ОЦЕНКА ТОЧНОСТИ РЕШЕНИЯ ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ МЕТОДОМ КОЛЛОКАЦИИ

Бывшев В.А.
Финансовый университет при Правительстве Российской Федерации
VByvshev@fa.ru

Аннотация. Обсуждается приложение метода коллокации к решению задач машинного обучения, востребованного в экономике, финансах, медицине и других отраслях. Метод коллокации восходит к Л.В. Канторовичу и состоит в восстановлении искомых функций по значениям функционалов. Метод коллокации используется в математической физике, геофизике, геодезии. В данной работе обсуждается оценка точности решения задачи машинного обучения методом коллокации.

Ключевые слова: Задача и метод машинного обучения, задача и метод коллокации, искомая функция и функционалы, оценка точности решения.

Введение

Термин «коллокация» используется в вычислительной математике, геофизике, астрономии и геодезии после выхода работы [1] лауреата Нобелевской премии по экономике Л. В. Канторовича. С математической точки метод коллокации является обобщением метода наименьших квадратов на случай бесконечномерных гильбертовых пространств. В данной работе задача коллокации обсуждается во взаимосвязи с задачей машинного обучения, а затем строится процедура оценки точности решения методом коллокации.

1 Задача машинного обучения

Задачи машинного обучения по прецедентам состоят в следующем [2], [3]. Задано множество X объектов $x \in X$ и задано множество Y допустимых ответов у $\in Y$. Каждый объект $x \in X$ описывается набором *признаков* $a(x) = (a_1(x), \dots, a_q(x))$, и с этим набором x отождествляется: $x \sim a(x)$. Каждый признак $a_j(x)$ имеет смысл некоторой характеристики объекта x и формально может интерпретироваться как отображение множества объектов X в множество A_j возможных значений признака a_j, то есть $a_j: X \to A_j$. В зависимости от типа множества A_j признак a_j может быть бинарным, категориальным или количественным. Отметим, что категориальные признаки можно квантифицировать, поэтому будем предполагать, что все признаки в наборе $a(x)$ являются количественными, то есть $a(x) \in R_q$. Ниже под объектом $x \in X$ будет пониматься его *количественное* признаковое описание $a(x)$. Добавим, что операцию признакового описания объектов $x \in X$ можно интерпретировать как отображение $a: X \to A \subset R_q$.

По предположению, существует неизвестная функция $f: X \to Y$, значения которой $y_i = f(x_i)$ известны на конечном множестве объектов $(x_1, x_2, \dots, x_n) \subset X$. Пары $(a(x_i), y_i)$ именуются *прецедентами;* совокупность $S = (a(x_i), y_i)_{i=1}^N$ прецедентов образуют *выборку*. Выборку разделяют на две части: на обучающую выборку $S_l = (a(x_i), y_i)_{i=1}^n$ и контролирующую выборку $S_c = \left(a(x_{(i)}), y_{(i)}\right)_{(i)=1}^{(n)}$. Требуется по выборке S построить функцию $\hat{f}: X \to Y$, приближающую неизвестную функцию $f: X \to Y$ на всём множестве X.

Структура множества ответов Y определяет тип задачи машинного обучения. Если множество Y является *бинарным или категориальным*, то задача машинного обучения называется *задачей классификации*; если же множество Y является *множеством действительных чисел* R_1, то задача машинного обучения именуется *задачей восстановления регрессии.*

Как решается задача машинного обучения? Схема решения состоит из трёх шагов [2].

Шаг 1. Задаётся некоторое параметрическое семейство Ф функций $\varphi(a; \theta)$:

$$(1.1) \qquad \Phi = \{\varphi(a; \theta) \big| \varphi: R_q \times \Theta \to \mathrm{Y}\}.$$

Аргументом функции $\varphi(a; \theta) \in \Phi$ служит вектор $a = a(x)$ признакового описания объекта $x \in X$, компоненты вектора θ являются искомыми параметрами функции $\varphi(a; \theta)$, при этом $\theta \in \Theta \subset R_m$, где $\Theta -$ множество допустимых значений параметров. Областью изменения функции $\varphi(a; \theta)$ служит множество ответов Y. Вот два примера [2] функции $\varphi(a; \theta)$:

$$(1.2) \ \ \varphi(a(x); \theta) = \textstyle\sum_{j=1}^q \theta_j \cdot a_j(x).$$

$$(1.3) \ \ \varphi(a(x); \theta) = sign(\textstyle\sum_{j=1}^q \theta_j \cdot a_j(x))).$$

Функция, определённая по правилу (1.2), используется в задаче восстановления регрессии; функция (1.3) применяется в задачах классификации на два класса.

Пусть $\theta -$ некоторый произвольный фиксированный вектор параметров, $a(x) -$ признаковое описание некоторого объекта $x \in X$. Значение

$$(1.4) \ \ \tilde{y} = \varphi(a(x); \theta)$$

имеет смысл *оценки* истинного ответа $y = f(x)$ на объекте x; разность

$$(1.5) \quad \tilde{y} - y = e(a(x), \varphi, \theta)$$

является истинной ошибкой оценки $\tilde{y}$ ответа $y = f(x)$ на объекте $x \in X$ при векторе параметров θ. Качество вектора параметров θ измеряется значением некоторой функции ошибки $e(a(x), \varphi, \theta)$. Такую функцию принято именовать *функцией потерь* и обозначать $Loss$. В силу (1.5), функция потерь является функцией аргументов $(a(x), \varphi, \theta)$. Вот три примеры функции потерь [2].

$$(1.6) \ \ Loss(a(x), \varphi, \theta) = |e(a(x), \varphi, \theta)| -$$

абсолютная истинная ошибка оценки $\tilde{y}$.

$$(1.7) \quad Loss(a(x), \varphi, \theta) = e(a(x), \varphi, \theta)^2 -$$

квадрат истинной ошибки оценки $\tilde{y}$.

$$(1.8) \quad Loss(a(x), \varphi, \theta) = I\{e(a(x), \varphi, \theta) \neq 0\} -$$

индикатор наличия ошибки в оценке $\tilde{y}$.

Функции потерь (1.6) и (1.7) используются в задаче восстановления регрессии; функция (1.8) применяется в задаче классификации.

Пусть $Loss(a(x), \varphi, \theta)$ – некоторая функция потерь. Её значения зависят от типа функций $\varphi \in \Phi$, вектора параметров $\theta \in \Theta$ и объекта $x \in X$. Усреднённое по множеству X значение функции потерь обозначим $Loss(\varphi, \theta)$, то есть

$$(1.9) \quad Loss(\varphi, \theta) = M_x(Loss(a(x), \varphi, \theta)).$$

Значение $Loss(\varphi, \theta)$ зависит от типа функций $\varphi \in \Phi$ и вектора параметров $\theta \in \Theta$. Величина $Loss(\varphi, \theta)$ интерпретируется как *мера риска прогноза* (1.4) ответа $y = f(x)$ при заданном векторе параметров θ. На практике значение $Loss(\varphi, \theta)$, как правило, не известно, но всегда доступна его оценка, например, по обучающей выборке S_l:

$$(1.10)\ Loss(\widetilde{\varphi, \theta}, S_l) = \frac{1}{n}\sum_{i=1}^{n} Loss(a(x_i), \varphi, \theta),$$

именуемая эмпирическим риском.

Шаг 2. На обучающей выборке решается вариационная задача на условный экстремум

$$(1.11)\ \begin{cases} Loss(\widetilde{\varphi, \theta}, S_l) = \frac{1}{n}\sum_{i=1}^{n} Loss(a(x_i), \varphi, \theta) \to min \\ \theta \in \Theta \end{cases}.$$

То есть отыскивается такой вектор параметров $\hat{\theta} \in \Theta$, который минимизирует эмпирический риск на обучающей выборке:

$$(1.11)'\ \begin{cases} Loss\left(\widetilde{\varphi, \hat{\theta}}, S_l\right) = \frac{1}{n}\sum_{i=1}^{n} Loss(a(x_i), \varphi, \hat{\theta}) = min\{Loss(\widetilde{\varphi, \theta}, S_l)\} \\ \theta \in \Theta \end{cases}$$

Отметим, что если функция потерь определена равенством (1.7), то решение задачи (1.11) соответствует методу наименьших квадратов.

Шаг 3. При найденном векторе параметров $\hat{\theta}$ вычисляется эмпирический риск на контролирующей выборке:

$$(1.12) \quad Loss\left(\widetilde{\varphi, \hat{\theta}}, S_c\right) = \frac{1}{(n)}\sum_{(i)=1}^{(n)} Loss\left(a\left(x_{(i)}\right), \varphi, \hat{\theta}\right).$$

Если величина $Loss\left(\widetilde{\varphi, \hat{\theta}}, S_c\right)$ значительно не превосходит количество $Loss\left(\widetilde{\varphi, \hat{\theta}}, S_l\right)$, то есть если имеет место приближённое равенство

$$(1.13) \quad Loss\left(\widetilde{\varphi, \hat{\theta}}, S_c\right) \approx Loss\left(\widetilde{\varphi, \hat{\theta}}, S_l\right),$$

то задача машинного обучения считается решённой; это означает, что найдена функция

$$(1.14) \quad \tilde{y} = \hat{f}(x) = \varphi(a(x); \hat{\theta}),$$

приближающая неизвестную функцию $f: X \to Y$ на всём множестве X. Функцию (1.14) используют для прогноза ответов. В противном случае, когда величина $Loss\left(\widetilde{\varphi, \hat{\theta}}, S_c\right)$ значительно превосходит $Loss\left(\widetilde{\varphi, \hat{\theta}}, S_l\right)$, говорят [2], что произошло *переобучение* модели (1.14), меняют состав обучающей и контролирующей выборок и повторяют шаги 1, 2 и 3 до достижения условия (1.13).

2 Задача коллокации

Термин «коллокация» используется в вычислительной математике, геофизике, астрономии и геодезии после выхода работы [1] лауреата Нобелевской премии по экономике Л. В. Канторовича. С математической точки метод коллокации является обобщением метода наименьших квадратов на случай бесконечномерных гильбертовых пространств. Современное состояние этого метода отражено в монографии [4] выдающегося австрийского геофизика и геодезиста Г. Морица. Адаптированное изложение метода коллокации с примерами из вычислительной математики можно найти в работе [5]. Наконец, приложению метода коллокации в задачах прогнозирования финансово-экономической информации посвящена монография [7].

Ниже задача коллокации излагается во взаимосвязи с постановкой и методом решения задачи машинного обучения. Это, с одной стороны, позволяет увидеть сходство и различие методов решения данных задач, а с другой стороны, выявляет возможность включение метода коллокации в арсенал методов решения задачи машинного обучения.

Вернёмся к постановке задачи машинного обучения о восстановлении неизвестной функции $f: X \to Y$. Под символом X будем теперь понимать признаковый образ множества объектов в задаче машинного обучения, так что $X \subset R_q$. В задаче коллокации о восстановлении функции $f: X \to Y$ предполагается, что

1. искомая функция $y = f(x)$ определена на ограниченном замкнутом множестве X и принадлежит некоторому гильбертову пространству H с воспроизводящим ядром $K(x, x')$, где $(x, x') \in X \times X$;
2. известны значения $(l_i)_{i=1}^n$ ограниченных линейно-независимых функционалов $(L_i)_{i=1}^n$ на искомой функции $f(x)$:

$$(2.1) \quad L_1(f) = l_1,\ L_2(f) = l_2, \dots, L_n(f) = l_n.$$

Систему (2.1) можно интерпретировать как систему уравнений наблюдений над косвенными проявлениями искомой функции $f \in H$; эту систему удобно записать компактно

$$(2.1)' \quad L(f) = l,$$

где $L = (L_1, \dots, L_n)^T$ − векторный функционал, $l = (l_1, \dots, l_n)^T \in R_n$.

Различают [5] *локальную и глобальную* задачу коллокации. В *глобальной задаче* коллокации требуется по уравнениям наблюдений (2.1) восстановить функцию $f: X \to Y$, то есть найти такую аналитически заданную функцию $\hat{f}: X \to Y$, которая приближала бы f на X и удовлетворяла бы системе уравнений наблюдений (2.1)′:

$$(2.2) \quad L(\hat{f}) = l.$$

В *локальной задаче* коллокации требуется на основании уравнений наблюдений (2.1) вычислить оценку $\hat{F}(f)$ значения $F(f)$ заданного линейного функционала F на искомой функции f. Ниже отметим, что решение локальной задачи коллокации естественно вычисляется по найденному решению глобальной задачи:

$$(2.3) \quad \hat{F}(f) = F(\hat{f}).$$

Замечание 2.1. Простейшим примером линейного функционала служит $\delta_x -$ функционал, определённый по правилу

$$(2.4) \quad \delta_x(f) = f(x).$$

Значит, если в системе (2.1) все функционалы L_i являются $\delta_{x_i} -$ функционалами, то есть, если

$$(2.5) \quad L_i(f) = \delta_{x_i}(f) = f(x_i),$$

то система уравнений наблюдений (2.1) превращается в выборку $S = (x_i, y_i = f(x_i) = l_i)_{i=1}^n$, которая используется в процессе решении задачи машинного обучения. Следовательно, задачу машинного обучения можно трактовать как частный случай задачи коллокации. Однако метод решения задачи коллокации, как будет видно ниже, радикально отличается от метода решения задачи машинного обучения.

Рассмотрим метод решения глобальной задачи коллокации, который сводится к решению системы уравнений наблюдений (2.1)′. Это означает, что глобальная задача коллокации относится к классу обратных задач [8]. Так как неизвестная функция f представляет собой элемент *бесконечномерного* гильбертова пространства H, то система уравнений (2.1)′ всегда имеет решение, причём решений бесконечное множество. По этой причине глобальная задача коллокации относится к некорректным по Адамару задачам [8].

Обозначим множество всех решений системы (2.1)′ символом $L^-(l)$, то есть

$$(2.6) \quad L^-(l) = \{\varphi : \varphi \in H, L(\varphi) = l\}.$$

Подчеркнём, что искомая функция $f \in L^-(l)$. Можно доказать, что на множестве $L^-(l)$ существует единственная функция наименьшей нормы. Такая функция называется *нормальным решением* системы (2.1)′ и именно она принимается в качестве решения $\hat{f}$ глобальной задачи коллокации. Так что

$$(2.7) \qquad \hat{f} = \underset{\varphi \in L^-(l)}{arginf}(\| \varphi \|_H).$$

Вычисление функции $\hat{f}$ базируется на следующей теореме.

Теорема [4]. Если в системе уравнений наблюдений (2.1)′ компоненты $L_1, \dots, L_n$ векторного функционала L являются ограниченными и линейно-независимыми функционалами на гильбертовом пространстве H с воспроизводящим ядром $K(x, x')$, то

1) гильбертово пространство H разлагается в прямую сумму ортогональных подпространств $KerL = \{\varphi: \varphi \in H,\ L(\varphi) = 0\}$ и $Ker^{\perp}L$:

$$(2.8)\quad H = KerL \oplus Ker^{\perp}L,$$

где бесконечномерное подпространство $KerL$ является ядром векторного функционала L, размерность подпространства $Ker^{\perp}L$ совпадает с количеством n линейно-независимых компонентов векторного функционала L, базис подпространства $Ker^{\perp}L$ образуют функции

$$(2.9)\ (K(x, L_1),\ K(x, L_2), \dots, K(x, L_n)) = K(x, L)^T\ ,$$

где символом $K(x, L_i)$ обозначен результат действия функционала L_i на воспроизводящее ядро $K(x, x')$ по аргументу $x' \in X$; так что

$$(2.10)\quad Ker^{\perp}L = Span\{K(x, L_1), K(x, L_2), \dots, K(x, L_n)\};$$

2) решение $\hat{f}$ глобальной задачи коллокации является элементом подпространства $Ker^{\perp}L$, то есть

$$(2.11)\ \hat{f} \in Ker^{\perp}L.$$

Следствие. Согласно (2.10) и (2.11), решение глобальной задачи коллокации $\hat{f}(x)$ можно представить в виде

$$(2.12)\ \hat{f}(x) = \sum_{j=1}^{n} c_j \cdot K\big(x, L_j\big) = K(x, L)^T \cdot c,$$

где $c = (c_1,\ c_2, \dots, c_n)^T -$ вектор подлежащих определению коэффициентов. Для определения коэффициентов $c = (c_1,\ c_2, \dots, c_n)^T$ подставим правую часть (2.12) в уравнения наблюдений (2.1), в итоге получим систему линейных алгебраических уравнений с положительно определённой матрицей $K = \Big(K\big(L_i, L_j\big)\Big)$:

$$(2.13)\ L_i\left(\hat{f}(x)\right) = \sum_{j=1}^{n} c_j \cdot K\big(L_i, L_j\big) = l_i \Leftrightarrow K \cdot c = l \Rightarrow\ c = K^{-1} \cdot l.$$

$$i = 1,2, \dots, n$$

Используя решение $c = K^{-1} \cdot l$ системы (2.13) в (2.12), получим сначала окончательный вид решения глобальной задачи коллокации:

$$(2.14)\ \hat{f}(x) = K(x, L)^T \cdot K^{-1} \cdot l\ ,$$

а затем и решение (2.3) локальной задачи коллокации:

$$(2.15)\ \hat{F}(f) = F(\hat{f}) = K(F, L)^T \cdot K^{-1} \cdot l.$$

Предположим теперь, что $\hat{f}(x)$ имеет смысл восстановленной функции в задаче машинного обучения, где всегда

$$(2.16)\ L_i(f) = \delta_{x_i}(f) = l_i = f(x_i) = y_i\ ,\ \text{а}\ \ F(f) = \delta_x(f) = f(x) = y.$$

Тогда, согласно (2.9) и (2.12), оценка ответа для любого объекта $x \in X$ будет вычисляться методом коллокации по правилу

$$(2.17)\ \tilde{y} = \hat{f}(x) = \sum_{j=1}^{n} c_j \cdot K\big(x, x_j\big).$$

Из сказанного выше следует, что метод коллокации способен войти в арсенал машинного обучения. Для этого необходимо располагать методикой оценки точности решения (2.17) задачи машинного обучения методом коллокации. Такая методика обсуждается ниже.

3 Оценка точности решения задачи машинного обучения методом коллокации

Реальная практика [4] метода коллокации базируется на следующей теореме о воспроизводящем ядре.

Теорема [9, стр. 512] *(М. Лоэв)*. Функция $K(x, x')$ является воспроизводящим ядром некоторого гильбертова пространства H функций, определённых на множестве X, тогда и только тогда, когда она представляет собой ковариационную функцию некоторой случайной функции $y = f(x, \omega)$, определённой на том же множестве X.

Искомую функцию в задаче машинного обучения всегда можно интерпретировать как некоторую конкретную реализацию $y = f(x, \omega_0)$ случайной функции $y = f(x, \omega)$, постулируя, ради возможности приложения, что случайная функция $y = f(x, \omega)$ является *изотропной*, то есть её ковариационная функция $K(x, x')$ есть функцией только расстояния между точками множества X:

$$(3.1)\ K(x, x') = K(\rho(x, x')).$$

Теорема Лоэва освобождает прикладника от проблемы выбора гильбертова пространства H, к которому следует отнести восстанавливаемую функцию $y = f(x, \omega_0)$; нужно лишь подобрать подходящую модель $K(\rho(x, x'))$ ковариационной функции случайной функции $y = f(x, \omega)$, располагая выборочными значениями $S = (x_i, y_i = f(x_i, \omega_0) = l_i)_{i=1}^{n}$ её единственной реализации $y = f(x, \omega_0)$.

Будем предполагать, что 1) модель (3.1) построена, 2) случайная функция $y = f(x, \omega)$ имеет нулевое математическое ожидание,

$$(3.2)\ E\big(f(x,\omega)\big) = \int_{\Omega} f(x,\omega)\, d\Omega = 0,$$

3) случайная функция $y = f(x, \omega)$ является эргодической [10, стр. 47]. Это значит, что усреднение (3.2) функции $y = f(x, \omega)$ по пространству Ω элементарных исходов $\omega \in \Omega$ совпадает с усреднением $M(\cdot)$ реализации $y = f(x, \omega_0)$ по множеству X значений x её аргумента:

$$(3.3)\ E\big(f(x,\omega)\big) = M\big(f(x,\omega_0\,)\big) = \int_{X} f(x,\omega_0)\, dX = 0.$$

Вернёмся к прогнозу (2.17) решения методом коллокации задачи машинного обучения. С учётом

(2.13), (2.14) и (2.16) уравнение (2.17) можно переписать так

$$(3.4)\ \tilde{y} = \hat{f}(x) = K(x, \vec{x})^T \cdot K^{-1} \cdot \vec{y}.$$

Здесь $\vec{x} = (x_1, \dots, x_n)^T -$ выборочные объекты из множества X. Далее,

$$(3.5)\ K(x, \vec{x})^T = \Big(K\big(\rho(x, x_1)\big), \dots, K\big(\rho(x, x_n)\big)\Big) -$$

строка ковариаций искомого значения $y(x)$ и известных выборочных значений $y_i = y(x_i)$, $\vec{y} = (y_1, \dots, y_n)^T -$ вектор выборочных значений исследуемой функции,

$$(3.6)\ K = (K\left(\rho(x_i, x_j)\right)) -$$

ковариационная матрица вектора $\vec{y}$.

Согласно следующей из теоремы М. Лоэва интерпретации искомой функции $y(x) = f(x, \omega_0)$ как реализации случайной функции, прогноз (3.4) её значения $y(x)$ является линейным несмещённым оптимальным (в среднем квадратическом относительно операции усреднения $M(\cdot)$) прогнозом Колмогорова-Винера [10, стр. 4]. При этом дисперсия $\sigma_\varepsilon^2 = M(\tilde{y} - y)^2$ истинной ошибки $\varepsilon = (\tilde{y} - y)$ прогноза (3.4) вычисляется по правилу

$$(3.7)\ \sigma_\varepsilon^2 = K(0) - K(x, \vec{x})^T \cdot K^{-1} \cdot K(x, \vec{x}).$$

Формула (3.7) служит инструментом точностных расчётов решения (2.17) методом коллокации задачи машинного обучения. Подчеркнём, что обсуждённый выше метод машинного обучения восстановления функции $y(x)$ таким инструментом не обладает.

Литература

1. *Канторович Л.В.* Об одном методе приближённого решения дифференциальных уравнений в частных производных. – ДАН СССР, 1931, 2, № 9.
2. *Воронцов К.В.* Математические методы обучения по прецедентам (теория обучения машин), http://www.ccas.ru/voron..
3. *Соловьёв В.И.* Анализ данных в экономике. КНОРУС, М., 2019.
4. *Мориц Г.* Современная физическая геодезия. – М, «Недра», 1983.
5. *Нейман Ю.М., Бывшев В.А.* Обработка результатов измерений методом коллокации. – М., МИИГАиК, 1985.
6. *Нейман Ю.М., Бывшев В.А.* Геодезические приложения основ функционального анализа. – М., МИИГАиК, 1986.
7. *Бабешко Л.О., Бывшев В.А.* Прогнозирование финансово-экономических показателей по разнородным данным. М., ООО «РУСАЙНС», 2017.
8. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. М. – «Наука», 1979.
9. *Лоэв М.* Теория вероятностей, ИЛ, М, 1962.
10. *Колмогоров А.Н.* Интерполирование и экстраполирование стационарных случайных последовательностей. Известия. АН СССР. Сер. мат. 1941, т.5, с. 3-14.