \in \mathbb{R}$ и свободных переменных x. Требуется найти наиболее вероятные параметры $\overline{w}$:

$$\overline{w} = \underset{w \in \mathbb{R}^W}{\arg\max}\, p(y|x,w,f) = p(D|w,f).$$

Функция вероятности p зависит от гипотезы порождения данных и задается Байесовским выводом или методом наибольшего правдоподобия.

Линейная регрессия

Линейная регрессия предполагает, что функция f зависит от параметров w линейно. При этом линейная зависимость от свободной переменной x необязательна,

$$y = f(w,x) + \nu = \sum_{j=1}^{N} w_j g_j(x) + \nu.$$

В случае, когда функция $g \equiv \mathrm{id}$ линейная регрессия имеет вид

$$y = \sum_{j=1}^{N} w_j x_j + \nu = \langle w,x \rangle + \nu,$$

здесь x_j — компоненты вектора x.

Значения параметров в случае линейной регрессии находят с помощью метода наименьших квадратов. Использование этого метода обосновано предположением о гауссовском распределении случайной переменной.

Разности $y_i - f(x_i)$ между фактическими значениями зависимой переменной и восстановленными называются **регрессионными остатками** (residuals). В литературе используются также синонимы: *невязки* и *ошибки*. Одной из

важных оценок критерия качества полученной зависимости является сумма квадратов остатков:

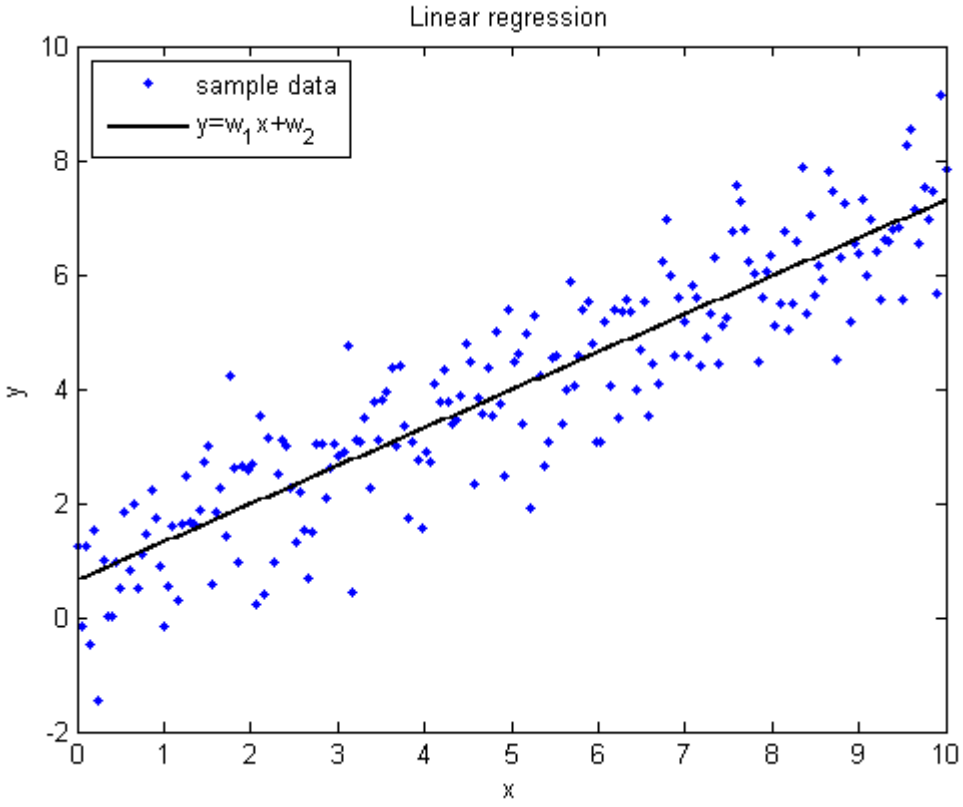
$$SSE = \left| f(x_i) - y_i \right|_2 = \sum_{i=1}^{N} (y_i - f(w, x_i))^2.$$

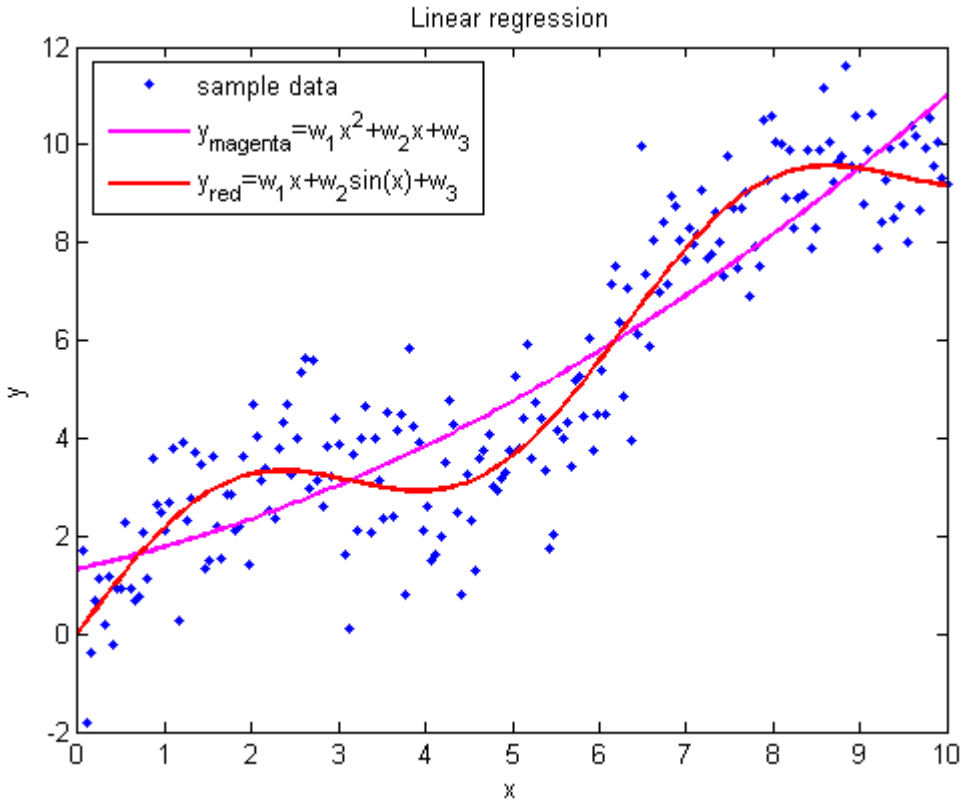
Здесь SSE — Sum of Squared Errors.

Дисперсия остатков вычисляется по формуле

$$\bar{\sigma}_\nu^2 = \frac{SSE}{N-2} = MSE.$$

Здесь MSE — Mean Square Error, среднеквадратичная ошибка.





На графиках представлены выборки, обозначенные синими точками, и регрессионные зависимости, обозначенные сплошными линиями. По оси абсцисс отложена свободная переменная, а по оси ординат — зависимая. Все три зависимости линейны относительно параметров.

Нелинейная регрессия

Нелинейные регрессионные модели — модели вида

$$y = f(w,x)+\nu,$$

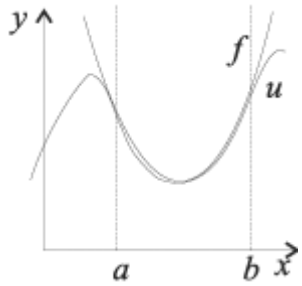
которые не могут быть представлены в виде скалярного произведения

$$f(w,x) = (w, g(x)) = \sum_{i=1}^{n} w_i g_i(x),$$

где $w = [w_1, \ldots, w_n]$ — параметры регрессионной модели, x — свободная переменная из пространства $\mathbb{R}^n$, y — зависимая переменная, ν — случайная величина и $g = [g_1, \ldots, g_n]$ — функция из некоторого заданного множества.

Значения параметров в случае нелинейной регрессии находят с помощью одного из методов градиентного спуска, например алгоритма Левенберга-Марквардта.

Термин "регрессия" был введён Фрэнсисом Гальтоном в конце 19-го века. Гальтон обнаружил, что дети родителей с высоким или низким ростом обычно не наследуют выдающийся рост и назвал этот феномен "регрессия к посредственности". Сначала этот термин использовался исключительно в биологическом смысле. После работ Карла Пирсона этот термин стали использовать и в статистике.

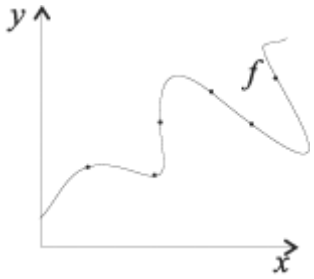


Аппроксимация функций: непрерывная функция f приближает непрерывную или дискретную функцию u

В статистической литературе различают регрессию с участием одной свободной переменной и с несколькими свободными переменными — *одномерную* и *многомерную* регрессию. Предполагается, что мы используем несколько свободных переменных, то есть, свободная переменная — вектор $x \in \Re^N$. В частных случаях, когда свободная переменная является скаляром, она будет обозначаться x.
Различают *линейную* и *нелинейную*регрессию. Если регрессионную модель не является линейной комбинацией функций от параметров, то говорят о нелинейной регрессии. При этом модель может быть произвольной суперпозицией функций g из некоторого набора. Нелинейными моделями являются, экспоненциальные, тригонометрические и другие (например, радиальные базисные функции или персептрон Розенблатта), полагающие зависимость между параметрами и зависимой переменной нелинейной.

Различают *параметрическую* и *непараметрическую* регрессию. Строгую границу между этими двумя типами регрессий провести сложно. Сейчас не существует общепринятого критерия отличия одного типа моделей от другого. Например, считается, что линейные модели являются параметрическими, а модели, включающие усреднение зависимой переменной по пространству свободной переменной —непараметрическими. Пример параметрической регресионной модели: линейный предиктор, многослойный персептрон. Примеры смешанной регрессионной модели: функции радиального базиса. Непараметрическая модель — скользящее усреднение в окне некоторой ширины. В целом, непараметрическая

регрессия отличается от параметрической тем, что зависимая переменная зависит не от одного значения свободной переменной, а от некоторой заданной окрестности этого значения.



Интерполяция: функция f задана значениями узловых точек

Есть различие между терминами: "приближение функций", "аппроксимация", "интерполяция", и "регрессия". Оно заключается в следующем.

Приближение функций. Дана функция u дискретного или непрерывного аргумента. Требуется найти функцию f из некоторого параметрическую семейства, например, среди алгебраических полиномов заданной степени. Параметры функции f должны доставлять минимум некоторому функционалу, например,

$$\rho(u,f)=\left(\frac{1}{b-a}\int_a^b |f(x)-g(x)|^2 dx\right)^{\frac{1}{2}}.$$

Термин *аппроксимация* — синоним термина "приближение функций". Чаще используется тогда, когда речь идет о заданной функции, как о функции дискретного аргумента. Здесь также требуется отыскать такую функцию f, которая проходит наиболее близко ко всем точкам заданной функции. При этом вводится понятие *невязки* — расстояния между точками непрерывной функции f и соответствующими точками функции u дискретного аргумента.

Интерполяция функций — частный случай задачи приближения, когда требуется, чтобы в определенных точках, называемых *узлами интерполяции* совпадали значения функции u и приближающей ее функции f. В более общем случае накладываются ограничения на значения некоторых производных f производных. То есть, дана функция u дискретного аргумента. Требуется отыскать такую функцию f, которая проходит через все точки u. При этом метрика обычно не используется, однако часто вводится понятие "гладкости" искомой функции.

Регрессия и классификация тесно связаны друг с другом. Термин *алгоритм* в классификации мог бы стать синонимом термина *модель* в регрессии, если бы алгоритм не оперировал с дискретным множеством ответов-классов, а модель — с непрерывно-определенной свободной переменной.

Метод наименьших квадратов — метод нахождения оптимальных параметров линейной регрессии, таких, что сумма квадратов ошибок (регрессионных остатков) минимальна. Метод заключается в минимизации евклидова расстояния $|Aw-y|$ между двумя векторами — вектором восстановленных значений зависимой переменной и вектором фактических значений зависимой переменной.

Постановка задачи

Задача метода наименьших квадратов состоит в выборе вектора w, минимизирующего ошибку $S=|Aw-y|^2$. Эта ошибка есть расстояние от вектора y до вектора Aw. Вектор Aw лежит в простанстве столбцов матрицы A, так как Aw есть линейная комбинация столбцов этой матрицы с коэффициентами $w_1,\dots,w_N$. Отыскание решения wпо методу наименьших квадратов эквивалентно задаче отыскания такой точки $p=Aw$, которая лежит ближе всего к y и находится при этом в пространстве столбцов матрицы A. Таким образом, вектор p должен быть проекцией y на пространство столбцов и вектор невязки $Aw-y$ должен быть ортогонален этому пространству. Ортогональность состоит в том, что каждый вектор в пространстве столбцов есть линейная комбинация столбцов с некоторыми коэффициентами $v_1,\dots,v_N$, то есть это вектор Av. Для всех v в пространстве Av, эти векторы должны быть перпендикулярны невязке $Aw-y$:

$$(Av)^T(Aw-y)=v^T(A^TAw-A^Ty)=0.$$

Так как это равенство должно быть справедливо для произвольного вектора v, то

$$A^TAw-A^Ty=0.$$

Решение по методу наименьших квадратов несовместной системы $Aw=y$, состоящей из M уравнений с N неизвестными, есть уравнение

$$A^TAw=A^Ty,$$

которое называется *нормальным уравнением*. Если столбцы матрицы A линейно независимы, то матрица A^TA обратима и единственное решение

$$w = (A^T A)^{-1} A^T y.$$

Проекция вектора y на пространство столбцов матрицы имеет вид

$$p = Aw = A(A^T A)^{-1} A^T y = Py.$$

Матрица $P = A(A^T A)^{-1} A^T$ называется матрицей проектирования вектора y на пространство столбцов матрицы A. Эта матрица имеет два основных свойства: она идемпотентна, $P^2 = P$, и симметрична, $P^T = P$. Обратное также верно: матрица, обладающая этими двумя свойствами есть матрица проектирования на свое пространство столбцов.

Пример построения линейной регрессии

Задана выборка — таблица

$$D = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \dots & \dots \\ x_M & y_M \end{pmatrix}.$$

Задана регрессионная модель — квадратичный полином

$$f = w_3 x^2 + w_2 x + w_1 = \sum_{j=1}^{3} w_j x^{j-1}.$$

Назначенная модель является линейной. Для нахождения оптимального значения вектора параметров $w = \langle w_1, \dots, w_3 \rangle^T$ выполняется следующая подстановка:

$x_i^0 \to a_{i1}, x_i^1 \to a_{i2}, x_i^2 \to a_{i3}$

Тогда матрица A значений подстановок свободной переменной x_i будет иметь вид

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \dots & \dots & \dots \\ a_{M1} & a_{M2} & a_{M3} \end{pmatrix}.$$

Задан критерий качества модели: функция ошибки

$$S = \sum_{i=1}^{M} (f(w, x_i) - y_i)^2 = |Aw - y|^2 \longrightarrow \min.$$

Здесь вектор $y = \langle y_1, \dots, y_M \rangle$. Требуется найти такие параметры w, которые бы доставляли минимум этому функционалу,

$$w = \arg\min_{w \in \Re^3}(S).$$

Требуется найти такие параметры w, которые доставляют минимум S — норме вектора невязок $Aw-y$.

$$S = |Aw-y|^2 = (Aw-y)^T(Aw-y) =$$
$$= y^T y - y^T Aw - w^T A^T y + w^T A^T Aw =$$
$$= y^T y - 2y^T Aw + w^T A^T Aw.$$

Для того, чтобы найти минимум функции невязки, требуется приравнять ее производные к нулю. Производные данной функции по w составляют

$$\frac{\partial S}{\partial w} = -2A^T y + 2A^T Aw = 0.$$

Это выражение совпадает с нормальным уравнением. Решение этой задачи должно удовлетворять системе линейных уравнений

$$A^T Aw = A^T y,$$

то есть,

$$w = (A^T A)^{-1}(A^T y).$$

После получения весов можно построить график найденной функции.

При обращении матрицы $(A^T A)^{-1}$ предполагается, что эта матрица невырождена и не плохо обусловлена. О том, как работать с плохо обусловленными матрицами см. в статье Сингулярное разложение.

Источник: http://www.machinelearning.ru