УДК 62-50

ПАВЛОВ А.А.,
КАЛАШНИК В.В.,
КОВАЛЕНКО Д.А.

ПОСТРОЕНИЕ МНОГОМЕРНОЙ ПОЛИНОМИАЛЬНОЙ РЕГРЕССИИ. РЕГРЕССИЯ С ПОВТОРЯЮЩИМИСЯ АРГУМЕНТАМИ ВО ВХОДНЫХ ДАННЫХ

В статье описывается метод проведения многомерной полиномиальной регрессии для случая одинаковых значениях аргументов во входных данных. Алгоритм основан на сведении многомерной регрессии к ряду одномерных и последующем решении переопределенной системы линейных уравнений. Также, рассматривается проведение одномерного регрессионного анализа при повторяющихся аргументах во входных данных.

The article describes method of multidimensional polynomial regression in case of equal arguments values in the input data. The algorithm is based on reducing multidimensional regression to a series of one-dimensional and then solving overdetermined system of linear equations. Also article features one-dimensional regression method in case of repeating arguments in the input data.

Вступление

Проблема построения многомерной полиномиальной регрессии по данным с шумом является одной из востребованных на практике задач. В [1] приведен метод построения многомерной полиномиальной регрессии по избыточному описанию в условиях активного эксперимента. Алгоритм основан на сведении многомерной регрессии к одномерной с помощью фиксации всех переменных кроме одной. Этот метод имеет тот недостаток, что в одномерных регрессиях имеется большое количество членов с небольшими значениями степеней при аргументах. Такие коэффициенты полиномов восстанавливаются плохо, что приводит к накоплению ошибок при нахождении коэффициентов исходной модели.

В данной работе приведен модифицированный алгоритм построения многомерной полиномиальной регрессии, основанный на алгоритме из [1]. Также детально рассматривается построение полиномиальной регрессии при претворяющихся аргументах во входных данных.

Построение многомерной полиномиальной регрессии

Пусть многомерная модель задается в виде

$$y\left(\bar{x}\right)=\sum_{\forall(i_1,\dots i_t)\in \mathrm{K}}\ \sum_{\forall(j_1,\dots j_t)\in \mathrm{K}(i_1,\dots i_t)} b_{i_1\ \dots i_t}^{j_1\dots j_t}\left(x_{i_1}\right)^{j_1}\cdot\left(x_{i_2}\right)^{j_2}\dots\left(x_{i_t}\right)^{j_t}+E, \qquad (1)$$

где $\bar{x}=\left(x_1\dots x_n\right)^{\mathrm{T}}$ – детерминированный вектор входных переменных, $b_{i_1\ \dots i_t}^{j_1\dots j_t}$ – неизвестные коэффициенты, j_l – натуральные числа; i_l – натуральные индексы из множества $\{1,\dots,n\}$; E – случайная величина с нулевым математическим ожиданием и ограниченной неизвестной дисперсией σ_E^2 (может быть известна верхняя оценка σ_E^2).

Модель (1) является избыточной – возможно, некоторые из коэффициентов $b_{i_1\ \dots i_t}^{j_1\dots j_t}$ равны нулю.

Пусть все значения аргументов изменяются одинаково $x_1=x_2=\dots x_n=x$.

Тогда, при такой схеме, многомерная регрессия превратится в одномерную регрессию вида:

$$\theta_0^0+\theta_1^0 x+\theta_2^0 x^2+\dots+\theta_{n_0}^0 x^{n_0} \qquad (2)$$

Оценки коэффициентов $\theta_0^0,\theta_1^0,\dots,\theta_{n_0}^0$ находим по результатам одномерного регрессионного анализа.

Модель (2) позволяет построить систему из n_0+1 линейных равенств, связывающих числа $\theta_0^0,\theta_1^0,\dots,\theta_{n_0}^0$ с коэффициентами $b_{i_1\ \dots i_t}^{j_1\dots j_t}$ модели (1).

В левую часть *i*-того уравнения входят коэффициенты, которые находятся при x в *i* -той степени, в правой части – оценки $\theta_0^0, \theta_1^0, ..., \theta_{n_0}^0$.

Далее, фиксируем одну переменную $x_1 = x_{1\Phi}$. Остальные переменные пусть изменяются одинаково $x_2 = ...x_n = x$. Аналогично к предыдущему шагу, получили одномерную регрессию вида $\theta_0^1 + \theta_1^1 x + \theta_2^1 x^2 + ... + \theta_{n_1}^1 x^{n_1}$.

Получив оценки $\theta_0^1, \theta_1^1, ..., \theta_{n_1}^1$, мы можем построить систему линейных уравнений, аналогично первому случаю, однако теперь в левой части уравнений будут находится не только коэффициенты начальной модели, но и фиксированная переменная $x_{1\Phi}$.

Аналогично последовательно проводим активные эксперименты для $i = \overline{2,n}$ с последующим применением регрессионного анализа. В результате получим *n-1* систем линейных уравнений.

Для нахождения оценок коэффициентов исходной модели нужно объединить полученные системы и решить их.

Запишем полученную систему в матричном виде:

$$Ab = \theta \qquad (3)$$

A - матрица коэффициентов системы;

b - вектор неизвестных коэффициентов $b_{i_1 \ldots i_t}^{j_1 \ldots j_t}$ модели (1), размерностью r;

θ - вектор коэффициентов одномерных регрессий.

Решая систему (3) методом наименьших квадратов, мы гарантированно получим оценки коэффициентов, если в матрице A найдется r линейно независимых строк, где r – количество неизвестных коэффициентов исходной модели (исключая свободный член). Поскольку, за построением матрица A не имеет комплексных чисел, для проверки количества линейно независимых строк можно использовать тот факт, что [2]

$$rang(A^T A) = rang(A) \qquad (4)$$

Если $\det(A^T A) \neq 0$, то $rang(A) = r$. Тогда матрица $A^T A$ является невырожденной, и решение (3) невырожденное и единственное.

Если в системе (3) есть *r* линейно независимых строк ($rang(A) = r$), то мы можем утверждать, что метод наименьших квадратов даст решение, которое и есть коэффициентами исходной модели. В противном случае решения задачи нет.

Примечания

1. Проблема точности вектора правых частей решается следующим образом. В [1] показано, что оценки коэффициентов одномерных регрессий при низких степенях имеют большую дисперсию. Таким образом, при наборе линейных уравнений из одномерных регрессий, стоит добавлять в общую систему не все уравнения, а только те, правые части которых содержат θ_i^j при степенях x не меньше 2. Это гарантирует высокую точность вектора правых частей и точное нахождение оценок неизвестных коэффициентов многомерной регрессии. Если при таком подходе в системе не набирается r линейно независимых строк, то в нее можно включить уравнения, в правых частях которых находится θ_i^j при x первой степени. В системе не должно быть уравнений, соответствующих свободному члену в построенных одномерных регрессиях, так как оценка свободного члена одномерной регрессии имеет большую дисперсию и ухудшит качество оценки коэффициентов многомерной регрессии. Свободный член исходной модели находиться при помощи оценок остальных коэффициентов, как среднее арифметическое отклонений значений регрессионной модели без него, от входных данных.

2. На практике возникают ситуации, когда число линейно независимых строк меньше количества неизвестных коэффициентов. В таком случае мы можем найти частичное решение, выбрав из полученной системы новую, невырожденную подсистему с меньшим количеством переменных. Здесь удобно использование человеко-машинной процедуры, так как человек может увидеть и выделить такие подсистемы. Нахождение некоторых коэффициентов начальной модели гораздо упрощает ее и позволяет решить меньшую задачу регрессии любым другим способом.

Поскольку точность нахождения оценок коэффициентов при многомерном регрессионном анализе зависит от качества одномерной регрессии, то далее будут рассмотрены методы улучшения одномерной регрессии.

Регрессия при данных с повторяющимися аргументами

В условиях задач, встречающихся на практике исследователю нужно проводить большое количество экспериментов на небольшом интервале значений аргументов. При равномерном распределении значений аргументов возникают некоторые неудобства, связанные с тем, что каждый следующий эксперимент нужно проводить, изменив значение аргумента на некоторое небольшое число Δx, что трудно на практике, при исследовании физических процессов. В этом случае на вход удобно подавать повторяющуюся последовательность

$$x_1,\cdots,x_{r+p},x_1,\cdots,x_{r+p},\cdots,$$

где $p\geq 1$, r – степень одномерного полинома, который задается избыточным описанием.

Докажем, что оценки коэффициентов $\hat{\theta}_j, j=\overline{0,r}$ не изменятся при усреднении результатов экспериментов, при выполнении одномерного полиномиального регрессионного анализа методом из [1]. Введем обозначения:

$$X=(x_1,x_2,...,x_n)$$

$$X'=(x_{11},x_{21},...x_{l1},...,x_{1n},x_{2n},...,x_{ln})$$

В X' находиться l копий X, т.е. $x_{ki}=x_i;\forall k=\overline{1,l};\forall i=\overline{1,n}$.

$$Y'=(y_{11},y_{21},...y_{l1},...,y_{1n},y_{2n},...,y_{ln})$$

$$Y=(\frac{\sum_{k=1}^{l}y_{k1}}{l},\frac{\sum_{k=1}^{l}y_{k2}}{l},...,\frac{\sum_{k=1}^{l}y_{kn}}{l})$$

Другими словами, вместо двух векторов X′ и Y′, размерностью $n\times l$ каждый, мы подаем X, Y, размерность которых – n, где n – количество каждого из различных значений аргумента.

$Q_j, j=\overline{0,r}$ – ортогональные полиномы, построенные на выборке X

$Q_j', j=\overline{0,r}$ – ортогональные полиномы, построенные на выборке X′

Методом математической индукции покажем, что

$$Q_j'(x)=\frac{Q_j(x)}{\sqrt{l}}, j=\overline{0,r} \quad (7)$$

При $j=0$ имеем:

$$Q_0'(x)=\frac{1}{\sqrt{n\times l}}=\frac{Q_0(x)}{\sqrt{l}}$$

При $j=1$ имеем:

$$Q_1'(x)=\frac{x-\overline{X}'}{\sqrt{\sum_{i=1}^{n\times l}(x_i-\overline{X}')^2}}=$$

$$=\left[\begin{array}{l}\overline{X}=\overline{X}';\\ \sum_{i=1}^{n\times l}(x_i-\overline{X}')^2=\sum_{k=1}^{l}\sum_{i=1}^{n}(x_i-\overline{X}')^2=l\times\sum_{i=1}^{n}(x_i-\overline{X}')^2\end{array}\right]=$$

$$=-\frac{(x-\overline{X}')}{\sqrt{l\times\sum_{i=1}^{n}(x_i-\overline{X}')^2}}=\frac{(x-\overline{X}')}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{X}')^2}}\times\frac{1}{\sqrt{l}}=\frac{Q_1(x)}{\sqrt{l}}$$

Пусть (7) выполняется для $Q_{j-1}'(x)$ и $Q_{j-2}'(x)$, тогда:

$$Q_{j-1}'(x)=\frac{Q_{j-1}(x)}{\sqrt{l}}$$

$$Q_{j-2}'(x)=\frac{Q_{j-2}(x)}{\sqrt{l}}$$

Покажем теперь, что (7) выполняется для любого $Q_j'(x)$.

Вначале докажем, что $\alpha'=\alpha$, (т.е. α не зависит от того, используем ли мы X или X'):

$$\alpha'=\sum_{i=1}^{n\times l}x_iQ_{j-1}'^2(x_i)=\sum_{i=1}^{n\times l}x_i\frac{Q_{j-1}{}^2(x_i)}{l}=\sum_{k=1}^{l}\sum_{i=1}^{n}x_i\frac{Q_{j-1}{}^2(x_i)}{l}$$

Так как каждый элемент суммы по индексу k не зависит от k, тогда мы можем заменить сумму на произведение:

$$\sum_{k=1}^{l}\sum_{i=1}^{n}x_i\frac{Q_{j-1}{}^2(x_i)}{l}=l\times\sum_{i=1}^{n}x_i\frac{Q_{j-1}{}^2(x_i)}{l}=$$

$$=\sum_{i=1}^{n}x_iQ_{j-1}{}^2(x_i)=\alpha$$

Аналогично покажем, что $\beta'=\beta$:

$$\beta' = \sum_{i=1}^{n\times l} x_i Q'_{j-1}(x_i) Q'_{j-2}(x_i) = \sum_{k=1}^{l}\sum_{i=1}^{n} x_i \frac{Q_{j-1}(x_i)}{\sqrt{l}} \times \frac{Q_{j-2}(x_i)}{\sqrt{l}} =$$

$$= l \times \sum_{i=1}^{n} x_i \frac{Q_{j-1}(x_i) Q_{j-2}(x_i)}{l} = \sum_{i=1}^{n} x_i Q_{j-1}(x_i) Q_{j-2}(x_i) = \beta$$

Аналогично, покажем что $\lambda' = \lambda$:

$$\lambda' = \sqrt{\sum_{i=1}^{n\times l} (x_i Q'_{j-1}(x_i) - \lambda' Q'_{j-1}(x_i) - \beta' Q'_{j-2}(x_i))^2} =$$

$$= \sqrt{\sum_{k=1}^{l}\sum_{i=1}^{n} (\frac{x_i Q_{j-1}(x_i)}{\sqrt{l}} - \frac{\lambda Q_{j-1}(x_i)}{\sqrt{l}} - \frac{\beta Q_{j-2}(x_i)}{\sqrt{l}})^2} =$$

$$= \sqrt{l \times \sum_{i=1}^{n} \frac{(x_i Q_{j-1}(x_i) - \lambda Q_{j-1}(x_i) - \beta Q_{j-2}(x_i))^2}{l}} =$$

$$= \sqrt{\sum_{i=1}^{n} (x_i Q_{j-1}(x_i) - \lambda Q_{j-1}(x_i) - \beta Q_{j-2}(x_i))^2} = \lambda$$

Теперь подставив значения $\lambda', \beta', \alpha'$, докажем что выполняется (7):

$$Q'_j(x) = \frac{xQ'_{j-1} - \alpha' Q'_{j-1} - \beta' Q'_{j-2}(x)}{\lambda'}$$

$$\frac{x\frac{Q_{j-1}}{\sqrt{l}} - \alpha\frac{Q_{j-1}}{\sqrt{l}} - \beta\frac{Q_{j-2}(x)}{\sqrt{l}}}{\lambda} =$$

$$= \frac{xQ_{j-1} - \alpha Q_{j-1} - \beta Q_{j-2}(x)}{\lambda} \times \frac{1}{\sqrt{l}} = \frac{Q_j(x)}{\sqrt{l}}$$

Мы показали, что при переходе от решения задачи регрессии на X' к решению на X, мы должны заменить все значения полиномов по формуле (7).

Покажем, что оценка регрессионного полинома не изменится, при переходе на данные, с усредненными результатами экспериментов.

Вначале, найдем зависимость значений оценок весовых коэффициентов $\hat{w}_j, j = \overline{1,n}$, полученных при усреднении результатов экспериментов и оценок полученных на данных без усреднения:

$$\hat{\omega}'_j = \sum_{i=1}^{n}\sum_{k=1}^{l} y_{ki} Q'_j(x_{ki}) =$$

$$= \left[Q'_j(x_{k_1 i}) = Q'_j(x_{k_2 i}), \forall k_1, k_2 \in \{1,...,l\} \right] =$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{l} y_{ki} Q'_j(x_i) = \sum_{i=1}^{n}\sum_{k=1}^{l} y_{ki} \frac{Q_j(x_i)}{\sqrt{l}} =$$

$$= [\times \frac{l}{l}] = \sum_{i=1}^{n} \left(\frac{\sum_{k=1}^{l} y_{ki}}{l} \right) \times l \times \frac{Q_j(x_i)}{\sqrt{l}} = \hat{\omega}_j \times \sqrt{l}$$

Получили, что:

$$\hat{\omega}'_j = \omega_j \times \sqrt{l}, \forall j = \overline{1,n}$$

Теперь покажем, что $\hat{\theta}_j' = \hat{\theta}_j, \forall j = \overline{0,r}$:

$$\hat{\theta}_j' = \hat{w}_r' q'_{rj} + \cdots + \hat{w}'_j q'_{jj} = \hat{w}_r \sqrt{l} \times \frac{q_{rj}}{\sqrt{l}} + \cdots + \hat{w}_j \sqrt{l} \times \frac{q_{jj}}{\sqrt{l}} = \hat{\theta}_j, \forall j = \overline{0,r} \qquad (8)$$

Как видно из (8), регрессионный анализ над данными после усреднения дает такой же результат, как и анализ над всем исходным массивом данных.

Также, для более точного нахождения оценок коэффициентов рекомендуется проводить активный эксперимент на более широком интервале, и чтобы центр интервала был как можно ближе к точке 0 [3].

Вывод

В статье приведен конструктивный метод восстановления многомерной полиномиальной регрессии, представленной избыточным описанием, с использованием ограниченного активного эксперимента. Он базируется на выполнении ряда одномерных полиномиальных регрессий и решении системы линейных уравнений. Так как, этот метод не опирается на оценки коэффициентов при низких степенях аргумента, в отличии от метода из [1], он лучше восстанавливает истинную закономерность. Приведенный метод используется в случаи, когда входные аргументы изменяются одинаково.

В исследованиях с повторяющимися сериями экспериментов при одинаковых значениях аргументов, вместо всех данных на вход алго-

ритма регрессии, можно подавать гораздо меньше точек, усредняя значения y при одинаковом x. При этом оценки коэффициентов не изменятся. Это дает значительный выигрыш в количестве вычислений.

Список литературы

1. Згуровский М. З. Принятие решений в сетевых системах с ограниченными ресурсами [Текст] : [монография] / М. З. Згуровский, А. А. Павлов ; Нац. техн. ун-т "Киев. политехн. ин-т". – К. : Наукова думка, 2010. – 575 с. : рис., табл. – Бібліогр.: С. 560-569.
2. Meyer C. D. Matrix analysis and applied linear algebra / Carl Dean Meyer. – Philadelphia : Society for Industrial and Applied Mathematics, 2000. – 718 с.
3. Павлов А. А. Рекомендации по выбору зоны проведения активного эксперимента для одномерного полиномиального регрессионного анализа / А. А. Павлов, В. В. Калашник. // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка. – 2014. – №60. – С. 41-45.