

Поиск параметрической регрессионной модели в индуктивно заданном множестве*

В. В. СТРИЖОВ

Вычислительный центр имени А. А. Дородницына РАН

e-mail: `strijov@ccas.ru`

Описана процедура поиска параметрической регрессионной модели в классе моделей, определенном суперпозициями гладких функций из заданного множества. Для поиска используются оценки плотности распределения параметров элементов моделей. Параметры моделей оцениваются с помощью методов нелинейной оптимизации. Для иллюстрации приведена задача о моделировании изменения давления в камере внутреннего сгорания дизельного двигателя.

1. Введение

Проблема отыскания оптимальной параметрической регрессионной модели имеет большую историю, однако остается одной из самых актуальных в области распознавания образов. А.Г. Ивахненко, еще в 1968 году, предложил метод группового учета аргументов, МГУА [1]. Согласно этому методу модель, доставляющая наилучшее приближение, отыскивается во множестве последовательно порождаемых моделей. В частности, для построения моделей как суперпозиций функций, использовались полиномиальные функции, ряды Фурье и некоторые другие функции. А.Г. Ивахненко и его ученики создали ряд алгоритмов синтеза моделей и предложили методы оценки качества моделей.

При порождении конкурирующих моделей появляется задача определения значимости элементов модели. В работе К. Бишоп [2] предложен метод анализа распределения параметров однослойных нейронных сетей посредством гиперпараметров, то есть параметров распределения параметров аппроксимирующих функций. Для каждого элемента сети оценивается плотность Гауссовского распределения его параметров и делается вывод о том, насколько информативен данный элемент исследуемой регрессионной модели.

Для модификации моделей Ле Кюн предложил метод, называемый методом оптимального отсечения (optimal brain damage) [3]. Этот метод состоит в исключении некоторых, наименее информативных, элементов регрессионной модели с тем условием, что при таком исключении качество аппроксимации уменьшается незначительно. При исключении отдельных элементов модели становится возможным оценить вклад этих элементов по значениям заданной функции качества аппроксимации.

Проблема сравнения и выбора регрессионных моделей получила новое развитие после ряда публикаций Д. МакКая [4–6], предложившего при выборе модели из заданного множества использовать не информационные критерии, например AIC — Akaike Information Criterion, а двухуровневый Байесовский вывод и правило Оккама. На первом уровне вывода вычисляются плотности вероятностей распределения параметров каждой модели из за-

*Работа выполнена при финансовой поддержке РФФИ (грант № 04-01-00401).

данного множества. На втором уровне вывода вычисляется правдоподобие моделей. Правило Оккама состоит в том, что вероятность выбора более сложной модели меньше, чем вероятность выбора более простой модели при сравнимом значении функции качества аппроксимации.

Метод, предлагаемый в данной работе, заключается в следующем. Поиск моделей выполняется по итерационной схеме “порождение-выбор” в соответствии с определенными правилами порождения моделей и критерием выбора моделей. Последовательно порождаются наборы конкурирующих моделей. Каждая модель в наборе является суперпозицией элементов заданного множества гладких параметрических функций. После построения модели, каждому элементу суперпозиции ставится в соответствие гиперпараметр. Параметры и гиперпараметры модели последовательно настраиваются. Из набора выбираются наилучшие модели для последующей модификации. При модификации моделей, по значениям гиперпараметров делаются выводы о целесообразности включения того или иного элемента в модель следующего порождаемого набора.

Поставим задачу нахождения регрессионной модели нескольких свободных переменных следующим образом. Задана выборка — множество $\{\mathbf{x}_1, \dots, \mathbf{x}_N | \mathbf{x} \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1, \dots, y_N | y \in \mathbb{R}\}$ соответствующих им значений зависимой переменной. Обозначим оба эти множества как множество исходных данных D .

Также задано множество $G = \{g | g : \mathbb{R} \times \dots \times \mathbb{R} \rightarrow \mathbb{R}\}$ гладких параметрических функций $g = g(\mathbf{b}, \cdot, \cdot, \dots, \cdot)$. Первый аргумент функции g — вектор-строка параметров \mathbf{b} , последующие — переменные из множества действительных чисел, рассматриваемые как элементы вектора свободных переменных. Рассмотрим произвольную суперпозицию, состоящую из не более чем r функций g . Эта суперпозиция задает параметрическую регрессионную модель $f = f(\mathbf{w}, \mathbf{x})$. Модель f зависит от вектора свободных переменных \mathbf{x} и от вектора параметров \mathbf{w} . Вектор $\mathbf{w} \in \mathbb{R}^W$ состоит из присоединенных векторов-параметров функций g_1, \dots, g_r , то есть, $\mathbf{w} = \mathbf{b}_1 \dot{\vdots} \mathbf{b}_2 \dot{\vdots} \dots \dot{\vdots} \mathbf{b}_r$, где $\dot{\vdots}$ — знак присоединения векторов. Обозначим $\Phi = \{f_i\}$ — множество всех суперпозиций, индуктивно порожденное элементами множества G .

Требуется найти такую модель f_i , которая доставляет максимум функционала $p(\mathbf{w} | D, \alpha, \beta, f_i)$. Этот функционал, определяемый далее, включает искомую модель $f_i(\mathbf{w}, \mathbf{x})$ и ее дополнительные параметры α и β .

2. Выбор регрессионных моделей и гипотеза порождения данных

Общий подход к сравнению нелинейных моделей заключается в следующем. Рассмотрим набор конкурирующих моделей f_1, \dots, f_M . Априорная вероятность модели f_i определена как $P(f_i)$. При появлении данных D апостериорная вероятность модели $P(f_i | D)$ может быть найдена по теореме Байеса,

$$P(f_i | D) = \frac{P(f_i)p(D|f_i)}{\sum_{i=1}^M p(D|f_i)P(f_i)},$$

где $p(D|f_i)$ — функция соответствия модели данным. Знаменатель дроби обеспечивает выполнение условия $\sum_{i=1}^M P(f_i | D) = 1$.

Вероятности моделей f_1 и f_2 , параметры которых идентифицированы по данным D , сравнимы как

$$\frac{P(f_1|D)}{P(f_2|D)} = \frac{P(f_1)p(D|f_1)}{P(f_2)p(D|f_2)}. \quad (1)$$

Отношение $\frac{p(D|f_1)}{p(D|f_2)}$ — есть отношение правдоподобия моделей. Отношение $\frac{P(f_1)}{P(f_2)}$ является априорной оценкой предпочтения одной модели другой. При моделировании отдается предпочтение наиболее простым и устойчивым моделям. Если априорные оценки $P(f_i)$ моделей одинаковы, то есть, нет причины предпочитать одну модель другой, то их необходимо сравнивать по значениям $p(D|f_i)$:

$$p(D|f_i) = \int p(D|\mathbf{w}, f_i)p(\mathbf{w}|f_i)d\mathbf{w}.$$

Апостериорная плотность распределения параметров \mathbf{w} функции f_i при заданной выборке D равна

$$p(\mathbf{w}|D, f_i) = \frac{p(D|\mathbf{w}, f_i)p(\mathbf{w}|f_i)}{p(D|f_i)}, \quad (2)$$

где $p(\mathbf{w}|f_i)$ — априорно заданная плотность вероятности параметров начального приближения, $p(D|\mathbf{w}, f_i)$ — функция правдоподобия параметров модели, а знаменатель $p(D|f_i)$ обеспечивает выполнение условия $\int p(\mathbf{w}|D, f_i)d\mathbf{w} = 1$. Он задан интегралом в пространстве параметров $\int p(\mathbf{w}'|D, f_i)p(\mathbf{w}'|f_i)d\mathbf{w}'$. Формулы (2) и (1) называются формулами Байесовского вывода первого и второго уровня.

Рассмотрим регрессию $y = f_i(\mathbf{b}, \mathbf{x}) + \nu$ с аддитивным Гауссовским шумом с дисперсией σ_ν и с нулевым матожиданием. Тогда плотность вероятности появления данных

$$p(y|x, \mathbf{w}, \beta, f_i) \triangleq p(D|\mathbf{w}, \beta, f) = \frac{\exp(-\beta E_D(D|\mathbf{w}, f_i))}{Z_D(\beta)},$$

где $\beta = \frac{1}{\sigma_\nu^2}$. Нормирующий множитель $Z_D(\beta)$ задан выражением

$$Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}} \quad (3)$$

и взвешенный функционал ошибки в пространстве данных

$$\beta E_D = \frac{\beta}{2} \sum_{n=1}^N (f_i(\mathbf{x}_n) - y_n)^2. \quad (4)$$

Введем регуляризующий параметр α , который отвечает за то, насколько хорошо модель должна соответствовать зашумленным данным. Функция плотности вероятности параметров с заданным гиперпараметром α имеет вид

$$p(\mathbf{w}|\alpha, f_i) = \frac{\exp(-\alpha E_W(\mathbf{w}|f_i))}{Z_W(\alpha)},$$

где α — обратная дисперсии распределения параметров, $\alpha = \sigma_{\mathbf{w}}^{-2}$, а нормирующая константа Z_W определена дисперсией распределения параметров как

$$Z_W(\alpha) = \left(\frac{2\pi}{\alpha}\right)^{\frac{w}{2}}. \quad (5)$$

Требование к малым значениям параметров [7] предполагает Гауссовское априорное распределение с нулевым средним:

$$p(\mathbf{w}) = \frac{1}{Z_W} \exp\left(-\frac{\alpha}{2} \|\mathbf{w}\|^2\right).$$

Так как переменные α и β являются параметрами распределения параметров модели, в дальнейшем будем называть их гиперпараметрами. Исключая нормирующую константу Z_W , которая не зависит от параметров \mathbf{w} и логарифмируя, получаем

$$\alpha E_W = \frac{\alpha}{2} \|\mathbf{w}\|^2. \quad (6)$$

Эта ошибка регуляризирует параметры, начисляя штраф за их чрезмерно большие значения.

При заданных значениях гиперпараметров α и β выражение (2) для фиксированной функции f_i будет иметь вид

$$p(\mathbf{w}|D, \alpha, \beta) = \frac{p(D|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)}{p(D|\alpha, \beta)}.$$

Записывая функцию ошибки в виде $S(\mathbf{w}) = \alpha E_W + \beta E_D$, получаем

$$p(\mathbf{w}|D, \alpha, \beta, f_i) = \frac{\exp(-S(\mathbf{w}|f_i))}{Z_S(\alpha, \beta)}, \quad (7)$$

где Z_S — нормирующий множитель.

3. Нахождение параметров модели

Рассмотрим итеративный алгоритм для определения оптимальных параметров \mathbf{w} и гиперпараметров α, β при заданной модели f_i . Корректный подход заключается в интегрировании всех неизвестных параметров и гиперпараметров. Апостериорное распределение параметров определяется как

$$p(\mathbf{w}|D) = \iint p(\mathbf{w}, \alpha, \beta|D) d\alpha d\beta = \iint p(\mathbf{w}|\alpha, \beta, D) p(\alpha, \beta|D) d\alpha d\beta, \quad (8)$$

что требует выполнить интегрирование апостериорного распределения параметров $p(\mathbf{w}|\alpha, \beta, D)$ по пространству, размерность которого равна количеству параметров. Вычислительная сложность этого интегрирования весьма велика. Интеграл может быть упрощен при подходящем выборе начальных значений гиперпараметров.

Приближение интеграла заключается в том, что апостериорная плотность распределения гиперпараметров $p(\alpha, \beta|D)$ имеет выраженный пик в окрестности наиболее правдоподобных значений гиперпараметров $\alpha^{\text{MP}}, \beta^{\text{MP}}$. Это приближение известно как аппроксимация Лапласа [8]. При таком допущении интеграл (8) упрощается до

$$p(\mathbf{w}|D) \approx p(\mathbf{w}|\alpha^{\text{MP}}, \beta^{\text{MP}}, D) \iint p(\alpha, \beta|D) d\alpha d\beta \approx p(\mathbf{w}|\alpha^{\text{MP}}, \beta^{\text{MP}}, D).$$

Необходимо найти значения гиперпараметров, которые оптимизируют апостериорную плотность вероятности параметров, а затем выполнить все остальные расчеты, включающие $p(\mathbf{w}|D)$ при фиксированных значениях гиперпараметров.

Для нахождения функционала $p(\mathbf{w}|\alpha, \beta, D)$, который использует апостериорное распределение параметров, рассмотрим аппроксимацию ошибки $S(\mathbf{w})$ на основе рядов Тейлора второго порядка:

$$S(\mathbf{w}) \approx S(\mathbf{w}^{\text{MP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^{\text{MP}})^T A(\mathbf{w} - \mathbf{w}^{\text{MP}}). \quad (9)$$

В выражении (9) нет слагаемого первого порядка, так как предполагается, что \mathbf{w}^{MP} определяет локальный минимум функции ошибки, то есть

$$\frac{\partial S(\mathbf{w}^{\text{MP}})}{\partial w_\xi} = 0$$

для всех значений ξ . Матрица A — это матрица Гессе функции ошибок:

$$A = \nabla^2 S(\mathbf{w}^{\text{MP}}) = \beta \nabla^2 E_D(\mathbf{w}^{\text{MP}}) + \alpha I.$$

Обозначим первое слагаемое правой части через H , тогда $A = H + \alpha I$.

Подставив полученное приближенное значение $S(\mathbf{w})$ в (7) и обозначив $\Delta \mathbf{w} = \mathbf{w} - \mathbf{w}^{\text{MP}}$, получим

$$p(\mathbf{w}|\alpha, \beta, D) = \frac{1}{\hat{Z}_S} \exp\left(-S(\mathbf{w}^{\text{MP}}) - \frac{1}{2}\Delta \mathbf{w}^T A \Delta \mathbf{w}\right),$$

Оценим нормирующую константу \hat{Z}_S , необходимую для аппроксимации кривой Гаусса, как

$$\hat{Z}_S = \exp(-S(\mathbf{w}^{\text{MP}}))(2\pi)^{\frac{W}{2}} (\det A)^{-\frac{1}{2}}. \quad (10)$$

Максимизируем функцию $p(D|\alpha, \beta)$, изменяя значения гиперпараметров α и β . Это можно выполнить, интегрируя функцию плотности вероятности данных по пространству параметров \mathbf{w} :

$$p(D|\alpha, \beta) = \int p(D|\mathbf{w}, \alpha, \beta)p(\mathbf{w}|\alpha, \beta)d\mathbf{w} = \int p(D|\mathbf{w}, \alpha, \beta)p(\mathbf{w}|\alpha)d\mathbf{w}, \quad (11)$$

где второй интеграл справедлив по причине того, что распределение параметров не зависит от дисперсии шума в силу гипотезы о Гауссовском распределении шума. Для упрощения вычислений мы допускаем, что распределение $p(\alpha, \beta)$ является равномерным.

Используя (4), (6), запишем (11) в виде

$$p(D|\alpha, \beta) = \frac{1}{Z_D(\beta)} \frac{1}{Z_D(\alpha)} \int \exp(-S(\mathbf{w}))d\mathbf{w}.$$

Из (3), (5), (10) и предыдущего выражения получим

$$\ln p(D|\alpha, \beta) = -\alpha E_W^{\text{MP}} - \beta E_D^{\text{MP}} - \frac{1}{2} \ln |A| + \frac{W}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln (2\pi). \quad (12)$$

Для того, чтобы оптимизировать это выражение относительно α , найдем производную

$$\frac{d}{d\alpha} \ln |A| = \frac{d}{d\alpha} \ln \left(\prod_{j=1}^W \lambda_j + \alpha \right) = \frac{d}{d\alpha} \sum_{j=1}^W \ln(\lambda_j + \alpha) = \sum_{j=1}^W \frac{1}{\lambda_j + \alpha} = \text{tr}(A^{-1}). \quad (13)$$

В этом выражении $\lambda_1, \dots, \lambda_W$ — собственные значения матрицы H . Так как функция ошибки на данных не является квадратичной функцией параметров, как при линейной или

RVF регрессии, то непосредственно оптимизировать величину α невозможно, гессиан не является константой, а зависит от параметров \mathbf{w} . Так как мы принимаем $A = H + \alpha I$ для вектора \mathbf{w}^{MP} , который зависит от выбора α , то собственные значения H косвенным образом зависят от α . Таким образом, формула (13) игнорирует параметры модели.

С использованием этого приближения, производная (12) с учетом α равна

$$\ln p(D|\alpha, \beta) = -E_W^{\text{MP}} - \frac{1}{2} \sum_{j=1}^W \frac{1}{\lambda_j + \alpha} + \frac{W}{2\alpha}.$$

Приравнивая последнее выражение к нулю и преобразовывая, получаем выражение для α

$$2\alpha E_W^{\text{MP}} = W - \sum_{j=1}^W \frac{\alpha}{\lambda_j + \alpha}. \quad (14)$$

Обозначим вычитаемое правой части через γ

$$\gamma = \sum_{j=1}^W \frac{\alpha}{\lambda_j + \alpha}.$$

Те компоненты суммы, в которых $\lambda_j \gg \alpha$ приносят вклад, близкий к единице, а те компоненты суммы, в которых $0 < \lambda_j \ll \alpha$, приносят вклад, близкий к нулю. Таким образом γ может быть интерпретирована как мера числа хорошо обусловленных параметров модели.

Для нахождения гиперпараметра β рассмотрим задачу оптимизации (12). Обозначим через μ_j собственное значение матрицы $\nabla^2 E_D$. Так как $H = \beta \nabla^2 E_D$, то $\lambda_j = \beta \mu_j$ и следовательно,

$$\frac{d\lambda_j}{d\beta} = \mu_j = \frac{\lambda_j}{\beta}.$$

Отсюда,

$$\frac{d}{d\beta} \ln |A| = \frac{d}{d\beta} \sum_{j=1}^W \ln(\lambda_j + \alpha) = \frac{1}{\beta} \sum_{j=1}^W \frac{\lambda_j}{\lambda_j + \alpha}.$$

Дифференцируя, как и в случае нахождения α , мы находим, что оптимальное значение β определено как

$$2\beta E_D^{\text{MP}} = N - \sum_{j=1}^W \frac{\lambda_j}{\lambda_j + \alpha} = N - \gamma. \quad (15)$$

Способ вычисления оптимальных значений гиперпараметров α и β описан в следующем разделе.

4. Процедура поиска оптимальной модели

Поиск оптимальной модели происходит на множестве порождаемых моделей на каждой итерации алгоритма. Перед работой алгоритма заданы множество измеряемых данных D и множество гладких функций G . Задан начальный набор конкурирующих моделей, $F_0 = \{f_1, \dots, f_M | f \in \Phi\}$, в котором каждая модель f_i есть суперпозиция функций $\{g_{ij}\}_{j=1}^{r_i}$.

Каждой функции g_{ij} — элементу модели f_i ставится в соответствие гиперпараметр α_{ij} , характеризующий начальную плотность распределения вектора параметров \mathbf{b}_{ij} этой функции. Каждой модели f_i поставлен в соответствие гиперпараметр β_i начального приближения. Параметры начального приближения для i -й модели назначаются исходя из априорного распределения данных, определяемых значением β_i . Далее выполняется последовательность шагов, приведенных ниже, которые повторяются заданное количество раз.

1. Методом сопряженных градиентов [9] минимизируются штрафные функции $S_i(\mathbf{w})$ для каждой модели $f_i, i = 1, \dots, M$. Отыскиваются параметры моделей \mathbf{w}_i^{MP} .

2. После нахождения параметров \mathbf{w}_i^{MP} определяются, исходя из (14) и (15), новые значения гиперпараметров α_{ij}^{new} и β_i^{new} . Гиперпараметр β_i функции f_i вычисляется для всего набора данных и равен

$$\beta_i^{\text{new}} = \frac{N - \gamma_i}{E_D(f_i)}.$$

Гиперпараметр α_{ij} вычисляется для каждой функции g_{ij} из суперпозиции f_i и равен

$$\alpha_{ij}^{\text{new}} = \frac{W - \gamma_i}{E_W(\mathbf{b}_{ij})}.$$

Здесь значения функционалов γ_i и $E_W(\mathbf{b}_{ij})$ вычисляется только для подмножества тех параметров \mathbf{b}_{ij} из множества \mathbf{w}_i , которые являются параметрами функции g_{ij} . Изменение гиперпараметров повторяется итерационно до тех пор пока локальный минимум $S_i(\mathbf{w})$ не останется постоянным.

3. Заданы следующие правила построения производных моделей f'_1, \dots, f'_M . Для каждой модели f_i строится производная модель f'_i . В f_i выбирается функция g_{ij} с наименьшим значением α_{ij} . Выбирается произвольная модель f_ξ из $F_0 \setminus \{f_i\}$ и ее произвольная функция $g_{\xi\zeta}$. Модель f' порождается из модели f путем замещения функции g_{ij} с ее аргументами на функцию $g_{\xi\zeta}$ с ее аргументами.

4. С заданной вероятностью η каждая модель f'_i подвергается изменениям. В изменяемой модели выбирается j -тая функция, причем закон распределения вероятности выбора

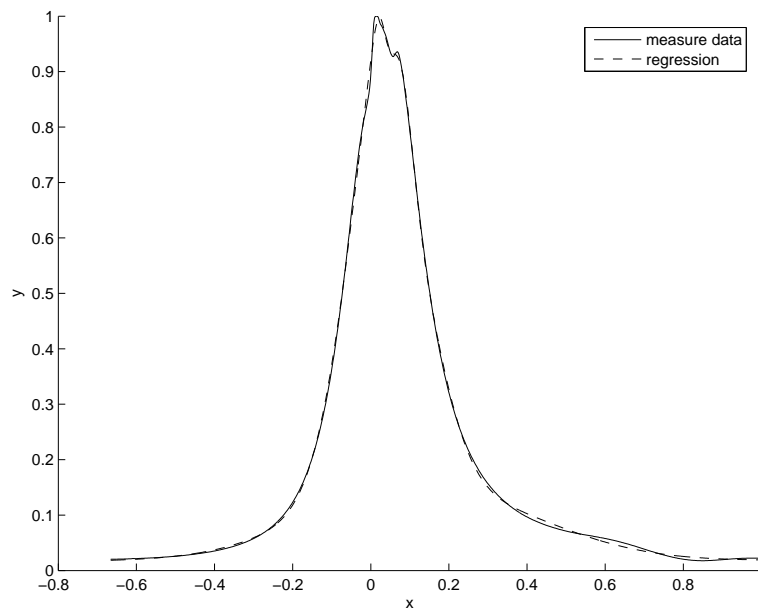


Рис. 1. Исходная выборка и восстановленная выборка, полученная моделью № 2

функции $p(j)$ задан. Из множества G случайным образом выбирается функция g' и замещает функцию g_j . Гиперпараметр α_{ij} этой функции определяется как $\max(\alpha_{ij})$. Вектор параметров этой функции \mathbf{b}_{ij} равен нулю или назначается при задании G^j .

5. При выборе моделей из объединенного множества родительских и порожденных моделей в соответствии с критерием $S(\mathbf{w})$ выбираются M наилучших, которые используются в дальнейших итерациях.

5. Численный эксперимент

Ниже описывается пример построения регрессионной модели. Объектом моделирования является кривая одной свободной переменной, представленная набором измерений давления в камере внутреннего сгорания дизельного двигателя. На рис. 1 сплошной кривой показаны исходные данные, пунктиром показаны значения модели № 2. По оси абсцисс отложено значение свободной переменной, по оси ординат — значение зависимой переменной. Выборка, представленная данной кривой, содержит четыре тысячи отсчетов. Для верификации полученных моделей использовалось 118 выборок.

Экспертами было задано множество базовых функций G из элементов которого порождаются регрессионные модели. Список функций приведен в таблице 1. Множество F_0 моделей начального приближения также было задано экспертами.

№	Функция	Описание	Параметры
Функции двух переменных аргументов, $g(\mathbf{b}, x_1, x_2)$			
1	plus	$y = x_1 + x_2$	—
2	times	$y = x_1 x_2$	—
3	divide	$y = x_1 / x_2$	—
Функции одного переменного аргумента, $g(\mathbf{b}, x_1)$			
4	multiply	$y = ax$	a
5	add	$y = x + a$	a
6	gaussian	$y = \frac{\lambda}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) + a$	λ, σ, ξ, a
7	linear	$y = ax + b$	a, b
8	parabolic	$y = ax^2 + bx + c$	a, b, c
9	cubic	$y = ax^3 + bx^2 + cx + d$	a, b, c, d
10	logsig	$y = \frac{\lambda}{1 + \exp(-\sigma(x-\xi))} + a$	λ, σ, ξ, a

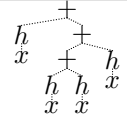
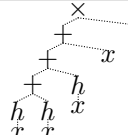
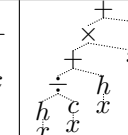
Т а б л и ц а 1. Множество G базовых функций

Выбор моделей производился из более тысячи порожденных моделей. В таблице 2 приведены три модели, полученные в результате работы алгоритма. Качество моделей оценивалось по ошибкам ρ_1, ρ_2 и числу параметров в векторе параметров \mathbf{w} . Значения ошибок каждой модели получены путем усреднения результатов оптимально настроенной модели по 118 выборкам. Ошибка ρ_1 — среднеквадратичная относительная ошибка

$$\rho_1 = \sqrt{\frac{1}{N} \sum_{i=1}^n \left(\frac{y_i - f(\mathbf{x}_i)}{\max(y_i)} \right)^2},$$

ошибка ρ_2 — максимальная относительная ошибка

$$\rho_2 = \max_{i=1, \dots, N} \frac{|y_i - f(\mathbf{x}_i)|}{\max(y_i)}.$$

№ модели	1	2	3
Ошибка ρ_1	0.0034	0.0037	0.0035
Ошибка ρ_2	0.0421	0.0325	0.00338
Число параметров	16	16	16
Описание			

Легенда: h — gaussian, c — cubic, l — linear,
+ — plus, × — times, ÷ — divide

Т а б л и ц а 2. Описание выбранных моделей

В строке «Описание» таблицы 2 показана структура модели в виде дерева. В качестве примера рассмотрим модель №2. Эта модель состоит из суперпозиции восьми функций $f_2 = g_1(g_2(g_3(g_4(g_5(x), g_6(x)), g_7(x)), x), g_8(x))$. Функции $g_1 = \times(\emptyset, \cdot, \cdot)$ и $g_2, \dots, g_4 = +(\emptyset, \cdot, \cdot)$, сложения и умножения, имеют первым аргументом пустой вектор параметров; $g_5, \dots, g_7 = h(\mathbf{b}_i, \cdot)$, $i = 1, \dots, 3$, и $g_8 = l(\mathbf{b}_4, \cdot)$. Функции $h = \frac{\lambda_i}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{(x-\xi_i)^2}{2\sigma_i^2}\right)$ имеют векторы параметров $\mathbf{b}_i = \langle \lambda_i, \mu_i, \sigma_i, a_i \rangle$, и функция $l = (ax + b)$ имеет вектор параметров $\mathbf{b}_4 = \langle a, b \rangle$.

Модель f_2 можно переписать в виде $f(\mathbf{w}, \mathbf{x}) = l(\mathbf{b}_4, x)^{-1} \times (x + \sum_{i=1}^3 h(\mathbf{b}_i, x))$, где $\mathbf{x} = x$ и $\mathbf{w} = \mathbf{b}_1:\mathbf{b}_2:\mathbf{b}_3:\mathbf{b}_4$. Развернутый вид модели:

$$y = (ax + b)^{-1} \left(x + \sum_{i=1}^3 \frac{\lambda_i}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{(x-\xi_i)^2}{2\sigma_i^2}\right) + a_i \right).$$

Модель f_2 была использована экспертами для анализа и прогноза концентрации кислорода в выхлопных газах дизельного двигателя.

Заключение

Универсальные регрессионные модели, например, нейронные сети или радиальные базисные функции, при обработке результатов измерений часто имеют большое число параметров и получаются переобученными. Для достижения результатов в построении несложных и достаточно точных моделей была поставлена задача о выборе регрессионной модели, которая состоит из суперпозиции гладких функций.

Для выбора наилучшей модели из индуктивно заданного множества использован двухуровневый Байесовский вывод. В связи со сложностью вычисления значений интегралов вывода, были предложены процедуры приближения, которые позволяют отыскивать адекватные модели за приемлемое время вычислений.

Предложенная процедура выбора регрессионных моделей использует гиперпараметры, поставленные в соответствие элементам модели. Эти гиперпараметры указывают на важность элементов модели. На основе информации о важности элементов итеративно порождаются новые модели. Сложность моделей ограничивается автоматически при сравнении моделей.

Описанный метод был протестирован на задаче по аппроксимации кривой, полученной в результате измерений давления в камере внутреннего сгорания дизельного двигателя. Получена модель с удовлетворительной погрешностью аппроксимации.

Список литературы

- [1] *Malada, H.R., Ivakhnenko, A. G.* Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, 1994.
- [2] *Bishop, C.M., Tipping, M.E.* Bayesian regression and classification // Suykens, J., Horvath, G. et. al., eds. Advances in Learning Theory: Methods, Models and Applications, Volume 190. IOS Press, NATO Science Series III: Computer and Systems Sciences, 2000. P 267–285.
- [3] *LeCun, Y., Denker, J. S., and Solla, S. A.* Optimal brain damage // Touretzky, D.S., ed. Advances in Neural Information Processing Systems. Morgan Kaufmann, San Mateo, CA, 1990. P. 598–605.
- [4] *MacKay, D.* Information, inference, learning algorithms. Cambridge University Press, 2003.
- [5] *MacKay, D.* Hyperparameters: optimise or integrate out? // Heidberger, G., ed. Maximum entropy and Bayesian Methods. Santa Barbara, Dordrecht: Kluwer, 1993.
- [6] *MacKay, D.* Bayesian interpolation // Neural Computation 4(3), 1992. P. 415–447.
- [7] *Nabney, I.T.* NETLAB: Algorithms for pattern recognition. Springer, 2004. P. 330.
- [8] *MacKay, D.* Choice of basis for Laplace approximation // Machine Learning, vol. 33(1), 1998.
- [9] *Branch, M.A., Coleman, T.F., Li, Y.* A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems // SIAM Journal on Scientific Computing, vol. 21(1), 1999. P. 1–23.