Лекция 13

1. Решение задачи регрессии в рамках статистического и детерминистского подходов

Задача регрессии является одной из классических задач анализа данных и относится к классу предсказательных (predictive) задач. Как уже упоминалось, предсказательные задачи осуществляются в два этапа. Первый этап направлен на построение модели анализируемого объекта (процесса) путем обработки данных – получаемых в ходе экспериментов наблюдений, в той или иной степени, его характеризующих. На втором этапе модель используется для предсказания результатов по отношению к новым наборам данных.

Специфика задачи регрессии состоит в том, что модель объекта строится в виде функциональной зависимости набора непрерывных выходных переменных от заданного набора входных, а второй этап направлен на определение значений выходных переменных при появлении новых значений входных. Общность задач оценивания и регрессии определяется тем, что построение регрессионной модели фактически сводится к задаче оценивания неизвестных параметров модели объекта с использованием схожих методов и алгоритмов.

Итак, задача регрессии состоит в установлении функциональной зависимости между зависимыми и независимыми показателями и переменными в понятной форме. Обычно вектор входных (независимых) переменных (называемых также регрессорами или предикторами) $x \in X \subseteq \boldsymbol{R}^n$, а вектор выходных (зависимых) переменных $y \in Y \subseteq \boldsymbol{R}^1$. В случае, если $n = 1$, говорят о постановке задачи парной регрессии, а если $n > 1$, – множественной регрессии. В более широкой постановке может рассматриваться несколько зависимых переменных, которые объединяются в вектор $y \in R^h$. Если значение зависимой переменной определяется по отношению к будущему моменту времени, которое в этом случае присутствует в перечне компонентов вектора x, то задача регрессии называется задачей прогнозирования. При решении задачи требуется на основе имеющихся исходных данных установить функциональную зависимость $y = \tilde{\varphi}(x)$, которая с определенным приближением отражает реальную зависимость $y = \varphi(x)$ и которую можно далее использовать для предсказания неизвестного значения зависимой переменной для любых значений независимых (входных переменных). Приближение должно удовлетворять некоторому критерию качества.

1. Постановка и решение задачи регрессии в рамках статистического подхода

Первоначально рассмотрим решение задачи регрессии в статистической постановке. В этом случае x и y рассматриваются как значения случайных величин или случайных векторов. При этом считается, что известны полные статистические описания этих величин (векторов) в виде совместной плотности распределения вероятностей $p(x,y)$. Тогда регрессией СВ $\mathbf{y}$ на СВ $\mathbf{x}$ в широком смысле называется функция $\bar{\mathbf{y}} = \bar{\varphi}(\mathbf{x})$, приближенно отражающая статистическую зависимость

$$\mathbf{y} = \bar{\varphi}(\mathbf{x}) + \boldsymbol{\alpha}(\mathbf{x}),$$

где $\boldsymbol{\alpha}(\mathbf{x})$ – случайная ошибка такого представления.

Рассмотрим квадратичный критерий качества аппроксимации зависимости $\mathbf{y}$ от $\mathbf{x}$ для каждой конкретной пары значений y, x

$$J[\|y - \bar{\varphi}(x)\|] = (y - \bar{\varphi}(x))^T (y - \varphi(x)).$$

Тогда наилучшую в статистическом смысле функцию, которая минимизирует ошибку аппроксимации искомой функциональной зависимости, можно получить на основе следующих соотношений:

$$J = M[(\mathbf{y} - \bar{\varphi}(\mathbf{x}))^T (\mathbf{y} - \bar{\varphi}(\mathbf{x}))] = \int (y - \bar{\varphi}(x))^T (y - \varphi(x)) p(x,y) dx dy \to \min_{\varphi}. \quad (1)$$

Проводя выкладки, аналогичные ранее выполненным при получении оптимальных оценок в лекции 6, предусматривающие дифференцирование функционала по компонентам $\bar{\mathbf{y}} = \bar{\varphi}(\mathbf{x})$, нетрудно видеть, что функция, минимизирующая J, имеет вид

$$\bar{y} = \bar{\varphi}(x) = \int y p(y/x) dy. \quad (2)$$

Таким образом, регрессией СВ $\mathbf{y}$ на СВ $\mathbf{x}$ называют условное математическое ожидание $\mathbf{y}$ при фиксированном значении $\mathbf{x}$: $\bar{y} = M[\mathbf{y}/\mathbf{x} = x] = \bar{\varphi}(x)$. Данная функция является в общем случае нелинейной и ее вид зависит от распределения $p(x,y)$.

На практике при решении задачи регрессии изначально ограничивают класс функций, для которых осуществляется минимизация функционала (1). Чаще всего используется линейная функция

$$\bar{y} = \bar{\varphi}(x) = b_0 + bx = b_0 + \sum_{k=1}^{n} b_i x_i,$$

где $b_0 = (b_{10},..b_{h0})^T$ – вектор-столбец размера $h \times 1$; $b = (b_1,...,b_n)$ – матрица коэффициентов линейной регрессии размера $h \times n$, каждый столбец которой имеет вид $b_i = (b_{i1},...,b_{ih})^T$.

Для удобства представления иногда вводят расширение вектора x, добавляя фиктивную компоненту $x_0 \equiv 1$, $x_e = (x_0, x_1,...,x_n)^T$ и объединяют коэффициенты в общую матрицу $b_e = (b_0, b_1,...,b_n)$. Тогда

$$\tilde{y} = b_e x_e = \sum_{k=0}^{n} b_i x_i .$$

Решение задачи сводится к нахождению коэффициентов линейной регрессии путем минимизации функционала

$$J = M[(\mathbf{y} - b_0 - b\mathbf{x})^T(\mathbf{y} - b_0 - b\mathbf{x})] = \int (y - b_0 - bx)^T (y - b_0 - bx)\, p(x,y)dxdy \to \min_{b_e} .$$

Дифференцируя J по элементам $b_e = (b_0, b_1,...,b_n)^T$ и приравнивая частные производные нулю, получим следующие уравнения:

$$\left.\frac{\partial J}{\partial b_0}\right|_{b_0=\tilde{b}_0} = -2\int (y - b_0 - bx)\, p(x,y)dxdy = \mathbf{0} ,$$

$$\left.\frac{\partial J}{\partial b}\right|_{b=\tilde{b}} = 2\int (y - b_0 - bx)x^T p(x,y)dxdy = \mathbf{0}$$

или

$$\tilde{b}_0 = \int (y - bx)\, p(x,y)dy = M[\mathbf{y}] - bM[\mathbf{x}] = m_y - \tilde{b}\, m_x ,$$

$$\tilde{b}\, M[\mathbf{x}\mathbf{x}^T] = \int (yx^T - b_0 x^T)\, p(x,y)dy = M[\mathbf{y}\mathbf{x}^T] - \tilde{b}_0 m_x .$$

Введем матрицы ковариаций случайных векторов

$$R_{xx} = M[(\mathbf{x} - m_x)(\mathbf{x} - m_x)^T], R_{yx} = M[(\mathbf{y} - m_x)(\mathbf{x} - m_x)^T],$$

тогда полученные уравнения можно представить в виде

$$\tilde{b}_0 = m_y - R_{yx}R_{xx}^{-1}m_x, \ \tilde{b} = R_{yx}R_{xx}^{-1} .$$

В итоге получим функцию

$$\tilde{y} = \tilde{\varphi}(x) = m_y + R_{yx}R_{xx}^{-1}(x - m_x), \quad (3)$$

которая и является функцией линейной регрессии.

Следует отметить, что взаимная матрица ковариаций ошибки такого представления $\boldsymbol{\varepsilon} = \mathbf{y} - \tilde{\varphi}(\mathbf{x})$ и $\mathbf{x}$ равна нулю

$$R_{\varepsilon x} = M[(\mathbf{y} - \tilde{\varphi}(\mathbf{x}))(\mathbf{x} - m_x)^T] = M[(\mathbf{y} - m_y - R_{yx}R_{xx}^{-1}(x - m_x))(x - m_x)^T] =$$

$$= R_{yx} - R_{yx}R_{xx}^{-1}R_{xx} = \mathbf{0}.$$

Точно также равна нулю и взаимная ковариация ошибки и регрессии

$$R_{\varepsilon\varphi} = M[(\mathbf{y} - \tilde{\varphi}(\mathbf{x}))(\tilde{\varphi}(\mathbf{x}) - m_\varphi)^T] = M[(\mathbf{y} - m_y - R_{yx}R_{xx}^{-1}(x - m_x))(x - m_x)^T R_{xx}^{-1}R_{xy}] =$$

$$= R_{yx}R_{xx}^{-1}R_{xy} - R_{yx}R_{xx}^{-1}R_{xy} = \mathbf{0}.$$

Данное свойство свидетельствует об ортогональности в статистическом смысле указанных случайных величин. Так как

$$R_{\varphi x} = M[(\tilde{\varphi}(\mathbf{x}) - m_\varphi)(\mathbf{x} - m_x)^T] = M[R_{yx}R_{xx}^{-1}(x - m_x)(x - m_x)^T] = R_{yx},$$

то линейная регрессия соответствует ортогональной проекции $\mathbf{y}$ на $\mathbf{x}$.

В целом следует отметить, что представленные соотношения повторяют известный результат для оптимальной линейной оценки гауссовских случайных векторов, вытекающий из теоремы о нормальной корреляции. Это, в частности, отражает тесную связь задачи оценивания и задачи регрессии. Однако, при получении соотношений (3) ограничений на вид распределения не вводилось, что свидетельствует о том, что полученная оценка является *оптимальной в классе линейных*. В продолжение, следует отметить, что если рассматриваемые случайные векторы являются гауссовскими, то (3) одновременно определяет и уравнение для регрессии, соответствующее общему решению (2).

Таким образом, соотношения (2),(3) дают общее решение задачи построения нелинейной и линейной регрессии СВ при наличии необходимой статистической модели данных.

2. Постановка и решение задачи регрессии в рамках детерминистского подхода по методу наименьших квадратов

На практике чаще все-таки возникает ситуация, когда статистические описания переменных неизвестны или их не интерпретируют как случайные величины. В то же время имеются обучающие наборы измеряемых (наблюдаемых) данных, характеризующие реализации (выборочные значения) этих величин. При этом точная функциональная зависимость не может быть

получена, поскольку эти величины либо являются СВ, между которыми существует статистическая связь и на измерения накладываются случайные ошибки, либо между этими величинами существует функциональная связь, но наблюдения не в полной мере достоверны из-за ошибок измерения. Тогда задача регрессии решается в рамках детерминистского подхода на основе метода наименьших квадратов.

Пусть дана обучающая выборка данных $X^N = \{x^{(1)},...,x^{(N)}\}$, $x^{(i)} \in X \subseteq \mathbf{R}^n$, в которой каждому значению вектора-столбца независимых (входных) переменных соответствует совокупность значений зависимой (выходной) переменной $Y^N = \{y^{(1)},...,y^{(N)}\}$, $y^{(i)} \in Y \subseteq \mathbf{R}^1$ определены на множестве континуум. Требуется построить отображение вида $\tilde{\varphi}: X \to Y$, т.е. алгоритм восстановления регрессии, аппроксимирующие реально действующую зависимость переменных с заданным критерием качества.

Изначально всегда вводится ограничение на вид используемых для аппроксимации функций, которые задаются в виде семейства $\psi(x,a)$, $a \in R^m$, где a – вектор неопределенных параметров модели. В рамках метода наименьших квадратов требуется найти значение $\tilde{a}$, для которого минимизируется функционал

$$J(a) = \sum_{i=1}^{N} (\psi(x^{(i)},a) - y^{(i)})^2 \to \min_a, \tag{4}$$

или как необходимое условие выполняется

$$\left.\frac{\partial J}{\partial a}\right|_{a=\tilde{a}} = 2\sum_{i=1}^{N} (\psi(x^{(i)},a) - y^{(i)}) \frac{\partial \psi(x^{(i)},a)}{\partial a} = 0. \tag{5}$$

Полученная система из m уравнений имеет численное решение, которое позволяет построить регрессионную зависимость в виде

$$\tilde{y} = \psi(x,\tilde{a}). \tag{6}$$

Использование соотношений (6) имеет и статистическое обоснование. Пусть $\mathbf{x}$ и $\mathbf{y}$ СВ, между которыми существует статистическая связь, описываемая следующей регрессией

$$\mathbf{y} = \psi(\mathbf{x},a) + \boldsymbol{\alpha}(\mathbf{x}),$$

где $\boldsymbol{\alpha} \sim N(\alpha,0,\sigma_x^2)$ является гауссовской СВ с нулевым математическим ожиданием и дисперсией σ_x^2, которая в общем случае зависит от значения $\mathbf{x}$.

Тогда для обучающей выборки может быть записана следующая функция правдоподобия

$$P(Y^N / X^N, a) = \prod_{i=1}^{N} N(y^{(i)}, \psi(x^i, a), \sigma_i^2).$$

Максимум логарифма правдоподобия относительно вектора неизвестных параметров модели a

$$\ln P(Y^N / X^N, a) = -\frac{N}{2}\ln(2\pi) - \sum_{i=1}^{N}\sigma_i - \sum_{i=1}^{N}\frac{1}{2\sigma_i^2}\left(y^{(i)} - \psi(x^{(i)}, a)\right)^2$$

достигается при выполнении условия минимума последнего слагаемого, которое зависит от a. Это слагаемое, по сути, является эквивалентом функции, которая минимизируется при реализации критерия взвешенных наименьших квадратов при использовании в качестве весовых коэффициентов величин $w_i = 1/2\sigma_i^2, i = \overline{1,N}$. Если дисперсии одинаковы $\sigma_i^2 = \sigma_\alpha^2,\ i = \overline{1,N}$ и не зависят от $\mathbf{x}$, то решение в рамках статистической постановки имеет точное соответствие с решением (6), получаемым на основе стандартного МНК.

Линейная параметрическая регрессия. Рассмотрим первоначально решение задачи построения линейной регрессии на основе (4-6). Тогда в качестве $\psi(x,a)$ используется линейная функция

$$\psi(x,a) = a_0 + \sum_{k=1}^{n} a_k x_k,$$

где a_0 – скалярная величина; $a_k = (a_1,...,a_n)^T$ – вектор коэффициентов при компонентах входного вектора размера $n \times 1$. Введем расширение вектора x, добавляя фиктивную компоненту $x_0 \equiv 1$, $x_e = (x_0, x_1,...,x_n)^T$ и объединяют коэффициенты в общий вектор $a = (a_0, a_1,...a_n)^T$. Тогда

$$\psi(x,a) = a^T x_e = \sum_{k=0}^{n} a_i x_i.$$

Введем матрицу X размера $N \times (n+1)$, строками которой являются транспонированные расширенные векторы входной обучающей выборки, и вектор выходных данных обучающей выборки

$$X = \begin{pmatrix} x_e^{(1),T} \\ x_e^{(2),T} \\ \vdots \\ x_e^{(N),T} \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(N)} & \cdots & x_n^{(N)} \end{pmatrix},\ y = (y^{(1)},...,y^{(N)})^T.$$

Из общего решения (5) получим следующую систему $n+1$ уравнений в матричном виде, обеспечивающих минимизацию невязки $\|Xa-y\|^2$

$$J(a)=\|Xa-y\|^2=(Xa-y)^T(Xa-y)\to\min_a$$

$$\left.\frac{\partial J}{\partial a}\right|_{a=\tilde{a}}=2X^T(Xa-y)=2\sum_{i=1}^{N}(a^Tx_e^{(i)}-y^{(i)})x_e^{(i)}=0,\quad \frac{\partial\psi(x^{(i)},a)}{\partial a}=\frac{\partial a^Tx_e^{(i)}}{\partial a}=x_e^{(i)}$$

Тогда исходная система уравнений для нахождения $\tilde{a}$ перепишется в виде другой системы, называемой нормальной системой, которая в матричной форме имеет следующий вид:

$$X^Ty-X^TX\tilde{a}=0.$$

Нормальная система всегда совместна. Матрица $H=X^TX$ является квадратной матрицей размера $(n+1)\times(n+1)$. В случае, если матрица $H=X^TX$ является невырожденной, существует единственное, так называемое нормальной решение

$$\tilde{a}=(X^TX)^{-1}X^Ty=X^+y,\ X^+=(X^TX)^{-1}X^T. \qquad (7)$$

Матрица $X^+=(X^TX)^{-1}X^T$ является в этом случае псевдобратной матрицей для X. Она обладает свойствами, во многом схожими со свойствами обратной матрицы, а именно

$$X^+X=(X^TX)^{-1}X^TX=I,\quad X^+XX^+=(X^TX)^{-1}X^TX(X^TX)^{-1}X^T=X^+,$$

$$XX^+X=X(X^TX)^{-1}X^TX=X.$$

Отметим, что если матрица X является квадратной (объем обучающей выборки равен числу оцениваемых параметров регрессии) и невырожденной, то полученное решение является решением системы линейных алгебраических уравнений (СЛАУ) $Xa=y$ по правилу Крамера: $\tilde{a}=X^{-1}y$. При этом невязка тождественно равна нулю.

Однако на практике, в основном, СЛАУ $Xa=y$, является переопределенной (объем обучающей выборки превышает число оцениваемых параметров регрессии). В этом случае вектор $\tilde{a}$ может быть определен на основе (7).

Таким образом, окончательное решение задачи имеет вид

$$\tilde{y}=\psi(x,\tilde{a})=\tilde{a}^Tx=\tilde{a}_0+\sum_{k=1}^{n}\tilde{a}_ix_i. \qquad (8)$$

Величины

$$e^{(i)} = y^{(i)} - \bar{y}^{(i)}, i = \overline{1,N}\ , \quad \bar{y}^{(i)} = \bar{a}^T x^{(i)}$$

в (8) называются остатками. Анализ остатков весьма важен для анализа полученного решения с точки зрения его адекватности и пригодности полученной модели к использованию. При реализации МНК сумма остатков равна нулю

$$\sum_{i=1}^{N} e^{(i)} = \sum_{i=1}^{N} (y^{(i)} - \bar{y}^{(i)}) = 0.$$

В качестве самостоятельного задания можно предложить рассмотреть решение задачи линейной регрессии для случая, когда выходная реакция является вектором $y^{(i)} \in Y \subseteq R^h$.

Проверка значимости модели регрессии. Анализ полученной математической модели регрессии предполагает использование различных методов и алгоритмов математической статистики. Одной из центральных задачей анализа уравнения регрессии является проверка адекватности и значимости выбранной модели. Проверить адекватность уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным, и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной. Это нужно сделать для того, чтобы убедиться в пригодности и надежности модели для использования в прикладных целях, в частности для прогнозирования переменной y в будущем. Проверка адекватности регрессионной модели основана на сравнении рассеивания экспериментальных значений наблюдаемой переменной относительно линии регрессии с рассеиванием этих значений относительно ее среднего значения $\bar{y}$, которое с учетом свойств суммы остатков равно

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)} = \frac{1}{N}\sum_{i=1}^{N} \bar{y}^{(i)} = \frac{1}{N}\sum_{i=1}^{N} \bar{a}^T x^{(i)} = \bar{a}^T \bar{x}, \quad \bar{y}^{(i)} = \bar{a}^T x^{(i)}.$$

В проверке модели используется разложение общей суммы квадратов отклонений переменной y от среднего, которое имеет вид

$$TSS = \sum_{i=1}^{N} (y^{(i)} - \bar{y})^2 = \sum_{i=1}^{N} (\bar{y}^{(i)} - \bar{y})^2 + \sum_{i=1}^{N} (y^{(i)} - \bar{y}^{(i)})^2 = ESS + RSS,$$

Левая часть равенства – это общая сумма квадратов отклонений (total sum of squared) –TSS, которая равна сумме квадратов отклонений от среднего,

объясненных регрессией (explained sum of squared) – ESS и остаточной сумме квадратов (residual sum of squared) – RSS, определяемая случайными остатками. Считается, что чем меньше RSS, тем регрессия лучше. И, наоборот, чем меньше ESS, тем регрессия хуже. Несмещенной оценкой дисперсии ошибки измерения выходной наблюдаемой переменной или случайной ошибки – шуму, действующего при получении наблюдений, является величина так называемой остаточной дисперсии

$$\sigma_{\tilde{y}}^2 = \frac{RSS}{N-n-1} = \sum_{i=1}^{N}(y^{(i)} - \bar{y}^{(i)})^2 .$$

где $N-n-1=m$ – число степеней свободы, которое определяется как разность между числом точек, в которых ставились эксперименты N, по которым оценивают коэффициенты a и числом этих коэффициентов $n+1$. Для проверки адекватности уравнения регрессии необходимо, чтобы число степеней свободы было больше нуля. При отрицательном числе степеней свободы, т. е. при ($N<n+1$) метод наименьших квадратов не может быть использован. При $m=0$ уравнение регрессии может быть получено, однако статистический анализ этого уравнения провести нельзя, ибо для проверки адекватности модели не остается степеней свободы. При $m>0$ остаются степени свободы для проверки адекватности регрессионной модели.

Вводится коэффициент детерминации

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} .$$

Этот коэффициент показывает качество подгонки регрессионной модели к наблюдаемым значениям. Если он равен нулю, то регрессия на y на x не улучшает качества предсказания значения y' по сравнению с тривиальным предсказанием $y'=\bar{y}$. Если он равен своему максимальному значению – единице, то это означает точную подгонку регрессии под экспериментальные данные (все остатки равны нулю и имеет место строгая функциональная зависимость). Для приемлемых моделей предполагается, что коэффициент детерминации должен быть не меньше 50 %. Иногда коэффициент детерминации дает слишком оптимистические результаты, поскольку значение R^2 возрастает с ростом числа переменных (регрессоров), что не означает улучшения качества предсказания, и потому вводится скорректированный коэффициент детерминации

$$R^2 = 1 - \frac{RSS/(N-n-1)}{TSS/(N-1)} .$$

При реализации МНК используются методики проверки гипотезы о нулевых значениях различных комбинаций коэффициентов регрессии. Проверка значимости всей модели предполагает проверку гипотезы H_0 о равенстве нулю всех коэффициентов регрессии, кроме коэффициента a_0 при фиктивной переменной $x_0 \equiv 1$. В этом случае рассматривается отношение оценок дисперсий

$$F = \frac{ESS/n}{RSS/(N-n-1)}.$$

Данная величина при несмещенном, гауссовском характере случайных ошибок измерения выходной переменной имеет статистику распределения Фишера $F(n, N-n-1)$. В общем случае произвольного распределения ошибок такая статистика является асимптотической. Если

$$F < F_\gamma(n, N-n-1),$$

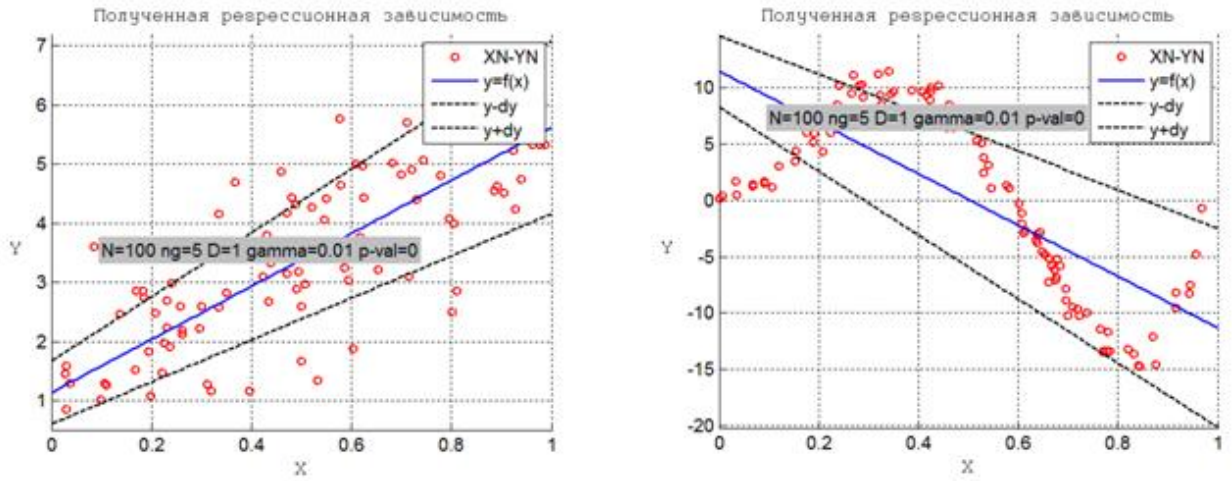
то гипотеза H_0 о равенстве нулю всех коэффициентов, кроме, может быть, a_0, подтверждается. Здесь F_γ – квантиль F - распределения уровня доверительной вероятности $1-\gamma$, показывающий пороговый уровень статистики, который превышается с вероятностью γ. Это означает, что необходимо исключать часть входных переменных из рассмотрения.

Следует отметить, что при применении критерия Фишера может использоваться и обратное отношение остаточной дисперсии к дисперсии отклонений регрессии от среднего. Тогда используют статистику $F(N-n-1, n)$. Обычно тест применяют так, чтобы в числителе находилось потенциально большая величина. В этом случае удобным способом проверки гипотез является использования p-значения – вероятности того, что случайная величина с данным распределением Фишера превысит данное значение статистики. Если p-значение меньше уровня значимости, то нулевая гипотеза отвергается, в противном случае принимается.

Далее будут рассмотрены несколько примеров, иллюстрирующих возможности программной реализации методов регрессионого анализа. Для этого использовались возможности среды MATLAB (раздел Statistics Toolbox), содержащей большое количество стандартных функций для построения регрессий различного вида, анализа регрессионных моделей и обеспечения возможностей построения и визуализации регрессионных характеристик в интерактивном режиме.

В представленных ниже примерах реализуется построение регрессионной линейной модели для различных вариантов задания исходной восстанавливаемой модели (в виде полинома, в виде гармонического ряда). Проводится генерация обучающей выборки данных $X^N = \{x^{(1)},...,x^{(N)}\}$, $x^{(i)} \in X \subseteq R^n$, как реализаций входного вектора x размерности $n = 1$ с равномерным распределением в заданном диапазоне значений. Значения выходной переменной $Y^N = \{y^{(1)},...,y^{(N)}\}$, $y^{(i)} \in Y \subseteq R^1$ генерируются как $y^{(i)} = a^T x^{(i)} + \alpha_i$, где α_i, $i = 1,N$ центрированные значения гауссовской ошибки с дисперсией $D = \sigma_\alpha^2$. Далее проводится построение линейной регрессии на основе описанной выше стандартной функции.

Кроме того, в примерах проводится вычисление коэффициента детерминации R^2, статистики Фишера $F(n, N-n-1)$, расчет значения F-статистики, который превышается с вероятностью gamma, расчет вероятности превышения полученного значения F (p-значение). Полученные значения этих величин, показывают,что полученная регрессионная модель является статистически значимой (p_value не превосходит gamma). На графиках рис.1а,б показаны зависимости для различных вариантов задания исходной восстанавливаемой модели.



а) б)

Рис.1. Примеры восстановления регрессии: исходная зависимость линейная (var1=1) – а; исходная зависимость в виде гармонического ряда (var1=2) – б

Как следует из представленных графиков, использование линейной регрессии возможно и в случае, если исходная модель является существенно нелинейной. Однако, если изначально известен характер зависимости, для

этого целесообразно использовать другие варианты алгоритма восстановления регрессии на основе МНК, которые будут рассмотрены ниже.

Метод псевдобратной матрицы и метод регуляризации в задачах регрессии. В общем случае, в том числе и в случае вырожденности матрицы $H = XX^T$, решение задачи на основе МНК можно получить в виде

$$\vec{a} = X^+ y,$$

где X^+– псевдообратная матрица, которая обладает ранее представленными свойствами.

Псевдообратная матрица, может вычисляться различными способами, например, с использованием известного алгоритма сингулярного разложения (SVD), в основе которого лежит следующее представление:

$$X = VDU^T,$$

где V– ортогональная матрица размера $(N+1)\times(n+1)$, столбцы которой $v^{(k)}$ являются собственными векторами матрицы XX^T; U– ортогональная матрица размера $(n+1)\times(n+1)$, столбцы которой$u^{(k)}$ являются собственными векторами матрицы $X^T X$; D– диагональная $(n+1)\times(n+1)$ матрица, диагональными элементами которой являются $\boldsymbol{\lambda}_1^{1/2},\ldots,\boldsymbol{\lambda}_n^{1/2}$, т.е. квадратные корни общих собственных чисел XX^T и $X^T X$. Число ненулевых собственных значений среди $\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_n$ равно рангу матриц X, XX^T $X^T X$.

Имея сингулярное разложение, можно записать псевдообратную матрицу в виде:

$$X^+ = UD^{-1}V^T = \sum_{k=1,\lambda_k\neq 0}^{n} \frac{1}{\sqrt{\boldsymbol{\lambda}_k}} u^{(k)} v^{(k),T},$$

где суммирование происходит только для отличных от нуля $\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_n$. В итоге общее решение задачи линейной регрессии на основе МНК может быть представлено в виде

$$\vec{a} = X^+ y = \sum_{k=1,\lambda_k\neq 0}^{n} \frac{1}{\sqrt{\boldsymbol{\lambda}_k}} u^{(k)} \left(v^{(k),T} y\right). \qquad (9)$$

Полученные соотношения (8),(9) не всегда позволяют обеспечить устойчивость решения задачи, особенно в случае, когда матрица $H = X^T X$ плохо обусловлена и возникает ситуация «мультиколлинеарности». Столбцы такой матрицы «почти линейно зависимы», а среди собственных чисел есть

числа близкие к нулю. Фактически это означает, что элементы обучающей выборки локализованы в пространстве меньшей размерности. В результате этих эффектов, обращение матрицы матрица численно неустойчиво, что может привести к возникновению существенных погрешностей при расчете коэффициентов регрессии.

В подобной ситуации наиболее эффективным методом решения задачи является метод регуляризации А.Н. Тихонова. В его основе лежит изменение функционала МНК путем добавления регуляризирующего слагаемого следующим образом:

$$J'(a) = \|Xa - y\|^2 + \beta\|a\|^2 = (Xa - y)^T (Xa - y) + \beta a^T a \to \min_a,$$

где β – малая величина, выбор которой представляет самостоятельную задачу. В результате в методе регуляризации реализуются две стратегии: минимизация невязки $\|Xa - y\|^2$, как и в МНК, и, одновременно, минимизация нормы решения, что обеспечивает гладкость устойчивость получаемого решения относительно используемых исходных данных. Приравнивая нулю производную $J'(a)$ относительно a, нетрудно видеть, что в данном случае решение задачи регрессии имеет вид

$$\tilde{a} = (\beta I + X^T X)^{-1} X^T y \quad . \tag{10}$$

Полученная в таком виде регрессия называется «гребневой» регрессией, поскольку при обращении к матрице $H = X^T X$ добавляется «гребень» βI, обеспечивающий регуляризацию решения. Если $\beta = 0$, то (10) переходит в (8). Увеличение β позволяет сделать решение более гладким и устойчивым, но при этом невязка увеличивается.

Для нахождения разумного значения β применяются различные методы, из которых наиболее часто применяемым является метод невязки. В нашем случае, если считать, что элементы выборки входных данных известны точно, а выходная реакция для каждого из них фиксируется с определенной погрешностью, этот метод предполагает выбор β на основе следующего соотношения:

$$\|X\tilde{a}_\beta - y\| = \delta, \ \ \|y\text{-}\hat{y}\| \le \delta,$$

где δ – определяет задаваемый уровень погрешности при получении исходных данных, которая, собственно, и определяет неустойчивость решении; $\tilde{a}_\beta$ – решение, полученное для конкретного значения β. Таким

образом, варьируя параметр β, можно найти компромиссное значение, при котором достигается требуемый уровень невязки.

Еще один вариант предполагает выбор значения параметра таким образом, чтобы обращаемая при регуляризации матрица $\beta I + X^T X$ имела бы заранее заданный «хороший» коэффициент обусловленности R_0 (например, не выше 10^3). Этот коэффициент определяется как отношение максимального собственного числа матрицы к минимальному собственному числу анализируемой матрицы. Тогда следует выбирать β исходя из следующего соотношения

$$cond(\beta I + X^T X) = \frac{\lambda_{\max} + \beta}{\lambda_{\min} + \beta} = R_0 ,$$

откуда следует рекомендация $\beta_э \approx \lambda_{\max} / R_0$.

В качестве примера реализации метода регуляризации при построении линейной регрессии рассмотрим пример, полученный на основе программы в среде Matlab. В ней проводится генерация обучающей выборки данных $X^N = \{x^{(1)},...,x^{(N)}\}$, $x^{(i)} \in X \subseteq R^n$, как реализаций гауссовского входного вектора x размерности $n = 3$. При этом матрица ковариаций задается следующим образом

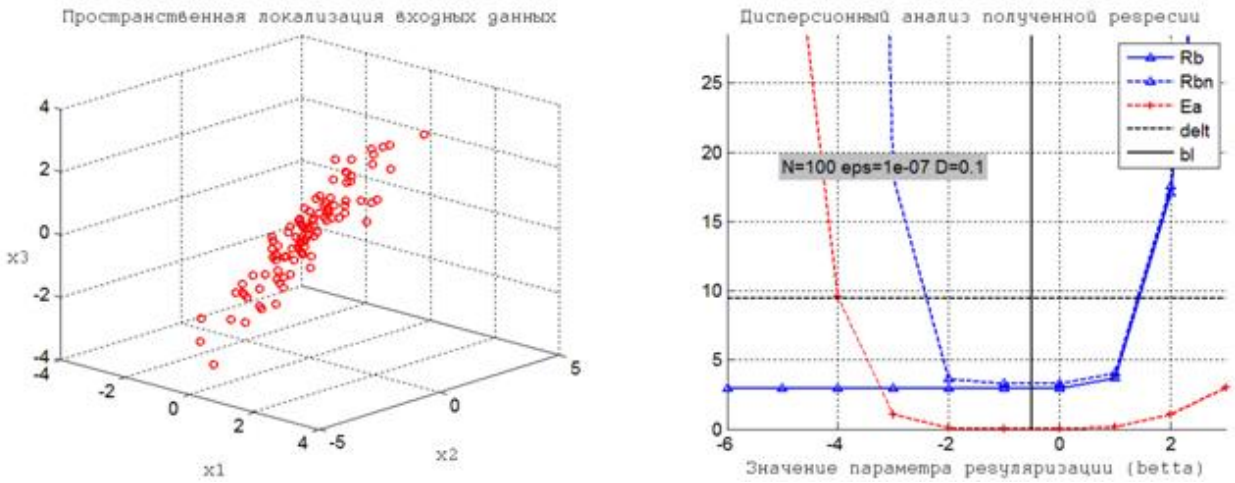
$$C = \begin{pmatrix} 1 & \cdots & 0 & 1-\varepsilon \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \vdots & 1 & 1-\varepsilon \\ 1-\varepsilon & \cdots & 1-\varepsilon & n-1 \end{pmatrix},$$

где $\varepsilon > 0$ – малая величина, определяющая степень обусловленности матрицы C. Такой вид матрицы означает, что первые $n-1$ компоненты вектора x независимы, и $n-$ ая компонента весьма сильно зависит от их суммы (при $\varepsilon = 0$ она в точности является суммой первых $n-1$ компонент вектора x). Данный прием позволяет проиллюстрировать эффект мультиколлинеарности. Значения выходной переменной $Y^N = \{y^{(1)},...,y^{(N)}\}$, $y^{(i)} \in Y \subseteq R^1$ генерируются как $y^{(i)} = a^T x^{(i)} + \alpha_i$, где α_i, $i = 1,N$ центрированные значения гауссовской ощибки с дисперсией $D = \sigma_\alpha^2$. Далее в цикле с изменяющимся значением β проводится построение линейной регрессии на основе описанных выше соотношений (10) и проводится ее дисперсионный анализ. При этом рассчитываются: величина суммы квадратов остатков для исходных данных (Rb), величина суммы квадратов остатков для новых данных, не участвовавших в обучении (Rbn), величина суммы квадратов ошибок оценки

коэффициентов регрессии (Ea). Вне цикла рассчитываются значение порогового уровня невязки как $\delta = 3 * sqrt(ND)$, а также величина $\beta_э \approx \lambda_{max} / R_0$.

При выполнении программы выводятся значения чисел обусловленности матриц $C, X^T X, X^T X + \beta I$, определяемых на основе стандартной функции cond(…), проводится визуализация данных в трехмерном пространстве для отображения мультиколлинеарности, а также графики зависимостей Rb, Rbn, Ea от β с отображением линий постоянного уровня δ по горизонтали и $\beta_э$ по вертикали.

Полученные при запуске программы результаты показывают, что за при использовании значения $\beta_э \approx \lambda_{max} / R_0$ обеспечивается необходимый уровень обусловленности обращаемой матрицы $X^T X + \beta I$. Одновременно производится визуализация области локализации данных и графиков зависимостей Rb, Rbn, Ea от β, представленные на рис.2а,б.



а) б)

Рис.2. Пример применения гребневой регрессии: область локализации данных – а; зависимости характеристик ошибок регрессии от параметра регуляризации – б

Анализ первого рисунка показывает, что данные локализованы фактически в плоскости, что есть отражение их мультиколлинеарности. На втором рисунке мы видим характерный минимум дисперсии ошибки для значений регрессии при использовании данных, не участвовавших в обучении, а также – для суммарной СКО для оценки параметров регрессионной модели, которые достигается в окрестности рекомендуемого

значения параметра регуляризации. При очень малых β полученная модель регрессии работает крайне неустойчиво, а при больших β ошибки возрастают вследствие того, что значения получаемых оценок a_i, $i=\overline{0,n}$ стремятся к нулю.

Расширения линейной регрессии и нелинейная регрессия. Очевидно, что полученное выше решение задачи регрессии может быть использовано при произвольном нелинейном преобразовании $x:\bar{x}=\psi(x)$, если неизвестные параметры входят в (4-6) линейным образом:

$$\psi(x,a)=a^T\psi(x),\ \ \bar{y}=\bar{a}^T\psi(x).$$

Пусть отображение $\psi: R^n \to R^{s+1}$ осуществляет перевод исходного пространства входных переменных в новое пространство в общем случае другой размерности. Тогда $\psi(x)=(1,\psi_1(x),...,\psi_s(x))^T$. Вводя для всех элементов обучающей выборки замену $\bar{x}^{(i)}=\psi(x^{(i)}),\ i=\overline{1,N},\bar{x}\in\bar{X}\subseteq R^s$, можно перейти к поиску решения задачи линейной регрессии для новых переменных

$$\bar{y}=\psi(x,\bar{a})=\bar{a}_0+\sum_{k=1}^{s}\bar{a}_i\bar{x}_i=\bar{a}^T\bar{x}_e,$$

где $\bar{a}$ ищется на основе соотношений, аналогичных (4)-(6).

В качестве примера можно привести использование представление полиномов n переменных различной степени до d включительно в виде линейной функции $s+1$ переменных $s+1=C_{n+d}^n$ (с добавлением фиктивной переменной $\bar{x}_0\equiv 1$)

$$\bar{x}_0=1,\ \bar{x}_i=x_i,\ i=\overline{1,n},\ \bar{x}_{n+1}=x_1x_1,\ \bar{x}_{n+2}=x_1x_2,...,\bar{x}_{n(n+1)-1}=x_nx_{n-1},\bar{x}_{n(n+1)}=x_nx_n,...$$

При таком представлении каждое слагаемое искомой регрессии имеет вид

$$a_k\psi_k(x)=a_{i_1..i_n}x_1^{i1}\times x_2^{i2}\times...x_n^{in},i_1+i_2+..i_n\le d,\ \ k=\overline{1,s+1},$$

где $a_{i_1..i_n}$ – исходные коэффициенты, которые приравниваются к коэффициентам линейной регрессии a_k после замены переменных. В результате всегда можно представить исходный полином, описывающий линейную регрессию по отношению к s новым входным переменным. Следует еще раз обратить внимание, что при такой замене элементы исходной обучающей выборки преобразуются в элементы обучающей выборки для новых переменных эквивалентным образом.

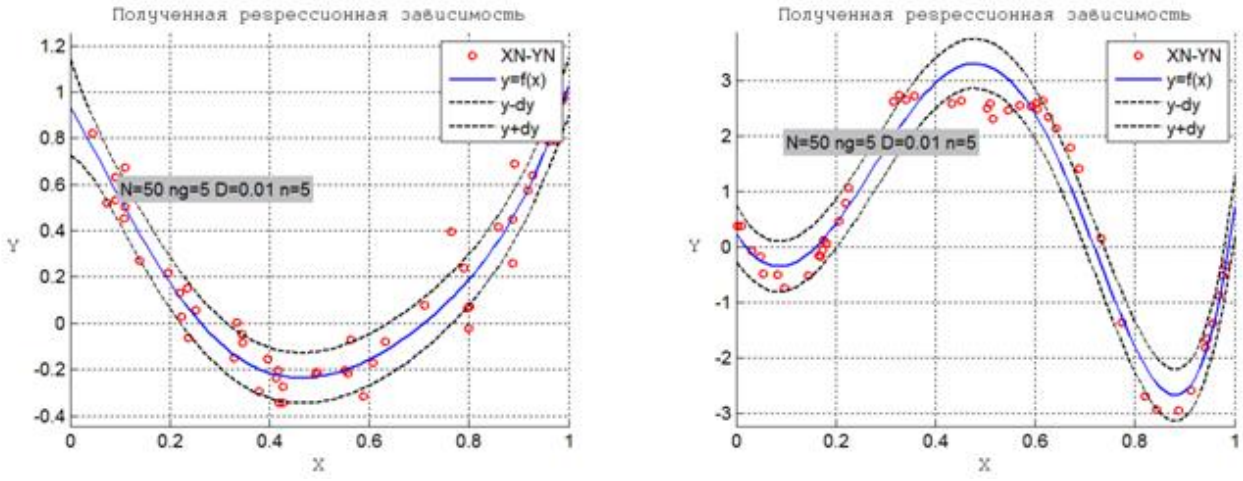
В частном случае, когда входная переменная является скалярной величиной, регрессия ищется изначально в виде

$$\varphi(x,a) = a^T \varphi(x) = \sum_{k=0}^{d} a_k x^d .$$

Такая регрессия в литературе называется полиномиальной и ее использование позволяет повысить точность представления данных в случаях, когда исходная модель, определяющая вид восстанавливаемой зависимости, является нелинейной.

В представленной ниже примере реализуется построение полиномиальной регрессии модели для различных вариантов задания исходной, восстанавливаемой модели (в виде полинома, в виде гармонического ряда). Проводится генерация обучающей выборки данных $X^N = \{x^{(1)},...,x^{(N)}\}$, $x^{(i)} \in X \subseteq R^n$, как реализаций входного вектора x размерности $n = 1$ с равномерным распределением в заданном диапазоне значений. Значения выходной переменной $Y^N = \{y^{(1)},...,y^{(N)}\}$, $y^{(i)} \in Y \subseteq R^1$ генерируются как $y^{(i)} = a^T x^{(i)} + \alpha_i$, где α_i, $i = 1,N$ центрированные значения гауссовской ощибки с дисперсией $D = \sigma_\alpha^2$.

На рис.3а,б представлены полученные зависимости построенной полиномиальной регрессии и границ доверительных интервалов для двух случаев: когда в качестве восстанавливаемой исходной модели используется полиномиальная модель, и, когда в качестве восстанавливаемой исходной модели используется модель в виде гармонического ряда.



а) б)

Рис.3. Примеры восстановления полиномиальной регрессии: исходная полиномиальная зависимость – а; исходная зависимость в виде гармонического ряда – б

Анализ представленных графиков показывает, что в данном случае использование регрессии полиномиального типа позволяет достаточно эффективно восстанавливать нелинейные зависимости различного вида.

Таким образом, можно видеть, что использование линейной регрессии после нелинейного преобразования исходных входных переменных в большинстве случаев позволяет решить задачу восстановления нелинейной функциональной зависимости произвольного вида.

Тем не менее, иногда при решении задачи регрессии в рамках МНК пользуются исходным нелинейным представлением $y = \psi(x,a)$. Пусть задан функционал

$$J(a) = \sum_{i=1}^{N} (\psi(x^{(i)},a) - y^{(i)})^2 \to \min_a ,$$

при этом его минимум не может быть найден точно на основе системы (4). В этом случае для нахождения МНК применяется численная минимизация на основе итеративного алгоритма Ньютона-Рафсона

$$a^{(t+1)} = a^{(t)} + \eta_t \left(J''(a^{(t)})\right)^{-1} J'(a^{(t)}), \quad t = 1,2,...$$

где η_t – скалярная величина, определяющая величину шага итераций; $J'(a^{(t)})$ – вектор - градиент функционала $J(a)$ в точке $a^{(t)}$

$$J'(a^{(t)}) = \left.\frac{\partial J}{\partial a}\right|_{a=a^{(t)}} = 2\sum_{i=1}^{N} (\psi(x^{(i)},a) - y^{(i)}) \frac{\partial \psi(x^{(i)},a)}{\partial a} ,$$

а $J''(a^{(t)})$– матрица вторых частных производных (гессиан)

$$J''(a^{(t)}) = \left\| \frac{\partial^2 J(a)}{\partial a_k \partial a_s} \right\|_{a=a^{(t)}}, \quad \frac{\partial^2 J(a)}{\partial a_k \partial a_s} = 2\sum_{i=1}^{N} \frac{\partial \psi(x^{(i)},a)}{\partial a_k} \frac{\partial \psi(x^{(i)},a)}{\partial a_s} - 2\sum_{i=1}^{N} (\psi(x^{(i)},a) - y^{(i)}) \frac{\partial \psi(x^{(i)},a)}{\partial a_k \partial a_s}$$

.

Таким образом, можно, вычисляя градиент, гессиан и обращая последний, вычислить необходимые коэффициенты для реализации итеративного алгоритма.

Еще более простое и практичное решение может быть использовано, если выполняются необходимые условия дифференцируемости функции $y = \psi(x,a)$. Используют прием линеаризации, предполагающий на каждом шаге выполняемой итерации применение следующего приближения

$$\psi(x^{(i)},a) = \psi(x^{(i)},a^{(t)}) + \sum_{k=1}^{n} \frac{\partial \psi(x^{(i)},a^{(t)})}{\partial a_k}(a_k - a_k^{(t)})$$

При использовании подобной линеаризации второе слагаемое в выражении для гессиана становится равным нулю и, тогда, итеративный алгоритм подбора параметров регрессии будет выглядеть следующим образом:

$$a^{(t+1)} = a^{(t)} + \eta_t \left(F_t^T F_t\right)^{-1} F_t^T (\psi_t - y), \quad t = 1,2,\ldots,$$

$$F_t = \|f_{ik}\|, \quad f_{ik} = \left.\frac{\partial \psi(x^{(i)},a)}{\partial a_k}\right|_{a=a^{(t)}}, \quad i = \overline{1,N}, k = \overline{1,n}$$

где $\psi_t = (\psi(x^{(1)},a^{(t)}),\ldots,\psi(x^{(N)},a^{(t)}))^T$ – вектор значений аппроксимирующей функции для полученного на t - ой итерации значения параметра $a^{(t)}$; F_t – матрица частных производных размера $N \times n$, вычисляемых на t - ой итерации. Этот метод называется методом Ньютона–Гаусса. Анализ полученного уравнения показывает, что в этом методе задача нелинейной регрессии фактически сводится к решению последовательности задач линейной регрессии относительно разности $\psi_t - y$, рассматриваемой как новые наблюдения.

◀ Лекция 12
Перейти на...
Перейти на...