Тематический раздел журнала: Системный анализ

УДК 004.855:519.853

О.Ю. Мытник

Построение байесовской регрессии опорных векторов в характеристическом пространстве полиномиальных функций Безье-Бернштейна.

Ключевые слова: регрессия опорных векторов, байесовское умозаключение, полиномы в форме Бернштейна, характеристическое пространство, подтверждение.

Key words: support vector regression, Bayesian inference, polynomials in Bernstein form, feature space, evidence.

1. Введение

Метод опорных векторов впервые был предложен В.Н.Вапником для решения задач классификации образов [1]. Вскоре этот подход был расширен на некорректно поставленные обратные задачи восстановления регрессии по данным наблюдений, что привело к появлению регрессии опорных векторов (РОВ) [2]. Достаточно полный исторический обзор по алгоритмам опорных векторов представлен в работе А.Смолы [3]. Поскольку производительность РОВ зависит от значений параметров регуляризации (гиперпараметров), которые отражают характеристики шума в обучающей выборке, то последующие исследования посвящены преимущественно разработке методов для определения этих гиперпараметров. Так, М.Лоу и Дж.Квок в своей работе [4] применили метод байесовского подтверждения адекватности модели РОВ с гиперпараметрами, который был предложен Д.Маккеем [5] для регуляризации нейронных сетей. Основным препятствием при использовании этого метода есть недифференцируемость ϵ-нечувствительной функции потерь. С целью обеспечения требуемого уровня гладкости, как правило, предлагается использовать различные приближения, такие как мягкая нечувствительная функция потерь [6]. Однако это приводит к нарушению свойства робастности РОВ. В настоящей работе предлагается подход, с помощью которого можно сохранить робастность и получить простой критерий байесовского подтверждения адекватности модели РОВ. Другим направлением исследований РОВ есть выбор характеристического пространства. Так в работе С.Гунна [7] предложен алгоритм

SUPANOVA, где характеристическое пространство порождается составляющими разложения анализа вариаций (ANOVA). Таким образом, такие преимущества ANOVA разложения как структурированное представление зависимостей и их интерпретация использованы в РОВ. Дальнейшее развитие разложений на основе ANOVA привело к появлению нейронечетких моделей в форме Бернштейна [8], которые отличаются способностью интерпретировать функциональные зависимости на языке нечеткой логики за счет использования базисных полиномов Бернштейна как функций принадлежности. Основной задачей данной работы является использование обобщающих свойств байесовской модели РОВ с преимуществами представления нейронечетких моделей в форме Бернштейна.

2. Постановка задачи

Пусть исследуемый нелинейный процесс описывается неизвестной скалярной функцией $y(\mathbf{x})$, где $\mathbf{x} \in \mathcal{X} = \mathbb{R}^n$. Допустим, что $y(\mathbf{x})$ принадлежит некоторому семейству функций $\mathcal{M}$, которое далее будем называть моделью $\mathcal{M}$, из заданного пространства моделей $\mathcal{H}$. Модель $\mathcal{M}(\mathbf{x}, \mathbf{w}, b)$, параметризованная вектором параметров $\mathbf{w}$ и смещением b, имеет вид линейной в параметрах модели:

$$\mathcal{M}(\mathbf{x}, \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{x}) \rangle_{\mathcal{F}} + b, \quad \mathbf{w} \in \mathcal{F}, \quad b \in \mathbb{R}, \quad \mathbf{\Phi} : \mathcal{X} \to \mathcal{F}, \tag{1}$$

где $\mathbf{\Phi}$ – нелинейное отображение пространства входных переменных в некоторое характеристическое пространство $\mathcal{F}$ размерности m со скалярным произведением $\langle \cdot, \cdot \rangle_{\mathcal{F}}$. Пространство $\mathcal{F}$ также называют пространством признаков или спрямляющим пространством, так как задача нелинейной регрессии во входном пространстве сводится к линейной в характеристическом. Необходимо по данным наблюдений $\mathcal{D} = \{(\mathbf{x}_j, y_j) : j = 1, \ldots, N\}$ выбрать адекватную модель $\mathcal{M}$ из пространства $\mathcal{H}$ и найти оценки параметров модели, которые минимизируют некоторую меру погрешности между $y(\mathbf{x})$ и $\mathcal{M}(\mathbf{x}, \mathbf{w}, b)$.

3. Байесовская модель регрессии опорных векторов

Следуя М.Лоу [4] допустим, что распределение шума подчиняется закону:

$$P(\delta_j | \mathbf{w}, \epsilon, \beta, \mathcal{M}) = \frac{\beta}{2(1 + \epsilon\beta)} \exp\left(-\beta |\delta_j|_\epsilon\right), \quad \epsilon \geqslant 0, \quad \beta > 0, \tag{2}$$

где $|\cdot|_\epsilon$ – ϵ-нечувствительная функция потерь Вапника: $|\delta|_\epsilon = \max\{0, |\delta| - \epsilon\}$. Параметры ϵ и β принято называть гиперпараметрами. Предположим, что шум носит аддитивный характер:

$y_j = \mathcal{M}(\mathbf{x}_j, \mathbf{w}, b) + \delta_j$, где δ_j – независимые и одинаково распределенные случайные величины. Тогда функция правдоподобия, представляющая модель шума, имеет вид:

$$P(\mathcal{D}|\mathbf{w}, \epsilon, \beta, \mathcal{M}) = \prod_{j=1}^{N} P(\delta_j|\mathbf{w}, \epsilon, \beta, \mathcal{M}) = \left(\frac{\beta}{2(1+\epsilon\beta)}\right)^N \cdot \exp\left(-\beta \sum_{j=1}^{N} |\delta_j|_\epsilon\right).$$

Правдоподобие соответствует вероятности порождения данных $\mathcal{D}$ моделью $\mathcal{M}$ при заданных значениях $\mathbf{w}$, ϵ и β. Зададим априорное распределение вектора параметров $\mathbf{w}$ модели $\mathcal{M}$ в виде многомерного нормального распределения с нулевым средним и единичной ковариационной матрицей: $P(\mathbf{w}|\mathcal{M}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I})$. Байесовское умозаключение для выбора модели из заданного пространства моделей и определения ее параметров и гиперпараметров как правило рассматривают на трех уровнях.

3.1. Байесовское умозаключение 1-го уровня. Оценка параметров

Для модели $\mathcal{M}$ с гиперпараметрами ϵ и β, наиболее вероятные значения вектора параметров $\mathbf{w}_{mp}$, находятся из условия максимальности апостериорной вероятности параметров, которая согласно теореме Байеса равна (знак $\propto$ – прямо пропорциональна):

$$P(\mathbf{w}|\mathcal{D}, \epsilon, \beta, \mathcal{M}) \propto P(\mathcal{D}|\mathbf{w}, \epsilon, \beta, \mathcal{M}) P(\mathbf{w}|\mathcal{M}), \quad \mathbf{w}_{mp} = \arg\max_{\mathbf{w}} P(\mathbf{w}|\mathcal{D}, \epsilon, \beta, \mathcal{M}). \tag{3}$$

В противоположном логарифмическом масштабе (3) запишется в виде:

$$\mathbf{w}_{mp} = \arg\min_{\mathbf{w}} \left(-N \ln \frac{\beta}{2(1+\epsilon\beta)} + \beta \sum_{j=1}^{N} |\delta_j|_\epsilon - \frac{m}{2} \ln \frac{1}{2\pi} + \frac{1}{2}\|\mathbf{w}\|_{\mathcal{F}}^2\right). \tag{4}$$

Тут $\|\mathbf{w}\|_{\mathcal{F}}^2 = \langle \mathbf{w}, \mathbf{w} \rangle_{\mathcal{F}}$. Задача (4) в контексте статистической теории обучения эквивалентна минимизации функционала регуляризованного риска [3]:

$$R_{reg}(\mathbf{w}) = \beta N R_{emp}(\mathbf{w}) + R_{\mathbf{w}}(\mathbf{w}) = \beta \sum_{j=1}^{N} |\delta_j)|_\epsilon + \frac{1}{2}\|\mathbf{w}\|_{\mathcal{F}}^2.$$

Это классическая задача РОВ [2], которую как правило сводят к задаче выпуклой оптимизации. При этом находят двойственные переменные α_j, α_j^* (множители Лагранжа прямой задачи). Наиболее вероятный вектор параметров определяется как линейная комбинация обучающих векторов в характеристическом пространстве:

$$\mathbf{w}_{mp} = \sum_{j=1}^{N} \left(\alpha_j - \alpha_j^*\right) \mathbf{\Phi}(\mathbf{x}_j).$$

Одним из простых способов определения смещения модели b_{mp} является решение уравнения $y_j - \langle \mathbf{w}_{mp}, \mathbf{\Phi}(\mathbf{x}_j) \rangle_{\mathcal{F}} - b_{mp} = \epsilon$ при $\alpha_j \in (0; \beta)$. Элементы выборки, которые находятся

вне ϵ-полосы или на ее границах называются *опорными векторами*. Обозначим множество опорных векторов как $SV = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : |\delta_i)| \geqslant \epsilon\}$. Каждый опорный вектор имеет один ненулевой множитель Лагранжа, то есть или $\alpha_j > 0$ или $\alpha_j^* > 0$. Таким образом, модель (1) записывается в виде так называемого разложения опорных векторов (неопорные векторы имеют $\alpha_j = 0$ и $\alpha_j^* = 0$, и обнуляют соответствующие слагаемые):

$$\mathcal{M}(\mathbf{x}, \mathbf{w}_{mp}, b_{mp}) = \sum_{SV} \left(\alpha_j - \alpha_j^*\right) \langle \mathbf{\Phi}(\mathbf{x}_j), \mathbf{\Phi}(\mathbf{x}) \rangle_{\mathcal{F}} + b_{mp}. \tag{5}$$

Из разложения (5) легко видеть, что искомая функция может быть полностью описана линейной комбинацией разреженной входной выборки в характеристическом пространстве.

3.2. Байесовское умозаключение 2-го уровня. Оценка гиперпараметров

Рассмотрим теперь каким образом, имея наиболее вероятные вектор параметров и смещение модели, можно адекватно скорректировать значения гиперпараметров β и ϵ. Наиболее вероятные значения гиперпараметров находятся из условия максимальности их апостериорной плотности вероятности:

$$P(\epsilon, \beta | \mathcal{D}, \mathcal{M}) \propto P(\mathcal{D} | \epsilon, \beta, \mathcal{M}) P(\epsilon, \beta | \mathcal{M}).$$

Величина $P(\mathcal{D}|\epsilon, \beta, \mathcal{M})$ называется подтверждением адекватности (или маргинальным правдоподобием) модели с гиперпараметрами ϵ и β. Предполагая, что сами значения гиперпараметров β и ϵ равновероятны, задача сводится к максимизации подтверждения $P(\mathcal{D}|\epsilon, \beta, \mathcal{M})$, которое определяется как:

$$P(\mathcal{D} | \epsilon, \beta, \mathcal{M}) = \int_{\mathcal{F}} P(\mathcal{D} | \mathbf{w}, \epsilon, \beta, \mathcal{M}) P(\mathbf{w} | \mathcal{M}) d\mathbf{w}. \tag{6}$$

Подтверждение имеет смысл меры согласия между моделью и данными наблюдений и является именно той функцией, которую следует использовать как предпочтение для тех или иных β и ϵ [9].

3.3. Байесовское умозаключение 3-го уровня. Выбор модели

Согласно байесовскому анализу выбирают модель с максимальной апостериорной вероятностью $P(\mathcal{M}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{M})P(\mathcal{M})$. Предполагая равновероятность моделей $\mathcal{M}$ в пространстве $\mathcal{H}$, выбирают модель с наибольшим подтверждением $P(\mathcal{D}|\mathcal{M})$ равным:

$$P(\mathcal{D} | \mathcal{M}) = \iint P(\mathcal{D} | \epsilon, \beta, \mathcal{M}) P(\epsilon, \beta | \mathcal{M}) d\epsilon d\beta.$$

В предположении о равновероятности гиперпараметров максимизация подтверждения $P(\mathcal{D}|\mathcal{M})$ эквивалентна максимизации подтверждения адекватности модели с гиперпараметрами (6). Таким образом второй и третий уровни байесовского умозаключения вырождаются в один, на котором ищут максимум $P(\mathcal{D}|\epsilon, \beta, \mathcal{M})$ одновременно и по моделям и по гиперпараметрам. Поскольку в аналитическом виде функция (6) труднопредставима, то используются различные приближения [10]: метод Монте-Карло, метод Лапласа, распространение ожиданий. Известно, что метод Лапласа является самым быстрым методом приближения и, вместе с тем, он систематически переоценивает значение подтверждения. Несмотря на этот недостаток метод Лапласа все еще можно использовать при сравнении уровня адекватности моделей. Так, используя метод Лапласа, можно показать, что подтверждение допускает следующее приближение в окрестности точки $\mathbf{w}_{mp}$:

$$-\ln P(\mathcal{D}|\epsilon, \beta, \mathcal{M}) \approx R_{reg}(\mathbf{w}_{mp}) - N \ln \frac{\beta}{2(1+\epsilon\beta)} + \frac{1}{2} \ln \det \mathbf{H}_{mp}, \tag{7}$$

где $\mathbf{H}_{mp}$ – значение гессиана $\mathbf{H}$ регуляризованного риска R_{reg} в точке $\mathbf{w}_{mp}$:

$$\mathbf{H} = \nabla^2_{\mathbf{w}} R_{reg}(\mathbf{w}) = \beta N \nabla^2_{\mathbf{w}} R_{emp}(\mathbf{w}) + \mathbf{I}_m,$$

где $\mathbf{I}_m$ – единичная матрица размерности m. Наиболее вероятные значения гиперпараметров минимизируют приближение (7). Легко видеть, что $\nabla^2_{\mathbf{w}} R_{emp}(\mathbf{w}) \equiv \mathbf{0}$ везде кроме множества критических точек, для которых существует хотя бы один элемент выборки, который лежит на верхней или нижней границах ϵ-полосы: $CP = \{\mathbf{w} \in \mathcal{F} : \exists i : |\delta_i| = \epsilon\}$. В этих точках функция риска недифференцируема. Для того, чтобы избежать этой проблемы в [4] предложено приблизить ϵ-нечувствительную функцию потерь следующей гладкой функцией вида $\mathcal{C}_\eta(u) = \zeta_\eta(u - \epsilon) + \zeta_\eta(-u - \epsilon)$, где $\zeta_\eta(u) = \frac{1}{\eta}\log(1 + \exp(\eta u))$, $\eta > 0$. В работе [6] предлагается мягкая нечувствительная функция потерь. Но такие подходы приводят как к нарушению робастности моделей РОВ так и к сложным интегралам, которые не вычисляются в аналитическом виде. В следующем разделе мы покажем каким образом можно обойти проблему недостаточной гладкости функции потерь и получить простой критерий байесовского подтверждения адекватности моделей.

4. Критерий байесовского подтверждения адекватности моделей

Лемма (о локальном сглаживании)**.** *Для сколь угодно малых η-окрестностей $W_\eta(-\epsilon) = (-\epsilon - \eta, -\epsilon + \eta)$, $W_\eta(\epsilon) = (\epsilon - \eta, \epsilon + \eta)$ точек $\frac{}{5}\epsilon$, ϵ существует такая C^2 гладкая функция,*

которая совпадает с ϵ-нечувствительной функцией потерь вне этих окрестностей.

Доказательство этой леммы можно найти в приложении.

Теорема. *Для любой ϵ-нечувствительной функции потерь регрессии опорных векторов существует сколь угодно малое отклонение $\Delta\epsilon$, такое что для ϵ_1-нечувствительной функции потерь ($\epsilon_1 = \epsilon - \Delta\epsilon$) гессиан регуляризованного риска $\mathbf{H}$ тождественно равен $\mathbf{I}_m$ и при этом все элементы выборки сохраняют свойства опорности.*

Доказательство. Обозначим множество ошибок приближения как $\{|y_i - \mathcal{M}(\mathbf{x}_i)| = \delta_i | i = 1, \ldots, N\}$. Если $\nexists i : \delta_i = \epsilon$, тогда положим $\Delta\epsilon = 0$. Если существует хотя бы одно $i : \delta_i = \epsilon$, тогда выберем такое подмножество ошибок $\{\delta_j\}_j$, для которых $\delta_j < \epsilon$. Очевидно, что существует $\delta_{max} = \max\{\delta_j : \delta_j < \epsilon\} < \epsilon$. Выберем сколь угодно малое значение $\Delta\epsilon$ из интервала $(0, \epsilon - \delta_{max})$. Поскольку полоса нечувствительности уменьшается ($\epsilon_1 = \epsilon - \Delta\epsilon$), то лежавшие на границах ϵ-полосы опорные векторы остаются опорными и за пределами ϵ_1-полосы. Таким образом выбранное отклонение $\Delta\epsilon$ обеспечивает то, что не существует элементов выборки, которые лежат на верхней или нижней границах ϵ_1-нечувствительной полосы (рис. 1). От-

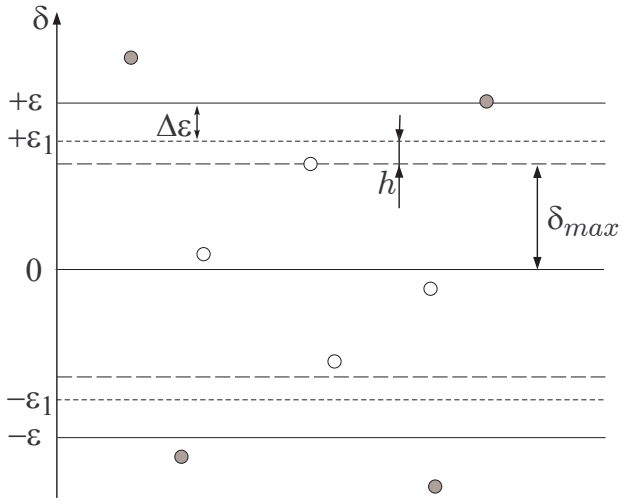

Рис. 1. Выбор отклонения $\Delta\epsilon$. Опорные векторы изображены закрашенными кругами.

сюда ясно, что $\min_i |\delta_i - \epsilon_1| = h > 0$. Выберем η такое, что $0 < \eta < h$, тогда согласно лемме существует функция S_{ϵ_1}, которая совпадает с ϵ_1-нечувствительной функцией потерь за границами окрестностей $W_\eta(\pm\epsilon_1)$. Поскольку не существует точек выборки, которые лежат в окрестностях $W_\eta(\pm\epsilon_1)$, а за пределами этих окрестностей вторая производная функции потерь равняется нулю, то можно сделать вывод о том, что $\nabla^2_{\mathbf{w}} R_{emp}(\mathbf{w}) \equiv \mathbf{0}$. Отсюда следует, что под общими условиями гессиан регуляризованного риска допустимо считать во всех

точках характеристического пространства равным единичной матрице $\mathbf{H} = \mathbf{I}_m$ поскольку для вычислительных методов можно пренебречь вариациями ширины полосы нечувствительности в пределах вычислительной точности технических средств. □

Так как $\mathbf{H}_{mp} = \mathbf{I}_m$, то непосредственно из (7) следует критерий байесовского подтверждения (КБП) адекватности модели РОВ:

$$-\ln P(\mathcal{D}|\epsilon, \beta, \mathcal{M}) \approx \text{КБП}(\epsilon, \beta, \mathcal{M}) = R_{reg}(\mathbf{w}_{mp}) - N \ln \frac{\beta}{2(1+\epsilon\beta)}. \tag{8}$$

Наиболее вероятные модель и значения гиперпараметров соответствуют минимальному значению КБП. В общем случае задача минимизации (8) это задача нелинейной оптимизации. Исследуем необходимые условия экстремума КБП.

Утверждение. *Скорость уменьшения эмпирического риска при увеличении ширины полосы нечувствительности прямо пропорциональна количеству опорных векторов.*

Доказательство. Для начала найдем производную функции потерь Вапника по ϵ:

$$(|\delta|_\epsilon)'_\epsilon = \lim_{\Delta\epsilon \to 0} \frac{|\delta|_{\epsilon+\Delta\epsilon} - |\delta|_\epsilon}{\Delta\epsilon} = -\frac{|\delta|_\epsilon}{|\delta| - \epsilon} = \begin{cases} 0, & \text{для} \quad |\delta| < \epsilon \\ -1, & \text{для} \quad |\delta| > \epsilon \end{cases}.$$

В силу теоремы считаем, что $\forall j: \ |\delta_j| \neq \epsilon$, а значит:

$$(R_{emp})'_\epsilon = \left(\frac{1}{N} \sum_{j=1}^{N} |\delta_j|_\epsilon \right)'_\epsilon = -\frac{N_{sv}}{N}.$$

где $N_{sv} = |\{(\mathbf{x}_j, y_j) \in D : |\delta_j| > \epsilon\}|$ – количество опорных векторов. □

Взяв производную в (8) относительно гиперпараметра β и приравняв ее нулю, получим первое условие:

$$-\frac{1}{\beta(1+\epsilon\beta)} + R_{emp}(\mathbf{w}_{mp}) = 0. \tag{9}$$

Взяв производную в (8) относительно гиперпараметра ϵ и приравняв ее нулю, получим второе условие:

$$\frac{N\beta}{1+\epsilon\beta} - \beta N_{sv} = 0 \Leftrightarrow N_{sv} = \frac{N}{1+\epsilon\beta}, \quad \beta \neq 0. \tag{10}$$

Необходимые условия экстремума (9-10) приводят в общем случае к системе нелинейных алгебраических уравнений (СНАР):

$$\begin{cases} -1 + \beta(1+\epsilon\beta) \cdot R_{emp}(\epsilon, \beta) & = 0, \\ -N + (1+\epsilon\beta) \cdot N_{sv}(\epsilon, \beta) & = 0. \end{cases} \tag{11}$$

Вместо минимизации (8) допустимо решать систему (11), однако это не гарантирует нахождение глобального минимума. Для решения выше указанных задач нелинейной оптимизации можно использовать например методы простой итерации, Ньютона, Левенберга-Магвердта. Отметим, что условие (10) имеет также и содержательную интерпретацию. Оно отражает априорную вероятность опорности вектора поскольку если настоящий шум действительно подчиняется закону (2), тогда вероятность того, что конкретному $\mathbf{x}$ будет отвечать значение выхода y вне ϵ-полосы нечувствительности равна:

$$p_{sv} = 1 - \int_{-\epsilon}^{\epsilon} \frac{\beta}{2(1+\epsilon\beta)} \exp(-\beta \cdot 0) d\delta = \frac{1}{1+\epsilon\beta}.$$

В работе [11] предложен интуитивный метод адаптации гиперпараметров РОВ на основе оценивания априорной и апостериорной вероятностей опорности векторов. Показана эффективность интуитивного метода на примере эталонной модели.

5. Характеристическое пространство полиномиальных функций Безье-Бернштейна

Одним из самых важных вопросов при использовании РОВ является вопрос о выборе характеристического пространства. В работе [3] рассмотрены различные типы характеристических пространств, описаны их преимущества и недостатки. С целью повышения структурированности байесовской модели РОВ в настоящей работе в качестве характеристического пространства предлагается использовать пространство полиномиальных функций Безье-Бернштейна, известных также как полиномы в форме Бернштейна. Впервые полиномы в форме Бернштейна одной и двух переменных были использованы для представления выхода нелинейной системы в виде разложения ANOVA в алгоритме построения нейронечетких моделей в форме Бернштейна [8]:

$$\mathcal{M}(\mathbf{x}, \mathbf{w}, b) = b + \sum_{k=1}^{n} B_k(x^k) + \sum_{p=1}^{n-1} \sum_{q=p+1}^{n} B_{pq}\left(x^p, x^q\right), \tag{12}$$

где вектор входа нелинейной системы $\mathbf{x} = (x^1, \ldots, x^n)^\top$. Для формирования полиномов в форме Бернштейна одной и двух переменных используются линейные комбинации соответствующих базисных полиномов Бернштейна от барицентрических координат s и $\mathbf{u}$:

$$B_k(x^k) = \sum_{j=0}^{d} \omega_j \phi_j^d\big(s(x^k)\big), \quad B_{pq}(x^p, x^q) = \sum_{i+r+t=d} \omega_{irt} \phi_{irt}^d\big(\mathbf{u}(x^p, x^q)\big),$$

где $s \in S_1$, $\mathbf{u} = (u, v)^\top \in U_2$. S_1, U_2 – пространства барицентрических координат. Базисные полиномы Бернштейна порядка d одной и двух переменных соответственно имеют вид:

$$\phi_j^d(s) = \binom{d}{j} \cdot s^j (1-s)^{d-j}, \quad \phi_{irt}^d(\mathbf{u}) = \binom{d}{i,r,t} \cdot u^i v^r (1-u-v)^t. \tag{13}$$

Барицентрические координаты в свою очередь вычисляются с помощью быстрого обратного алгоритма Кастельжо [12], который реализует отображения:

$$\Psi_k : x^k \mapsto s(x^k) \in [0;1], \qquad \Psi_{pq} : (x^p, x^q) \mapsto \mathbf{u}(x^p, x^q) \in \triangle\{u \geqslant 0, v \geqslant 0, u+v \leqslant 1\}.$$

Основным преимуществом использования такого подхода является возможность интерпретации моделей с помощью нечетких правил. При этом базисные полиномы Бернштейна суть функции принадлежности. Отметим также, что сложность базы нечетких правил квадратичная, а не экспонентная как у известного аналога ANFIS.

Определим конфигурацию характеристического пространства $\mathcal{F}$ полиномиальных функций Безье-Бернштейна как верхнетреугольную матрицу $\mathbf{C} = (c_{ij})$ размерности $n \times n$. Каждый элемент $c_{i \geqslant j} \geqslant 0$ отображает степень влияния фактора $x^{i=j}$ или пары факторов x^i, x^j на переменную выхода y и определяет степень соответствующего полинома в форме Бернштейна. Определим нелинейное отображение $\mathbf{\Phi}$ в виде:

$$\mathbf{\Phi}(\mathbf{C}) : \mathbf{x} \mapsto (\ldots, \phi_j^{c_k}(x^k), \ldots, \phi_{irt}^{c_{pq}}(x^p, x^q), \ldots)^\top, \quad \text{при этом } \langle \mathbf{x}, \mathbf{z} \rangle_{\mathcal{F}} = \sum_{i=1}^{m} x_i \bar{z}_i.$$

Это отображение индуцирует соответствующее пространство моделей в форме Бернштейна, которое обобщает разложение (12):

$$\mathcal{M}(\mathbf{x}, \mathbf{w}, b) = \mathcal{H}(\mathbf{x}, \mathbf{w}, b, \mathbf{C}) = b + \sum_{k=1}^{n} B_k^{c_k}(x^k) + \sum_{p=1}^{n-1} \sum_{q=p+1}^{n} B_{pq}^{c_{pq}}(x^p, x^q), \tag{14}$$

6. Индуктивный метод построения байесовской модели POB в форме Бернштейна

Схема алгоритма ПРИАМ (полиномиальной регрессии индуктивный алгоритм моделирования), который реализует данный метод имеет вид.

Заданы: пространство моделей в форме Бернштейна $\mathcal{H}$, данные наблюдений $\mathcal{D}$, уровень сходимости $\mu > 0$, начальная модель $\mathcal{M}^{(0)} = \mathcal{H}\left(\mathbf{C}^{(0)}\right)$, которая отвечает априорным ожиданиям и существующей информации.

Результат: субоптимальная модель $\mathcal{M}_{\text{opt}}$.

Алгоритм начинает работу с

нормировки входного и выходного пространств; итератор $t := 0$;

повторять

$\mathcal{M}_{\text{opt}} := \mathcal{M}^{(t)}$;

генерация множества моделей кандидатов: $\left\{\mathcal{M}_{ij}^{(t+1)} = \mathcal{H}\left(\mathbf{C}_{ij}^{(t+1)}\right)\right\}_{ij}$,

где $\mathbf{C}_{ij}^{(t+1)} = \mathbf{C}^{(t)} \pm \mathbf{1}_{ij}$, $1 \leqslant i \leqslant j \leqslant n$, $\mathbf{1}_{ij}$ — нулевая матрица с единицей в строке i в столбике j, что соответствует минимальному изменению конфигурации пространства;

для каждой модели кандидата $\mathcal{M}_{ij}^{(t+1)}$ **выполнить**

вычисление КБП $\left(\mathcal{M}_{ij}^{(t+1)}\right)$ с минимизацией (8) по гиперпараметрам;

конец перебора моделей кандидатов;

выбор модели с наименьшим КБП: $\mathcal{M}^{(t+1)} = \arg\left(\text{КБП}^{(t+1)} = \min \text{КБП}\left(\mathcal{M}_{ij}^{(t+1)}\right)\right)$;

$t := t + 1$;

пока не выполнится критерий останова: $\text{КБП}^{(t)} + \mu > \text{КБП}^{(t-1)}$;

конец работы алгоритма.

Определим размерность задачи обучения размерностью входного пространства n и объемом выборки данных наблюдений N. Тогда сложность ПРИАМ состоит из перебора моделей кандидатов $\mathcal{O}(n(n+1))$ и решения задачи квадратичного программирования методом активных ограничений $\mathcal{O}(N^3)$ и определяется как $\mathcal{O}(n^2N^3)$. Сходимость ПРИАМ обеспечивается глобальной и квадратичной сходимостью рефлексивного метода Ньютона для минимизации КБП по гиперпараметрам а также, в силу оценки $\text{КБП} > -N\ln(\beta_{\max}/2)$ для $\beta_{\max} < \infty$, линейной сходимостью поиска субоптимальной модели.

7. Выводы

Рассмотрена байесовская модель РОВ, основными преимуществами которой является ее независимость от размерности входного пространства, решение задачи квадратичного программирования вместо обращения плохо обусловленных матриц, способность делать статистическое обобщение на коротких выборках и давать оценки планок погрешностей при прогнозировании. Это позволяет восстанавливать зависимости с большим количеством переменных в условиях ограниченности существующей информации при сильной мультиколлинеарности и шуме. Локальное сглаживание функции потерь значительно упрощает критерий байесовского подтверждения адекватности моделей.

Разработан индуктивный метод построения байесовской модели РОВ в форме Бернштейна, где использованы преимущества как байесовской РОВ так и нейронечеткого представления в форме Бернштейна. Метод за полиномиальное время индуцирует структурированные модели с прозрачными зависимостями, которые легко можно интерпретировать с помощью нечеткой логики. Имеется возможность задавать априорную информацию о структуре модели. Практические исследования предложенного метода (ПРИАМ) на искусственных и реальных эталонных моделях нелинейной регрессии показывают его высокую продуктивность в сравнении с работой метода группового учета (МГУА) аргументов в пакете "NeuroShell 2" и рекуррентных нейронных сетей (РНС) в пакете "NeuroSolutions 5". Результаты сравнительного анализа приведены в табл. 1. Как не трудно видеть показатель среднеквадратичной ошибки (СКО) прогнозирования для алгоритма ПРИАМ в среднем меньше на 50%. Реальные и искусственные эталонные модели Longley, Filippelli, Friedman, AMPG можно найти в проекте StRD (Statistical Reference Datasets: http://www.itl.nist.gov/div898/strd/), который содержит сертифицированные эталонные модели, в проекте Delve (http://www.cs.toronto.edu/ d
и в базе данных для тестирования алгоритмов машинного обучения (http://www.ics.uci.edu/ mlearn/M
Данные по индексу потребительских цен (ИПЦ) и реальному потреблению населения (РПН) за 1998-1999 гг предоставлены Госкомстатом Украины.

Таблица 1

Сравнение нормированной СКО для различных методов

Модели	Longley	Filippelli	Friedman	AMPG	ИПЦ	РПН
ПРИАМ	0.017	0.004	0.016	0.015	0.003	0.028
МГУА	0.051	0.015	0.064	0.206	0.133	0.041
РНС	0.018	0.011	0.117	0.042	0.048	0.102

Приложение

Доказательство леммы о локальном сглаживании. Построим такую функцию S_ϵ, которая совпадает с ϵ-нечувствительной функцией потерь вне заданных окрестностей. Поскольку функция потерь симметрична, достаточно рассмотреть только неотрицательную действительную полуось. Определим везде гладкую первую производную $(S_\epsilon)'_u(u)$, в виде:

$$(S_\epsilon)'_u(u) = \begin{cases} 0, & u \in [0, \epsilon - \eta] \\ l(u), & u \in (\epsilon - \eta, \epsilon + \eta) \\ 1, & u \in [\epsilon + \eta, \infty) \end{cases},$$

где $l(u)$ – некоторая гладкая функция в окрестности $W_\eta(\epsilon)$, то есть на интервале $(\epsilon-\eta, \epsilon+\eta)$, которая удовлетворяет условиям:

$$\int_{\epsilon-\eta}^{\epsilon+\eta} l(u)du = \eta, \quad \lim_{u\to\epsilon-\eta} l(u) = 0, \quad \lim_{u\to\epsilon+\eta} l(u) = 1, \quad l'_+(\epsilon-\eta) = l'_-(\epsilon+\eta) = 0,$$

где l'_+, l'_- – производные справа и слева соответственно. Очевидно, что таким образом построенная первая производная $(S_\epsilon)'_u(u)$ обеспечивает C^2 гладкую функцию потерь S_ϵ, которая равна ϵ-нечувствительной функции потерь за пределами окрестности $W_\eta(\epsilon)$ на положительной полуоси. Поэтому для доказательства леммы найдем функцию $l(u)$ в классе логистических функций, то есть функций вида:

$$l(\pi(u)) = \frac{e^{\pi(u)}}{1+e^{\pi(u)}}. \tag{15}$$

Очевидно, что для таких функций выполняются условия: $\lim\limits_{\pi\to-\infty} l(\pi) = 0$, $\lim\limits_{\pi\to+\infty} l(\pi) = 1$. Применим дробно-линейное преобразование вида

$$\pi(u) = \frac{u-\epsilon}{\eta^2-(u-\epsilon)^2}$$

для того, чтобы конформно отобразить интервал $(\epsilon-\eta, \epsilon+\eta)$ на действительную ось $(-\infty, +\infty)$. Легко убедиться, что отображение $\pi(u)$ переводит точки $\epsilon-\eta$, ϵ, $\epsilon+\eta$ соответственно в точки $-\infty$, 0, $+\infty$. Это дает возможность записать логистическую функцию (15) переменной u в виде:

$$l(\pi(u)) = \frac{e^{\pi(u)}}{1+e^{\pi(u)}} = \frac{1}{1+e^{-\pi(u)}} = \left[1+\exp\left(\frac{u-\epsilon}{(u-\epsilon)^2-\eta^2}\right)\right]^{-1}. \tag{16}$$

В силу антисимметричности этой функции относительно точки $(\epsilon; 0.5)$, оценив площадь под графиком в η-окрестности, легко видеть, что $\int_{\epsilon-\eta}^{\epsilon+\eta} l(u)du = \int_{\epsilon}^{\epsilon+\eta} 1\cdot du = \eta$. Осталось проверить значение производных на концах η-окрестности. Производная логистической функции:

$$l'(u) = \frac{e^{-\pi(u)}}{(1+e^{-\pi(u)})^2}\cdot\pi'_u(u), \quad \pi'_u(u) = \frac{\eta^2+(u-\epsilon)^2}{[\eta^2-(u-\epsilon)^2]^2}.$$

Значение производных на концах соответственно составляют:

$$l'_+(\epsilon-\eta) = \lim_{u\downarrow\epsilon-\eta} l'(u) = 0, \quad l'_-(\epsilon+\eta) = \lim_{u\uparrow\epsilon+\eta} l'(u) = 0.$$

Таким образом, если выбрать первую производную $(S_\epsilon)'_u(u)$ с логистической функцией $l(u)$ вида (16), тогда обеспечивается существование такой функции потерь $S_\epsilon(u)$, которая совпадает с ϵ-нечувствительной функцией потерь вне окрестности W_η и имеет вид:

$$S_\epsilon(u) = \int_{-\infty}^{u} (S_\epsilon)'_t(t)dt.$$

□

Список литературы

[1] Boser B.E., Guyon I.M., Vapnik V.N. A training algorithm for optimal margin classifiers / In: Haussler D. (Ed.), Proceedings of the Annual Conference on Computational Learning Theory. ACM Press, Pitts-burgh, PA. – 1992. – P. 144–152.

[2] Vapnik V., Golowich S., Smola A. Support vector method for function approximation, regression estimation, and signal processing // Advances in Neural Information Processing Systems, Cambridge, MA. MIT Press. – 1997. – Vol. 9. – P. 281–287.

[3] Smola A.J., Schölkopf B. A tutorial on support vector regression // Statistics and Computing. – 2004. – №14. – P. 199–222.

[4] Law M.H., Kwok J.T. Applying the Bayesian evidence framework to ν-support vector regression / Proceedings of the European Conference on Machine Learning (ECML), Freiburg, Germany, September 2001. – P. 312–323.

[5] Mackay D. A practical Bayesian framework for backprop networks // Neural Computation. – 1992. – Vol. 4. – P. 448–472.

[6] Chu W., Keerthi S., Ong C. J. Bayesian support vector regression using a unified loss function // IEEE Transactions on Neural Networks. – 2004. – Vol. 15, №1. – P. 29–44.

[7] Gunn S.R., Brown M. SUPANOVA – a sparse, transparent modelling approach / In Proceedings of IEEE International Workshop on Neural Networks for Signal Processing, Madison, Wisconsin. – 1999. – P. 21–30.

[8] Hong X., Harris C.J. Generalized neurofuzzy network modeling algorithms using Bézier-Bernstein polynomial functions and additive decomposition // IEEE Transactions Neural Networks. – 2000. – Vol. 11, №4. – P. 889–902.

[9] Mackay D. Bayesian methods for adaptive models. Dissertation. California Institute of Technology. Pasadena. – 1992. – 98 p.

[10] Kuss M., Rasmussen C.E. Assessing approximate inference for binary gaussian process classification // Journal of Machine Learning Research. – 2005. – №6. – P. 1679–1704.

[11] Мытник О.Ю. Локальное сглаживание ϵ-нечувствительной функции потерь в байесовской модели опорных векторов / Сборник трудов VI международной научной конференции “Интеллектуальный анализ информации”, 16 – 19 мая 2006 года, г. Київ. – С. 189–198.

[12] Митник О.Ю., Бідюк П.І. Обернене відображення Кастельжо в нечітких нейронних моделях // Системні дослідження та інформаційні технології. – 2004. – №2. – С. 24–34.

Аннотации

Мытник О.Ю. Построение байесовской регрессии опорных векторов в характеристическом пространстве полиномиальных функций Безье-Бернштейна.

Исследуются некорректно поставленные обратные задачи восстановления нелинейных зависимостей по данным наблюдений с помощью метода опорных векторов. Разработан индуктивный метод построения байесовской модели регрессии опорных векторов в форме Бернштейна. Для сравнения уровня адекватности моделей в форме Бернштейна используется новый критерий байесовского подтверждения.

Митник О.Ю. Побудова байєсівської регресії опорних векторів в характеристичному просторі поліноміальних функцій Без'є-Бернштейна.

Досліджуються некоректно поставлені обернені задачі відновлення нелінійних залежностей за даними спостережень за допомогою методу опорних векторів. Розроблено індуктивний метод побудови байєсівської моделі регресії опорних векторів в формі Бернштейна. Для порівняння рівня адекватності моделей в формі Бернштейна використовується новий критерій байєсівського підтвердження.

Mytnyk O.Yu. Construction of Bayesian support vector regression in feature space spanned by Bézier-Bernstein polynomial functions.

The ill-posed inverse problems for recovery of non-linear dependencies based on observational data are under consideration. The inductive construction method of Bayesian model of support vector regression in Bernstein form is developed. New Bayesian evidence criterion for model adequacy comparison is used.