Министерство образования и науки Российской Федерации
Московский физико-технический институт
(государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Гребенников Евгений Владимирович

Метод дифференциальной бесперебoрной кросс-валидации в задачах восстановления зависимостей по эмпирическим данным

511656 - Математические и информационные технологии

Выпускная квалификационная работа бакалавра

Научный руководитель:
д.т.н., профессор
в.н.с. ВЦ РАН Моттль Вадим Вячеславович

Москва
2012

Содержание

Аннотация

При решении задач восстановления зависимостей в классе линейных моделей, содержащих структурный параметр, стоит необходимость настройки этого параметра. Классическим подходом для этого является использование метода скользящего контроля, в частности метода скользящего контроля по отдельным объектам. Недостатком данного метода является высокая вычислительная сложность, связанная с необходимостью проводить обучение столько раз, сколько объектов в обучающей совокупности. В работе предложен дифференциальной бесперебороной кросс-валидации для моделей числовых зависимостей, линейных по параметрам, при которой необходимо обучаться лишь один раз на всей выборке обучающих объектов. Объекты при таком подходе не удаляются полностью из обучения, а отбрасывается лишь часть объекта. Показано, что ошибка оценивания числовой функции на частично удаленном обучающем объекте линейно зависит от оставшегося веса этого объекта. Основываясь на этом факте, предложен функционал качества, использующий частные производные ошибок на объектах по их весам, и приведен метод его вычисления. Проведена серия вычислительных экспериментов, подтверждающая уместность использования нового метода дифференциальной кросс-валидации для задач восстановления числовых зависимостей.

1 Введение

Проблема восстановления зависимостей по эмпирическим данным принадлежит к одной из важнейших проблем современной информатики. Пусть $\omega \in \Omega$ - набор объектов реального мира, каждый из которых имеет ненаблюдаемую характеристику $y \in \mathbb{Y}$. Функция $y(\omega) : \Omega \to \mathbb{Y}$ известна наблюдателю только на конечном наборе обучающих объектов

$$\Omega^* = (\omega_j, y(\omega_j)), j = 1, \ldots, N.$$

Необходимо продолжить функцию на весь набор $\hat{y}(\omega) : \Omega \to \mathbb{Y}$, и таким способом оценить значение ненаблюдаемой характеристики для других объектов $\omega \in \Omega \setminus \Omega^*$ [1]. В случае когда функция может принимать конечный набор значений $\hat{y}(\omega) : \Omega \to \{y^{(1)}, \ldots, y^{(m)}\}$ данная задача называется задачей распознавания образов. Если же областью значений функции есть вся вещественная ось $\hat{y}(\omega) : \Omega \to \mathbb{R}$ задача называется задачей оценивания числовой зависимости. В данной работе рассматривается задача второго типа.

В общем случае, на практике, все принципы представления объектов реального мира делятся на два типа - векторами признаков и базирующиеся на сходстве/отличии объектов.

Принцип, основанный на выделении признаков, связывает с каждым объектом переменную $x(\omega) : \Omega \to \mathbb{X}$. Неизвестная числовая зависимость $\hat{y}(x(\omega)) : \mathbb{X} \to \mathbb{R}$ оценивается в этом случае из обучающего набора следующего вида:

$$\Omega^* : (x(\omega_j), y(\omega_j)), j = 1, \ldots, N.$$

Одно из простейших и популярных предположений в выделении признаков - представление объектов вещественными векторами признаков $\mathbf{x}(\omega) = (x_1(\omega), \ldots, x_n(\omega)) \in \mathbb{R}^n$. Тогда, в регрессионной модели $\hat{y}(x(\omega)) : \mathbb{R}^n \to \mathbb{R}$ параметры $\hat{\mathbf{c}} \in \mathbb{R}^n$и $\hat{b} \in \mathbb{R}$ особенно легко находятся по обучающей выборке. Регрессионная функция ищется в следующем параметрическом семействе:

$$\hat{y}(\mathbf{x}(\omega)) = \hat{\mathbf{c}}^T \mathbf{x}(\omega) + \hat{b} = \sum_{i=1}^{n} \hat{c}_i x_i(\omega) + \hat{b}.$$

В данной работе используется альтернативный и более общий принцип сходства/отличия представления объектов, предполагающий,что единственный способ различать объекты состоит в их попарном сравнении (ω', ω'') с помощью вещественной функции двух переменных $K(\omega', \omega'') : \Omega \times \Omega \to \mathbb{R}$. Таким образом обучающая

выборка представлена не векторами признаков, а матрицей попарных сравнений объектов:

$$\Omega^* : (K(\omega_j, \omega_l), j, l = 1, \ldots, N.)$$

Конкретнее, рассматривается кернельное представление объектов, подразумевающее что функция сравнения $K(\omega_j, \omega_l)$ является потенциальной функцией. Это значит, что функция симметрична $K(\omega_j, \omega_l) = K(\omega_l, \omega_j)$ и матрица попарных сравнений на обучающей выборке положительно полуопределена [2].

Предполагается, что кернел зависит от некоторого параметра β, т.е. $K(\omega_j, \omega_l|\beta)$ и таким образом линейная модель числовой зависимости, использующая кернел, содержит β в качестве структурного параметра модели. Существуют различные методики настройки структурного параметра, одна из которых - процедура скользящего контроля, впервые предложенная Бонгардом М.М. и Вайнцвайгом М.Н. в работах [3,4]. Недостатком данного метода служит высокая вычислительная сложность, связанная с необходимостью многократного разделения выборки на обучающую и тестирующую, обучения алгоритма на обучающей части и тестировании на тестирующей. В работах [5,6] предложены новые методики так называемой бесперебороной кросс-валидации, позволяющей избежать процедуры многократного обучения. В данной работе предложен новый подход бесперебороного скользящего контроля - дифференциальная кросс-валидация.

2 Линейная кернельная модель

Будем рассматривать следующий класс параметрических линейных моделей:

$$\hat{y}(\mathbf{x}) = \sum_{j=1}^{N} c_j K(\mathbf{x}_j, \mathbf{x}|\beta), \tag{1}$$

где $\mathbf{x}_j, j = 1, \ldots, N$-объекты обучения, а β - структурный параметр модели, который требуется настроить.

2.1 Нормальная модель наблюдений

В общем, мы рассматриваем набор пар $(\omega, y(\omega)) \in \Omega \times \mathbb{R}$ как вероятностное пространство. Это значит, что наблюдаемый объект и его ненаблюдаемая характеристика являются парой случайных величин в пространстве $\Omega \times \mathbb{R}$. Нахождение регрессионной зависимости будем искать через нахождение неизвестной плотности $\varphi^*(y|\mathbf{x})$ в $\mathbb{R}$.

Предположим случай линейной регрессионной модели $E(y|x; c_1, \ldots, c_n)$ с неизвестным вектором коэффициентов $\mathbf{c} = (c_1, \ldots, c_n)$ и будем предполагать нормальность соответствующего параметрического семейства распределений с неизвестной шумовой дисперсией $\xi > 0$:

$$\varphi(y|\mathbf{x}; \mathbf{c}, \xi) = \frac{1}{(2\pi)^{1/2}\xi^{1/2}} \exp\Big(-\frac{1}{2\xi}\{y - \sum_{j=1}^{N} c_j K(\mathbf{x}_j, \mathbf{x}|\beta)\}^2\Big) \tag{2}$$

Если, вдобавок, случайные величины ненаблюдаемой характеристики в обучающей выборке зависят только от соответствующего j-го объекта, то нормальное распределение совокупности наблюдений будет выглядеть следующим образом:

$$\Phi(y_1, y_2, ..., y_N|\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_N; \mathbf{c}, \xi) = \prod_{j=1}^{N} \varphi(y_j|\mathbf{x}_j; \mathbf{c}, \xi) =$$

$$= \frac{1}{(2\pi)^{N/2}\xi^{N/2}} \exp\Big(-\frac{1}{2\xi}\sum_{j=1}^{N}\{y_j - \sum_{l=1}^{N} c_l K(\mathbf{x}_l, \mathbf{x}_j)\}^2\Big) \tag{3}$$

2.2 Нормальная модель коэффициентов регрессии

В свою очередь, неизвестные коэффициенты числовой зависимости $(c_1, \ldots, c_N)$ априори рассматриваются как независимые случайные величины в пространстве $\mathbb{R}$ с нулевым математическим ожиданием $\mathbb{E}(c_j) = 0$ и круговой дисперсией ξ, совпадающей по значению с ξ из (2). Таким образом $\psi_j(c_j|\xi) = \frac{1}{\xi^{1/2}} \exp(-\frac{1}{2\xi}c_j^2)$. Совместная априорная плотность распределения выражается как произведение плотностей по каждой из координат:

$$\Psi(c_1, \ldots, c_N|\xi) \propto (\prod_{i=1}^{N} \xi)^{-1/2} \exp(-\frac{1}{2\xi}\sum_{j=1}^{N} c_j^2). \tag{4}$$

2.3 Байесовское оценивание коэффициентов кернельной регрессии

Апостериорное распределение вектора коэффициентов регрессии выглядит следующим образом:

$$P(\mathbf{c}|y_1, ...y_N, \xi) \propto \psi(\mathbf{c}|\xi)\Phi(y_1, y_2, ..., y_N|\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_N; \mathbf{c}, \xi) \tag{5}$$

Байесовская оценка вектора коэффициентов регрессии выводится из условия максимума апостериорной вероятности (5) и будет выглядеть следующим образом:

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} P(\mathbf{c}|y_1, ...y_N, \xi) = \arg\max_{\mathbf{c}}\{\ln\psi(\mathbf{c}|\xi) + \ln\Phi(y_1, y_2, ..., y_N|\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_N; \mathbf{c}, \xi)\} =$$

$$\arg\max_{\mathbf{c}}\Big(-\frac{1}{2\xi}\mathbf{c}^T\mathbf{c}-\frac{1}{2\xi}\sum_{j=1}^{N}(y_j-\sum_{l=1}^{N}c_lK(\mathbf{x}_j,\mathbf{x}_l))^2\Big)=$$

$$=\arg\min_{\mathbf{c}}\Big(\frac{1}{2\xi}\mathbf{c}^T\mathbf{c}+\frac{1}{2\xi}\sum_{j=1}^{N}(y_j-\sum_{l=1}^{N}c_lK(\mathbf{x}_j,\mathbf{x}_l))^2\Big)$$

Введем обозначение $\delta_j = y_j - \sum_{l=1}^{N} c_l K(\mathbf{x}_j, \mathbf{x}_l))^2$. Тогда задача оптимизации запишется в виде:

$$\begin{cases}\sum_{j=1}^{N} c_j^2 + \sum_{j=1}^{N}\delta_j^2 \to \min(c_1,\dots c_N,\delta_1,\dots,\delta_N),\\ y_j - \sum_{l=1}^{N} c_l K(\mathbf{x}_j,\mathbf{x}_l|\beta) = \delta_j, j=1,\dots,N.\end{cases} \tag{6}$$

3 Настройка параметров в кернельной регрессии

Припишем каждому из N обучающих объектов вес $0 \leqslant p_j \leqslant 1, j = 1, \ldots, N$.. В этом случае задача оптимизации переписывается в следующем виде:

$$\begin{cases} \sum_{j=1}^{N} c_j^2 + \sum_{j=1}^{N} p_j \delta_j^2 \to \min(c_1, \ldots c_N, \delta_1, \ldots, \delta_N), \\ y_j - \sum_{l=1}^{N} c_l K(\mathbf{x}_j, \mathbf{x}_l | \beta) = \delta_j, j = 1, \ldots, N. \end{cases} \tag{7}$$

Для решения данной оптимизационной задачи выписываем функцию Лагранжа:

$$L(\mathbf{c}, \delta_1, \ldots, \delta_N, \lambda_1, \ldots, \lambda_N) = \frac{1}{2}\mathbf{c}^T\mathbf{c} + \frac{1}{2}\sum_{j=1}^{N} p_j \delta_j^2 + \sum_{j=1}^{N} \lambda_j (y_j - \sum_{l=1}^{N} c_l K(\mathbf{x}_j, \mathbf{x}_l | \beta) - \delta_j).$$

Приравнивая частные производные функции Лагранжа по c_l и δ_l к нулю, получим следующие условия:

$$\frac{\partial L}{\partial c_l} = c_l - \sum_{j=1}^{N} \lambda_j K(\mathbf{x}_j, \mathbf{x}_l | \beta) = 0; \quad \frac{\partial L}{\partial \delta_l} = p_l \delta_l - \lambda_l = 0, l = 1, \ldots, N.$$

Таким образом выражение для коэффициентов в функции регрессии принимает следующий вид:

$$c_l = \sum_{j=1}^{N} p_j \delta_j K(\mathbf{x}_j, \mathbf{x}_l | \beta), l = 1, \ldots, N. \tag{8}$$

Подставляя в равенство $\delta_j = y_j - \sum_{l=1}^{N} c_l K(\mathbf{x}_j, \mathbf{x}_l | \beta)$ выражение для c_l получим:

$$y_j - \sum_{l=1}^{N} p_l \delta_l K(\mathbf{x}_l, \mathbf{x}_j) = \delta_j, \quad j = 1, \ldots, N,$$

Переписывая выражения относительно переменных δ_j приходим к следующему виду:

$$\sum_{l=1}^{N} [p_l K(\mathbf{x}_j, \mathbf{x}_l)]\delta_l + \delta_j = y_j, \quad j = 1, \ldots, N,$$

$$\sum_{l=1, l \neq j}^{N} [p_l K(\mathbf{x}_l, \mathbf{x}_j)]\delta_l + \{[p_j K(\mathbf{x}_j, \mathbf{x}_j)] + 1\}\delta_j = y_j, \quad j = 1, \ldots, N. \tag{9}$$

Введем обозначения для векторов : $\mathbf{y} = (y_1, \ldots, y_N)^T$, $\delta = (\delta_1, \ldots, \delta_N)^T$, $\mathbf{p} = (p_1, \ldots, p_N)^T$

В этом случае система уравнений на $\delta_j, j = 1, \ldots, N$ запишется в простом виде:

$$\{A + I\}\delta = \mathbf{y} \tag{10}$$

где матрица I - единичная матрица размера $N \times N$, а матрица A имеет следующий вид:

$$A = \begin{pmatrix} p_1 K(\mathbf{x}_1, \mathbf{x}_1) & p_2 K(\mathbf{x}_1, \mathbf{x}_2) & \dots & p_N K(\mathbf{x}_1, \mathbf{x}_N) \\ p_1 K(\mathbf{x}_2, \mathbf{x}_1) & p_2 K(\mathbf{x}_2, \mathbf{x}_2) & \dots & p_N K(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ p_1 K(\mathbf{x}_N, \mathbf{x}_1) & p_2 K(\mathbf{x}_N, \mathbf{x}_2) & \dots & p_N K(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix}$$

Таким образом, вектор ошибок на обучающих объектах при линейном регрессионном оценивании находится по следующей формуле:

$$\delta(\mathbf{p}) = \{A + I\}^{-1}\mathbf{y}.$$

3.1 Процедура скользящего контроля

Для настройки параметров регрессионной оценки традиционно используется процедура скользящего контроля CV. Для этого обучающая выборка $\Omega^* = (\omega_j, y(\omega_j)), j = 1, \ldots, N.$ разбивается S различными способами на две непересекающиеся подвыборки: $\Omega^* = \Omega_n^m \cup \Omega_n^k$, где Ω_n^m - обучающая подвыборка длины m, Ω_n^k - контрольная подвыборка длины $k = N - m, n = 1, \ldots, S$-номер разбиения.

Для каждого разбиения n строится оценка регрессионной зависимости $\hat{y}_n = y(\Omega_n^m)$ и вычисляется значение функционала качества:

$$Q_n = \sum_{j:\omega_j \in \Omega_n^k} [\hat{y}_n(\omega_j) - y(\omega_j)]^2.$$

Среднее арифметическое значений Q_n по всем разбиениям называется оценкой скользящего контроля:

$$CV(\hat{y}, \Omega^N) = \frac{1}{S} \sum_{n=1}^{S} Q_n = \frac{1}{S} \sum_{n=1}^{S} [\hat{y}_n(\omega_j) - y(\omega_j)]^2$$

Пожалуй, самый распространенный вариант скользящего контроля - контроль по отдельным объектам (leave-one-out CV). В этом случае оценка скользящего контроля строится по всем $S = C_N^1 = N$ разбиениям. Преимущества LOO в том, что каждый объект ровно один раз участвует в контроле, а длина обучающих подвыборок лишь на единицу меньше длины полной выборки. В то же время, существенным недостатком LOO является большая ресурсоёмкость, так как обучаться приходиться Nраз. В данной работе предлагается метод существенно ускоряющий подсчет ошибок при использовании контроля по отдельным объектам LOO.

3.2 Критерий дифференциального leave-one-out

Теорема 1. *Зависимость ошибки на j-ом объекте δ_j от веса этого объекта p_j имеет линейный характер.*

Доказательство 1.

Доказательство основывается на непосредственном решении системы линейных уравнений на δ_j предыдущего раздела. Приведем доказательство для случая, когда в обучении находится всего $N = 2$ два объекта. Система (10) тогда перепишется в

виде:

$$\begin{cases} (K(\mathbf{x}_1, \mathbf{x}_1)p_1 + 1)\delta_1(\mathbf{p}) + K(\mathbf{x}_1, \mathbf{x}_2)\delta_2(\mathbf{p})p_2 = y_1, \\ K(\mathbf{x}_2, \mathbf{x}_1)p_1\delta_1(\mathbf{p}) + (K(\mathbf{x}_2, \mathbf{x}_2)p_2 + 1)\delta_2(\mathbf{p}) = y_2 \end{cases}$$

Для удобства сделаем следующие обозначения: $K(\mathbf{x}_i, \mathbf{x}_j) = K_{ij}$ Из второго уравнения выразим $\delta_2(\mathbf{p})$ и подставим его в первое уравнение:

$$\delta_2(\mathbf{p}) = \frac{y_2 - K_{21}p_1\delta_1(\mathbf{p})}{K_{22}p_2 + 1} \Rightarrow (K_{11}p_1 + 1)\delta_1 + K_{12}p_2\frac{y_2 - K_{21}p_1\delta_1}{K_{22}p_2 + 1} = y_1$$

Группируя слагаемые, содержащие δ_1 получаем:

$$\delta_1(1 + K_{11}p_1 - \frac{K_{12}K_{21}p_1p_2)}{K_{22}p_2 + 1}) = y_1 - \frac{K_{12}y_2p_2}{K_{22}p_2 + 1}$$

Разделив правую и левую часть на коэффициент при δ_1 и обозначив числитель за C_1, а коэффициент при p_1 за D_1 получим:

$$\delta_1 = \frac{C_1}{1 + D_1p_1}. \tag{11}$$

В силу малости веса p_1 можно разложить дробь, воспользовавшись формулой Маклорена $(1 + \alpha x)^{-1} \approx 1 - \alpha x$:

$$\delta_1 = C_1(1 - D_1p_1).$$

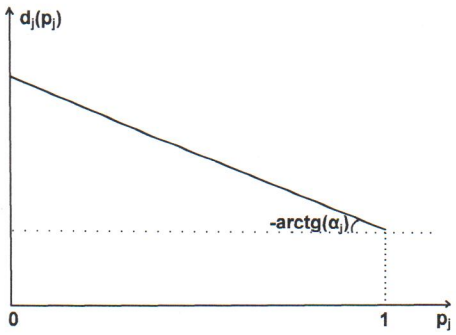
Аналогично выражая из первого уравнения $\delta_\mathbf{1}(\mathbf{p})$ и подставляя во второе уравнение придем к аналогичному выражению для $\delta_\mathbf{2}(\mathbf{p})$:

$$\delta_2 = C_2(1 - D_2p_2)$$

В силу громоздкости выкладок доказательство в полном объеме в работе не изложено, приведем лишь общую схему доказательства. Для случая, когда в обучении N объектов, система (10) будет иметь N линейных уравнений на $\delta_1, \ldots, \delta_N$. Докажем, что δ_N линейно зависит от p_N. Для этого в первых $N - 1$ уравнении перенесем слагаемые содержащие $\delta_N p_N$ в правую сторону, получим из первых $N - 1$ уравнения систему вида: $\mathbf{C}\delta^T = \mathbf{b}^T$, где $\mathbf{C}$ - некоторая матрица коэффициентов, независящих от δ_N и p_N, $\delta = (\delta_1, \ldots, \delta_{N-1})$, $\mathbf{b} = (y_1 - K_{1N}\delta_N p_N, \ldots, y_{N-1} - K_{1N}\delta_N p_N)$. Тогда линейная комбинация $(\delta_1, \ldots, \delta_{N-1})$ с вектором коэффициентов $\mathbf{a}$ в N-ом уравнении записывается в виде $\mathbf{a}\delta^T = \mathbf{a}\mathbf{C}^{-1}\mathbf{b}$, что в свою очередь представляется в виде $A + B\delta_N p_N$. Тогда N-ое уравнение системы можно переписать в виде:

$A + B\delta_N p_N + (K_{NN}p_N + 1)\delta_N = y_N \Rightarrow \delta_N = \frac{C}{1+Dp_N}$, что аналогично выражению (11). Раскладывая в силу малости p_N по формуле Маклорена, получаем утверждение теоремы. □

Из теоремы следует, что схематически график зависимости δ_j от p_j имеет следующий вид:



Заметим, что значение $\delta_j(0)$ соответствует ошибке на j-ом объекте, когда тот полностью исключен из обучения, т.е. ошибке в случае обычного leave-one-out, а значение $\delta_j(1)$ соответствует значению ошибки на j-ом объекте, когда тот полностью присутствует в обучении. Случай, когда $0 < p_j < 1$, означает частичное присутствие объекта в обучении. В силу линейности зависимости, ошибку на объекте при произвольном весе $0 \leqslant p_j \leqslant 1$ можно выписать через ошибку, в случае когда объект полностью присутствует в обучении:

$$\delta_j(p_j) = \delta_j(1) - \alpha_j(1 - p_j), \;\; \alpha_j = \frac{\partial \delta_j(p_j)}{\partial p_j}\Big|_{p_j=1}, j = 1, \ldots, N.$$

В частности, ошибка в случае обычного leave-one-out примет очень простой вид:

$$\delta_j^{LOO} = \delta_j(0) = \delta_j(1) - \alpha_j$$

В случае обычного leave-one-out, функционал качества записывается как усредненный квадрат ошибки, а именно:

$$Q_{LOO} = \frac{1}{N}\sum_{j=1}^{N}\delta_j^2(0) \to \min$$

Подставляя в это выражение, выражение для $\delta_j(0)$, получим:

$$Q_{LOO} = \frac{1}{N}\sum_{j=1}^{N}\Big(\delta_j(1) - \alpha_j\Big)^2 = \frac{1}{N}\sum_{j=1}^{N}\Big(\delta_j^2(1) - 2\delta_j(1)\alpha_j + \alpha_j^2\Big)$$

Учитывая что $\sum_{j=1}^{N}\delta_j^2(1) = const$ и отбрасывая числовые коэффициенты приходим к критерию качества дифференциальной кросс-валидации:

$$Q_{DIFF} = \sum_{j=1}^{N}\Big(\Big(\frac{\partial\delta_j(p_j)}{\partial p_j}\Big|_{p_j=1}\Big)^2 - \frac{\partial\delta_j(p_j)}{\partial p_j}\Big|_{p_j=1}\Big) \to \min \qquad (12)$$

3.3 Алгоритм подсчета частных производных

Полученный критерий существенно опирается на нахождение частных производных ошибок по весам объектов. Для решения данной проблемы, предлагается алгоритм нахождения матрицы Якоби следующего вида:

$$\mathbf{J}((\mathbf{x}_j, y_j)_{j=1}^{N}) = \begin{pmatrix} \frac{\partial}{\partial p_1}\delta_1(\mathbf{p}) & \frac{\partial}{\partial p_2}\delta_1(\mathbf{p}) & \dots & \frac{\partial}{\partial p_N}\delta_1(\mathbf{p}) \\ \frac{\partial}{\partial p_1}\delta_2(\mathbf{p}) & \frac{\partial}{\partial p_2}\delta_2(\mathbf{p}) & \dots & \frac{\partial}{\partial p_N}\delta_2(\mathbf{p}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial p_1}\delta_N(\mathbf{p}) & \frac{\partial}{\partial p_2}\delta_N(\mathbf{p}) & \dots & \frac{\partial}{\partial p_N}\delta_N(\mathbf{p}) \end{pmatrix}\Bigg|_{\mathbf{p}=(1,\dots,1)} \qquad (13)$$

Для подсчета критерия качества достаточно найти лишь диагональные элементы матрицы Якоби, но легче искать всю матрицу сразу. Запишем систему (10) в явном виде:

$$\begin{cases} \{K(\mathbf{x}_1,\mathbf{x}_1)p_1+1\}\delta_1(\mathbf{p}) + \{K(\mathbf{x}_1,\mathbf{x}_2)p_2\}\delta_2(\mathbf{p}) + \dots + \{K(\mathbf{x}_1,\mathbf{x}_N)p_N\}\delta_N(\mathbf{p}) & = y_1, \\ \{K(\mathbf{x}_2,\mathbf{x}_1)p_1\}\delta_1(\mathbf{p}) + \{K(\mathbf{x}_2,\mathbf{x}_2)p_2+1\}\delta_2(\mathbf{p}) + \dots + \{K(\mathbf{x}_2,\mathbf{x}_N)p_N\}\delta_N(\mathbf{p}) & = y_2, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots & \\ \{K(\mathbf{x}_N,\mathbf{x}_1)p_1\}\delta_1(\mathbf{p}) + \{K(\mathbf{x}_N,\mathbf{x}_2)p_2\}\delta_2(\mathbf{p}) + \dots + \{K(\mathbf{x}_N,\mathbf{x}_N)p_N+1\}\delta_N(\mathbf{p}) & = y_N. \end{cases}$$

Продифференцируем левые и правые части каждого из уравнений по p_j :

$$\begin{cases} \{K(\mathbf{x}_1,\mathbf{x}_1)\}\delta_1(\mathbf{p}) + \\ +\{K(\mathbf{x}_1,\mathbf{x}_1)p_1+1\}\frac{\partial\delta_1(\mathbf{p})}{\partial p_1} + \{K(\mathbf{x}_1,\mathbf{x}_2)p_2\}\frac{\partial\delta_2(\mathbf{p})}{\partial p_1} + \dots + \{K(\mathbf{x}_1,\mathbf{x}_N)p_N\}\frac{\partial\delta_N(\mathbf{p})}{\partial p_1} = 0, \\ \{K(\mathbf{x}_2,\mathbf{x}_1)\}\delta_1(\mathbf{p}) + \\ +\{K(\mathbf{x}_2,\mathbf{x}_1)p_1\}\frac{\partial\delta_1(\mathbf{p})}{\partial p_1} + \{K(\mathbf{x}_2,\mathbf{x}_2)p_2+1\}\frac{\partial\delta_2(\mathbf{p})}{\partial p_1} + \dots + \{K(\mathbf{x}_2,\mathbf{x}_N)p_N\}\frac{\partial\delta_N(\mathbf{p})}{\partial p_1} = 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \{K(\mathbf{x}_N,\mathbf{x}_1)\}\delta_1(\mathbf{p}) + \\ +\{K(\mathbf{x}_N,\mathbf{x}_1)p_1\}\frac{\partial\delta_1(\mathbf{p})}{\partial p_1} + \{K(\mathbf{x}_N,\mathbf{x}_2)p_2\}\frac{\partial\delta_2(\mathbf{p})}{\partial p_1} + \dots + \{K(\mathbf{x}_N,\mathbf{x}_N)p_N+1\}\frac{\partial\delta_N(\mathbf{p})}{\partial p_1} = 0. \end{cases}$$

Перенесем слагаемые, не зависящие от производных, в правую часть:

$$\begin{cases} \{K(\mathbf{x}_1,\mathbf{x}_1)p_1+1\}\frac{\partial\delta_1(\mathbf{p})}{\partial p_1}+\{K(\mathbf{x}_1,\mathbf{x}_2)p_2\}\frac{\partial\delta_2(\mathbf{p})}{\partial p_1}+\cdots+\{K(\mathbf{x}_1,\mathbf{x}_N)p_N\}\frac{\partial\delta_N(\mathbf{p})}{\partial p_1}= \\ -\{K(\mathbf{x}_1\mathbf{x}_1)\}\delta_1(\mathbf{p}), \\ \{K(\mathbf{x}_2,\mathbf{x}_1)p_1\}\frac{\partial\delta_1(\mathbf{p})}{\partial p_1}+\{K(\mathbf{x}_2,\mathbf{x}_2)p_2+1\}\frac{\partial\delta_2(\mathbf{p})}{\partial p_1}+\cdots+\{K(\mathbf{x}_2,\mathbf{x}_N)p_N\}\frac{\partial\delta_N(\mathbf{p})}{\partial p_1}= \\ -\{K(\mathbf{x}_2,\mathbf{x}_1)\}\delta_1(\mathbf{p}), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \{K(\mathbf{x}_N,\mathbf{x}_1)p_1\}\frac{\partial\delta_1(\mathbf{p})}{\partial p_1}+\{K(\mathbf{x}_N,\mathbf{x}_2)p_2\}\frac{\partial\delta_2(\mathbf{p})}{\partial p_1}+\cdots+\{K(\mathbf{x}_N,\mathbf{x}_N)p_N+1\}\frac{\partial\delta_N(\mathbf{p})}{\partial p_1}= \\ -\{K(\mathbf{x}_N,\mathbf{x}_1)\}\delta_1(\mathbf{p}). \end{cases}$$

Мы получили систему линейных уравнений относительно искомых производных по весу первого наблюдения $(\frac{\partial\delta_1(\mathbf{p})}{\partial p_1}),\ldots,\frac{\partial\delta_N(\mathbf{p})}{\partial p_1})$. Аналогичные системы получаются для производных по весам каждого из наблюдений $(\frac{\partial\delta_1(\mathbf{p})}{\partial p_j}),\ldots,\frac{\partial\delta_N(\mathbf{p})}{\partial p_j})$, $j=1,\ldots,N$, причем системы будут отличаться только правыми частями. Совокупность этих систем в общем виде может быть записана в виде матричного уравнения:

$$A\mathbf{J}=B, \tag{14}$$

где

$$A=\begin{pmatrix} p_1K(\mathbf{x}_1,\mathbf{x}_1)+1 & p_2K(\mathbf{x}_1,\mathbf{x}_2) & \ldots & p_NK(\mathbf{x}_1,\mathbf{x}_N) \\ p_1K(\mathbf{x}_2,\mathbf{x}_1) & p_2K(\mathbf{x}_2,\mathbf{x}_2)+1 & \ldots & p_NK(\mathbf{x}_2,\mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ p_1K(\mathbf{x}_N,\mathbf{x}_1) & p_2K(\mathbf{x}_N,\mathbf{x}_2) & \ldots & p_NK(\mathbf{x}_N,\mathbf{x}_N) \end{pmatrix}$$

$$B=\begin{pmatrix} -K(\mathbf{x}_1,\mathbf{x}_1)\delta_1(\mathbf{p}) & -K(\mathbf{x}_1,\mathbf{x}_2)\delta_2(\mathbf{p}) & \ldots & -K(\mathbf{x}_1,\mathbf{x}_N)\delta_N(\mathbf{p}) \\ -K(\mathbf{x}_2,\mathbf{x}_1)\delta_1(\mathbf{p}) & -K(\mathbf{x}_2,\mathbf{x}_2)\delta_2(\mathbf{p}) & \ldots & -K(\mathbf{x}_2,\mathbf{x}_N)\delta_N(\mathbf{p}) \\ \vdots & \vdots & \ddots & \vdots \\ -K(\mathbf{x}_N,\mathbf{x}_1)\delta_1(\mathbf{p}) & -K(\mathbf{x}_N,\mathbf{x}_2)\delta_2(\mathbf{p}) & \ldots & -K(\mathbf{x}_N,\mathbf{x}_N)\delta_N(\mathbf{p}) \end{pmatrix}$$

Таким образом, матрица Якоби $\mathbf{J}$ находится как решение матричного уравнения (14).

4 Вычислительный эксперимент

Проиллюстрируем работу алгоритма на примере синтетических данных , использованных Бишопом и Типпингом в [7]. Объекты $\omega \in \Omega$ представлены двумерными векторами признаков $\mathbf{x}(\omega) = (x_1(\omega), x_2(\omega)) \in \mathbb{R}^2$, а значение отклика на объекте задается функцией

$$y(\mathbf{x}) = \sum_{i=1}^{150} a_i K(\mathbf{x}_i, \mathbf{x}) + \varepsilon. \tag{15}$$

Здесь $\mathbf{x}_i, i = 1, \ldots 150$ - фиксированные точки, для каждой из которых известен коэффициент a_i, $\varepsilon \in N(0, 1/100)$ - нормальный шум, а $K(\mathbf{x}_i, \mathbf{x}_j) = exp\{-0.5\|\mathbf{x}_i - \mathbf{x}_j\|^2\}$ Для обучения сформируем набор $\mathbf{x}_j, j = 1, \ldots, N$из $N = 250$ объектов, равномерно распределенных в квадрате, а значения отклика для всех точек в квадрате подсчитываются по формуле (15).

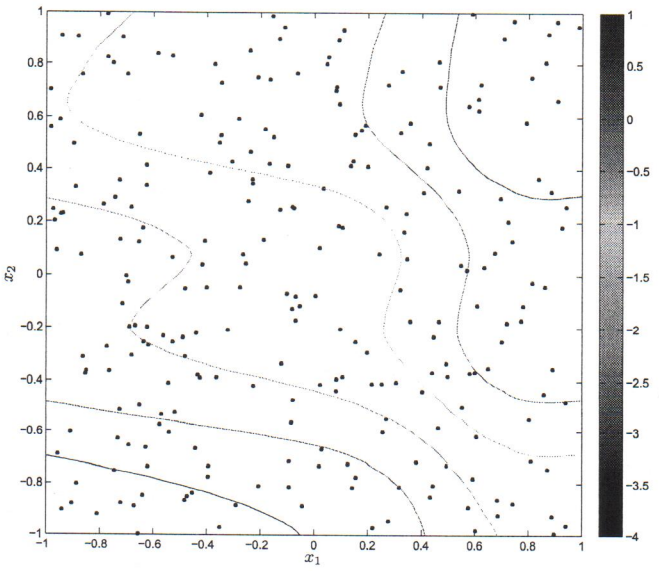


Рис. 1: Линии равных значений модельной зависимости и точки, используемые для обучения

Функцию регрессии будем искать в классе линейный параметрических моделей вида:

$$\hat{y}(\mathbf{x}|\mathbf{c}, \beta) = \sum_{j=1}^{N} c_j K(\mathbf{x}_j, \mathbf{x}|\beta)$$

Ошибка регрессионного оценивания подсчитывается с помощью функционала качества (12) Q_{DIFF}. Класс ядер рассматриваемый в эксперименте - экспоненциальное

ядро с параметром β в показателе экспоненты:

$$K(\mathbf{x}', \mathbf{x}''|\beta) = exp(-\beta\|\mathbf{x}' - \mathbf{x}''\|^2) \tag{16}$$

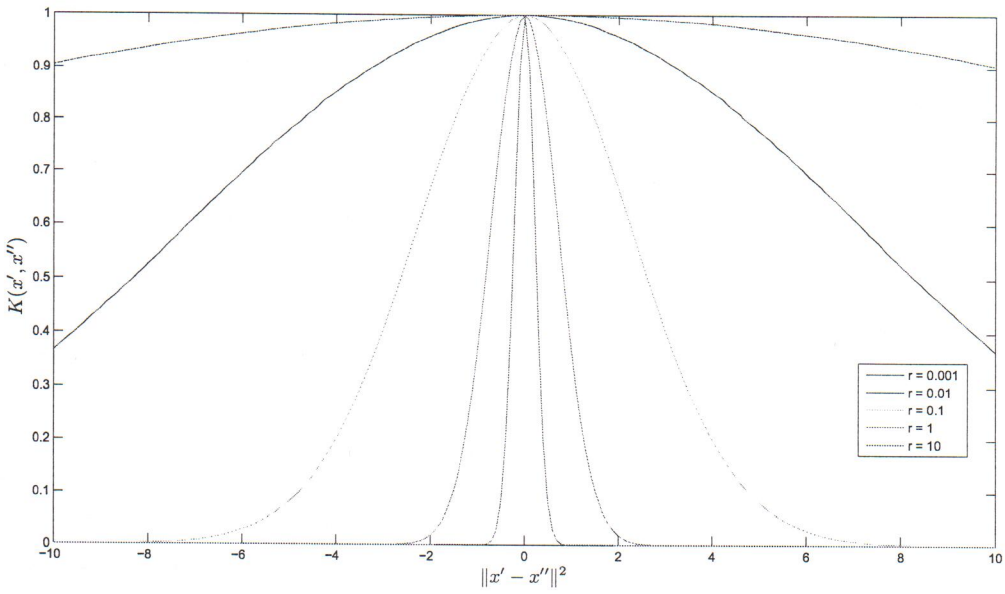


Рис. 2: Радиальное ядро при различных значениях параметра β

Данное ядро интересно тем, что в зависимости от значения параметра, объекты обучения по разному влияют на значения отклика в других точках (см. рис. 2). При малых положительных значениях β график функции очень пологий, следовательно, значение отклика на объекте зависит от многих обучающих объектов, в том числе неинформативных, а значит ошибка оценивания велика. В то же время, при слишком больших значениях параметра β ядро очень быстро спадает к значениям, близким к нулю, таким образом на значение отклика на объекте влияет крайне малое число обучающих объектов либо не влияет вообще и ошибка регрессионного оценивания также велика. Исходя из этого, должно существовать оптимальное β, при котором ошибка регрессионного оценивания минимальна.

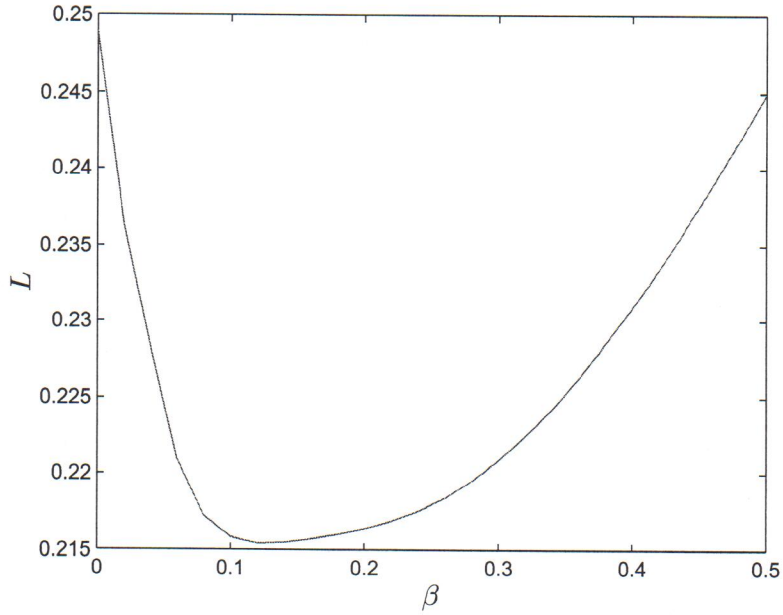


Рис. 3: Зависимость эмпирического риска Q_{DIFF} от параметра β

На графике наблюдается явно выраженный минимум в точке $\beta_0 = 0.11$, что подтверждает уместность использования дифференциальной кросс-валидации и функционала качества Q_{DIFF} вместо обычного контроля по отдельным объектам LOO.

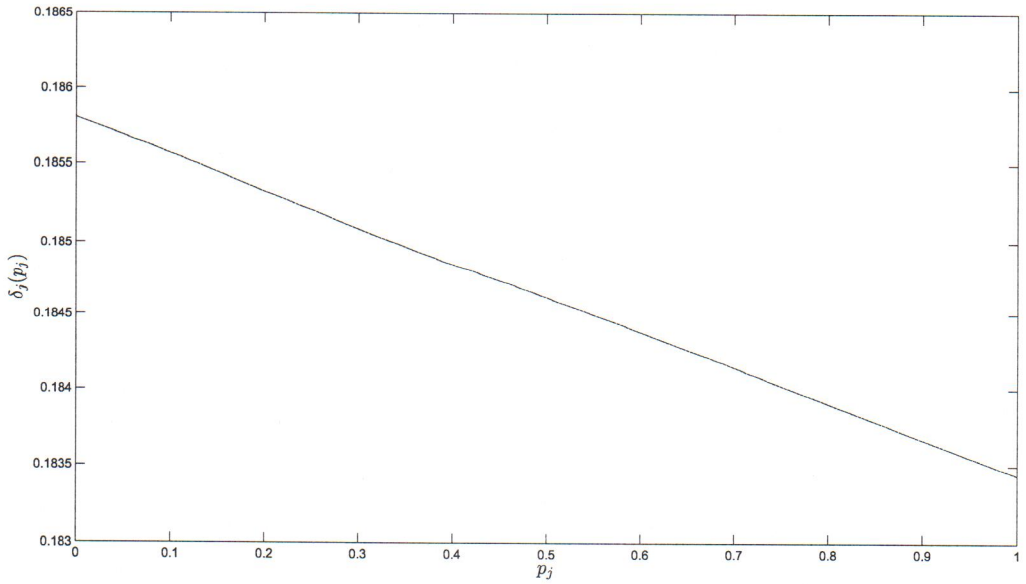


Рис. 4: Ошибка на объекте в зависимости от его веса

Также эксперимент подтверждает утверждение Теоремы 1 о том, что зависимость ошибки на обучающем объекте линейно зависит от его веса. На рис. 4 приведен график ошибки для случайно выбранного объекта из 250 обучающих объектов.

5 Заключение

В работе предложен новый подход дифференциальной кросс-валидации для настройки параметров кернельной регрессионной модели и проанализированы результата работы на модельных данных. Особенностью данного подхода является его бесперебборность, что существенно уменьшает время обсчета в отличии от обычного метода leave-one-out.

- В работе предложен новый функционал качества, использующий частные производные ошибок по весу, а также предложен алгоритм подсчета этих производных.
- Теоретически и эксперементально показано, что ошибка регрессии на обучающем объекте линейно зависит от его веса.
- Проведена серия численных экспериментов на модельных данных, результаты которых позволяют говорить об актуальности использования метода дифференциальной кросс-валидации при решении задач регрессионного оценивания.

Список литературы

[1] Vapnik V. *Estimation of Dependencies Based on Empirical Data*, Springer, 1982.

[2] Vapnik V. *Statistical Learning Theory*, John-Wiley & Sons, Inc., 1998.

[3] Бонгард М.М. *Проблема узнавания*, М.:Наука, 1967, 320 с.

[4] Бонгард М.М., Вайнцвайг М.Н. *Об оценках ожидаемого качества признаков* Проблемы кибернетики, вып. 20, 1968.

[5] E. Ezhova, V. Mottl, O. Krasotkina. Estimation of time-varying linear regression with unknown time-volatility via continuous generalization of the Akaike Information Criterion.Proceedings of WorldAcademy of Science, Engineering and Technology, No.51, Mar. 2009, pp. 144-150.

[6] Ежова Е.О., Красоткина О.В., Моттль В.В. Непрерывная коррекция информационного критерия Акаике для регуляризованного оценивания сверхбольшого числа параметров регрессионных моделей данных с неизвестной дисперсией. Доклады 8-й Международной конференции "Интеллектуализация обработки информации Кипр, Пафос, 17-24 октября 2010, с. 51-54.

[7] C.Bishop, M.Tipping *Variational Relevance Vector Machines.* , Proceedings of the 16th Conference on Uncertainty in Artificcial Intelligence, pp. 46-53. Morgan Kaufmann, 2000.

[8] Г. И. Ивченко, Ю. И. Медведев *Введение в математическую статитсику*, ЛКИ, 2009.