Анализ свойств ансамбля локально аппроксимирующих моделей

Р. И. Исламов[1]***, А. В. Грабовой***[1]***, В. В. Стрижов***[1]
islamov.ri@phystech.edu; grabovoy.av@phystech.edu; strijov@ccas.ru
[1]Московский физико-технический институт

Данная работа посвящена построению универсальной модели в виде ансамбля локальных моделей. Для решения задачи регрессии предлагается использовать многоуровневый подход. Множество объектов выборки разбивается на несколько подмножеств, каждому подмножеству ставится в соответствие одна локальная модель, оптимально аппроксимирующая данное подмножество. Для аппроксимации генеральной выборки строится универсальный аппроксиматор, который представлен в виде ансамбля локальных моделей. В качестве коэффициентов шлюзовой функции используется выпуклая комбинация локальных моделей, значение которой зависит от объекта, для которого производится предсказание. Ансамблевый подход позволяет описывать те выборки, которые затруднительно описать одной моделью. Для анализа свойств проводится вычислительный эксперимент. В качестве данных используются синтетические и реальные выборки.

Ключевые слова: *локальная модель; линейные модели; ансамбль моделей.*

1 Введение

В прикладных задачах данные порождены несколькими источниками. В таких случаях качество предсказания можно повышать, используя несколько моделей. Если моделей на самом деле меньше, чем нужно, то веса лишних моделей будут малы и их вклад будет несущественен. Преимуществом ансамбля является его способность описывать те выборки, которые затруднительно описывать одной моделью.

В данной работе исследуется проблема построения ансамбля локальных моделей. *Локальная модель* — модель, которая аппроксимирует объекты, находящиеся в одной связной области в пространстве объектов. В качестве решающей функции используется выпуклая комбинация локальных моделей, при этом веса локальных моделей не постоянны, а зависят от положения объекта в пространстве объектов.

В данной работе предполагается, что вклад каждой локальной модели в целевую переменную зависит от рассматриваемого объекта. Ансамбль локальных моделей использует шлюзовую функцию, которая определяет значимость предсказания каждой локальной модели, входящей в ансамбль.

В данной работе каждая локальная модель является линейной. В качестве функции ошибки используется логарифм правдоподобия модели. Оптимальные параметры ансамбля и локальных моделей находятся при решении двухуровневой задачи оптимизации. На первом шаге оптимизируются параметры локальных моделей при фиксированных параметрах шлюзовой функции, на втором шаге оптимизируется параметры шлюзовой при найденных фиксированных параметрах локальных моделей.

Алгоритмы тестировались на синтетических и реальных данных. Эксперименты показали преимущество использования ансамбля локальных моделей по сравнению с использованием одной модели.

Ансамблевый подход стал предметом многих исследований. Ансамбль моделей использовался в работах [1], [2], [3]. В работе [4] представлен обзор методов и моделей в задачах ансамбля моделей. В данной работе представлены виды шлюзовых функций. Приведен анализ разных моделей, которые могут выступать в качестве локальной модели.

Ансамблевый подход имеет множество приложений в прикладных задачах. В работе [5] предложен метод распознавания рукописных цифр. Метод распознания текстов при помощи ансамбля локальных моделей исследуется в работах [6], а для распознавания речи — в [7], [8]. В работе [9] исследуется смесь экспертов для задачи распознавания трехмерных движений человека.

Были предложены различные типы локальных моделей, такие как SVM [10], Гауссовский процесс [11] и нейронные сети [12]. Другие работы была сосредоточены на различных конфигурациях, таких как иерархическая структура [13], бесконечное число экспертов [14] и последовательное добавление экспертов [15]. [16] предлагает модель ансамбля локальных моделей для машинного перевода. Стробирующая сеть обучается на предварительно обученной модели NMT ансамбля.

2 Постановка задачи построения ансамбля локальных моделей

Задано множество объектов Ω, признаки которых описываются матрицей

$$\mathbf{X} \in \mathbb{R}^{N \times n}, \tag{2.1}$$

где N — число объектов во множестве, а n — размерность признакового пространства. Каждому объекту ω_i из Ω соответствует признаковое описание $\mathbf{x}_i \in \mathbb{R}^n$, которое является i-ой строкой матрицы $\mathbf{X}$, и значение целевой переменной $y_i \in \mathbb{R}$. Введем выборку данных $\mathfrak{D}$:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i) \mid i \in \overline{1, N}\}. \tag{2.2}$$

В данной работе предполагается, что выборка $\mathfrak{D}$ порождена K источниками. Это предположение индуцирует разбиение множества индексов $I = \{1, 2, \ldots, N\}$ на K непересекающихся подмножеств I_k:

$$I = \bigcup_{k=1}^{K} I_k. \tag{2.3}$$

Разбиение индексного множества I индуцирует разбиение множества объектов Ω на подмножества Ω_k

$$\Omega = \bigcup_{k=1}^{K} \Omega_k, \qquad \Omega_k = \{\omega_i \in \Omega \mid i \in I_k\} \tag{2.4}$$

и выборку $\mathfrak{D}$ на подвыборки $\mathfrak{D}_k$

$$\mathfrak{D} = \bigcup_{k=1}^{K} \mathfrak{D}_k, \qquad \mathfrak{D}_k = \{(\mathbf{x}_i, y_i) \in \mathfrak{D} \mid i \in I_k\} \tag{2.5}$$

Для каждого подмножества объектов Ω_k используется своя локальная модель.

Определение 1. *Модель $\mathbf{g}_k$ называется локальной, если она аппроксимирует аппроксимирует подвыборку $\mathfrak{D}_k = \{(\mathbf{x}_i, y_i) \in \mathfrak{D} \mid i \in I_k\}$.*

В данной работе локальный модели объединены в ансамбль локальных моделей.

Определение 2. *Ансамбль локальных моделей — мультимодель, определяющая правдоподобие веса π_k каждой локальной модели $\mathbf{g}_k$ на признаковом описании объекта $\mathbf{x}$.*

$$\mathbf{f} = \sum_{k=1}^{K} \pi_k \mathbf{g}_k(\mathbf{x}, \mathbf{w}_k), \qquad \pi_k\left(\mathbf{x}, \mathbf{V}\right) : \mathbb{R}^{n\times|\mathbf{V}|} \to [0,1], \qquad \sum_{k=1}^{K} \pi_k\left(\mathbf{x}, \mathbf{V}\right) = 1, \tag{2.6}$$

где $\mathbf{f}$ — ансамбль локальных моделей, $\mathbf{g}_k$ — локальная модель, π_k — шлюзовая функция, $\mathbf{V}$— параметры шлюзовой функции.

В работе в качестве локальной модели $\mathbf{g}_k$ используется линейная модель, а в качестве шлюзовой функции $\boldsymbol{\pi}$ используется двухслойная нейросеть:

$$\mathbf{g}_k(\mathbf{x}, \mathbf{w}_k) = \mathbf{w}_k^{\mathsf{T}}\mathbf{x}, \qquad \boldsymbol{\pi}\left(\mathbf{x}, \mathbf{V}\right) = \operatorname{softmax}\Big(\mathbf{V}_1^{\mathsf{T}}\sigma\left(\mathbf{V}_2^{\mathsf{T}}\mathbf{x}\right)\Big), \tag{2.7}$$

где $\mathbf{V} = \{\mathbf{V}_1, \mathbf{V}_2\}$ — параметры шлюзовой функции, $\sigma(x)$ — сигмоидная функция. Введем понятие расстояние между двумя объектами.

Определение 3. *Расстоянием между двумя объектами ω_1 и ω_2 из Ω называется число, равное расстоянию между векторами $\mathbf{x}_1, \mathbf{x}_2$ признаковых описаний этих объектов:*

$$\rho(\omega_1, \omega_2) = ||\mathbf{x}_1 - \mathbf{x}_2||_2. \tag{2.8}$$

Будем считать, что каждый объект имеет признаковое описание $\mathbf{x}$, взятый из некоторого вероятностного распределения. Пусть этому распределению соответствует вероятностная мера $\mathsf{F}(x)$ в пространстве признаков. Тогда пространство локальных моделей является гильбертовым пространством, в котором введено скалярное произведение.

Определение 4. *Скалярным произведением между двумя локальными моделями $\mathbf{g}_i$ и $\mathbf{g}_j$ называется число, вычисляемое по формуле*

$$\langle \mathbf{g}_i, \mathbf{g}_j \rangle = \operatorname{cov}\Big(\mathbf{g}_i(\mathbf{x}, \mathbf{w}_i), \mathbf{g}_j(\mathbf{x}, \mathbf{w}_j)\Big) = \mathsf{E}\Big(\mathring{\mathbf{g}}_i\left(\mathbf{x}\cdot\mathbf{w}_i\right)\cdot\mathring{\mathbf{g}}_j\left(\mathbf{x}, \mathbf{w}_j\right)\Big), \tag{2.9}$$

где $\mathsf{E}\left(\mathring{\xi}\cdot\mathring{\eta}\right) = \int\limits_{\mathbb{R}^n} \mathring{\xi}\left(\mathbf{x}\right)\mathring{\eta}\left(\mathbf{x}\right)\mathsf{dF}(\mathbf{x})$ — математическое ожидание произведения центрированных случайных величин $\xi(\mathbf{x})$ и $\eta(\mathbf{x})$.

Данное определение можно интерпретировать следующим образом: чем ближе скалярное произведение к нулю, тем дальше локальные модели друг от друга в пространстве моделей, и наоборот.

Для нахождения оптимальных параметров мультимодели используется функция ошибки следующего вида:

$$\mathcal{L}\left(\mathbf{V}, \mathbf{W}\right) = \sum_{(\mathbf{x},y)\in\mathcal{D}} \sum_{k=1}^{K} \pi_k\left(\mathbf{x}, \mathbf{V}\right)\left(y - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}\right)^2 + R\left(\mathbf{V}, \mathbf{W}\right), \tag{2.10}$$

где $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k]$ — параметры локальных моделей, $R\left(\mathbf{V}, \mathbf{W}\right)$ — регуляризация параметров. Оптимальные параметры определяются из выражения

$$\hat{\mathbf{V}}, \hat{\mathbf{W}} = \arg\min_{\mathbf{V},\mathbf{W}} \mathcal{L}(\mathbf{V}, \mathbf{W}). \tag{2.11}$$

В качестве базового алгоритма используется EM-алгоритм. Алгоритм подробно описан в [17]. Формулы, по которым происходит оптимизация функции ошибки (2.11), приведены в работе [18].

3 Вычислительный эксперимент

3.1 Постановка проблемы

Данный эксперимент ставится для того, чтобы показать, что одна линейная модель плохо аппроксимирует выборку, объекты которой порождены несколькими источниками.

В качестве данных используются синтетическая выборка. Рассмотрим две подвыборки объектов, имеющих по одному признаку x^k, описываемых линейной моделью с нормальным шумом:

$$\mathbf{y}_k = \alpha_k \mathbf{x}^k + \boldsymbol{\varepsilon}, \qquad k \in \{1, 2\}, \qquad x_k, y_k \in \mathbb{R}, \qquad \varepsilon \in \mathcal{N}(0, 1). \tag{3.1}$$

В качестве общей выборки рассматривается конкатенация двух подвыборок, описываемая вектором целевой переменной $\mathbf{y}$ и матрицей признаков $\mathbf{X}$:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x}^1 & \mathbf{0} \\ \mathbf{0} & \mathbf{x}^2 \end{pmatrix}. \tag{3.2}$$

На общей выборке $\mathbf{X}$ обучается линейная модель. Линейная модель хорошо аппроксимирует данную выборку (см. рис. 1а).

Во втором эксперименте две подвыборки сливаются в одну общую выборку, описываемая целевой переменной $\mathbf{y}$ и матрицей признаков $\hat{\mathbf{X}}$:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \qquad \hat{\mathbf{X}} = \begin{pmatrix} \mathbf{x}^1 & \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 & \mathbf{x}^2 \end{pmatrix}, \tag{3.3}$$

где $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in \mathcal{N}(0, 1)$. На построенной модели также обучается линейная модель. В этом случае линейная модель плохо аппроксимирует данную выборку (см. рис. 1б).

Данный эксперимент показывает, что для аппроксимации выборки, порожденной несколькими источниками, одна модель не подходит. Таким образом, при построение модели для обучения нужно учитывать гипотезу порождения данных. Нередко оказывается, что данные порождены несколькими источниками. В этом случае для лучшей аппроксимации можно использовать ансамбль локальных моделей, где каждая локальная модель обрабатывает свою область признакового пространства (в одной области объекты имеют схожие признаки, объекты из разных областей имеют разные признаковые описания).

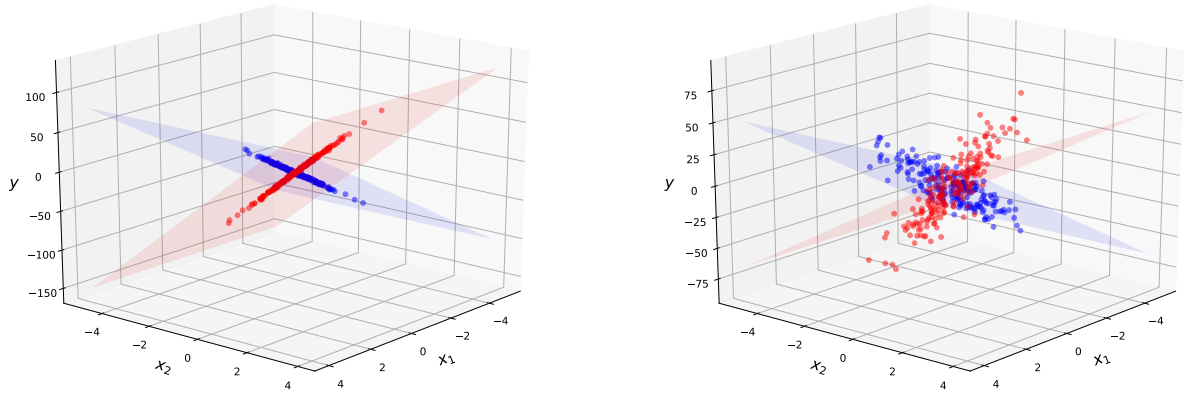
3.2 Решение проблемы при помощи ансамбля локальных моделей

Данный эксперимент ставится для того, чтобы показать, что ансамбль локальных моделей хорошо аппроксимирует выборку, порожденную несколькими источниками.

В данном эксперименте для аппроксимации используется ансамбль двух линейных локальных моделей.

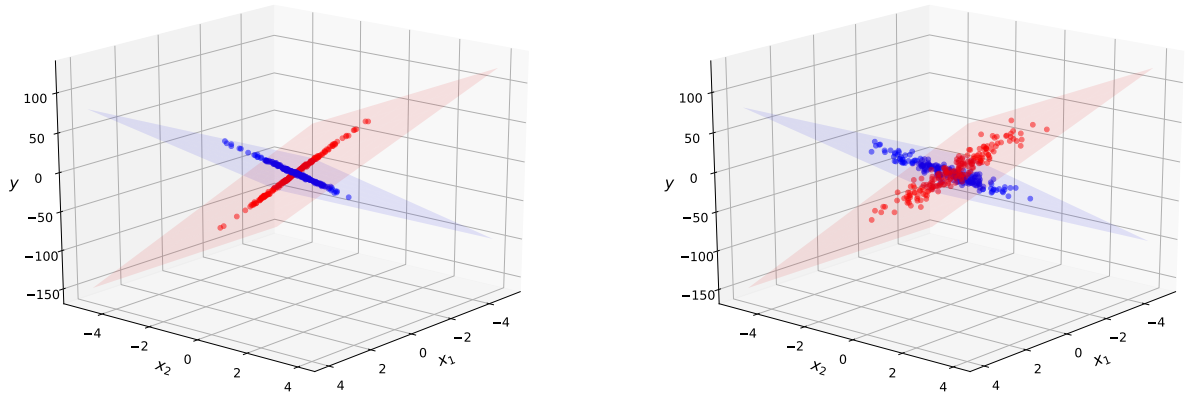
В первом эксперименте ансамбль обучается на выборке $(\mathbf{y}, \mathbf{X})$, а во втором — на выборке $(\mathbf{y}, \hat{\mathbf{X}})$. Ансамбль двух линейных локальных моделей хорошо аппроксимирует выборку (см. рис. 2 а, б).

Данный эксперимент показывает, что для аппроксимации выборки, порожденной несколькими источниками, подходит ансамбль локальных моделей. Качество аппроксимации ансамбля моделей выше, чем при использовании лишь одной модели.



а) б)

Рис. 1 а) Признаки, соответствующие другому подмножеству, заполнялись нулями, б) признаки, соответствующие другому подмножеству, заполнялись случайными числами. Точки соответствуют правильным ответам, плоскости задают предсказание линейной модели для каждого из подмножеств.



а) б)

Рис. 2 а) Признаки, соответствующие другому подмножеству, заполнялись нулями, б) признаки, соответствующие другому подмножеству, заполнялись случайными числами. Точки соответствуют правильным ответам, плоскости задают предсказание мультимодели для каждого из подмножеств.

3.3 Анализ ансамбля моделей в зависимости от уровня шума

Синтетические данные. В качестве данных используются синтетические данные. Используется две подвыборки, каждая из которых описывается линейной моделью с нормальным шумом:

$$\mathbf{y}_k = \alpha_k \mathbf{x}^k + \boldsymbol{\varepsilon}, \qquad k \in \{1, 2\}, \qquad x_k, y_k \in \mathbb{R}, \qquad \varepsilon \in \mathcal{N}(0, 1). \tag{3.4}$$

В качестве общей выборки рассматривается конкатенация двух подвыборок, описываемая вектором целевой переменной $\mathbf{y}$ и матрицей признаков $\hat{\mathbf{X}}$:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \qquad \hat{\mathbf{X}} = \begin{pmatrix} \mathbf{x}^1 & \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 & \mathbf{x}^2 \end{pmatrix}, \tag{3.5}$$

где $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in \mathcal{N}(0, \sigma)$. На общей выборке обучается ансамбль из двух локальных моделей, каждая из которых является линейной. Исследуется зависимость введенного расстояния (2.9) от параметра шума σ. График представлен на рисунке 3:

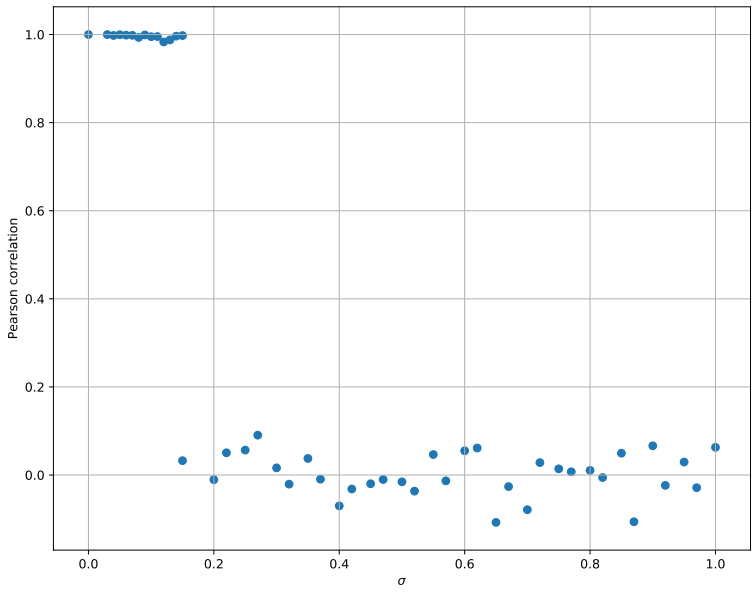


Рис. 3 График зависимости расстояние между локальными моделями от параметра шума σ для синтетических данных.

На графике видно, что при малом параметре шума σ (меньшем, чем пороговое значение) локальные модели близки, корреляция между ними приблизительно равна единице. Это означает, что при малом параметре σ шум практически не влияет на данные, для аппроксимации выборки достаточно одной линейной модели, поэтому параметры двух локальных моделей становятся приблизительно одинаковыми.

На графика также видно, что при параметре шума σ большем, чем пороговое значение, локальные модели становятся независимыми друг от друга, корреляция между ними приблизительно равна нулю. Это означает, что шумовые признаки сказываются на данных и для аппроксимации выборки необходимы две модели.

Данные на основе датасэтов Boston housing и Servo. Выборки Boston housing и Servo описываются матрицами признаков $\mathbf{X}_b \in \mathbb{R}^{506\times 13}$ и $\mathbf{X}_s \in \mathbb{R}^{167\times 4}$, а также векторами целевой переменной $\mathbf{y}_b \in \mathbb{R}^{506}$ и $\mathbf{y}_s \in \mathbb{R}^{167}$. В качестве общей выборки рассматривается конкатенация выборок Boston housing и Servo, описываемая вектором целевой переменной $\tilde{\mathbf{y}}$ и $\tilde{\mathbf{X}}$:

$$\tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_b \\ \mathbf{y}_s \end{pmatrix} \in \mathbb{R}^{673}, \qquad \tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_b & \\ \mathbf{X}_s & \mathcal{E} \end{pmatrix} \in \mathbb{R}^{673\times 13}, \tag{3.6}$$

где $\mathcal{E}$ — матрица размера 167×9, каждый элемент которой из $\mathcal{N}(0, \sigma)$. На общей выборке обучается ансамбль из двух локальных моделей, каждая из которой является линейной. Исследуется зависимость расстояния между локальными моделями от параметра шума σ. График представлен на рисунке 4:

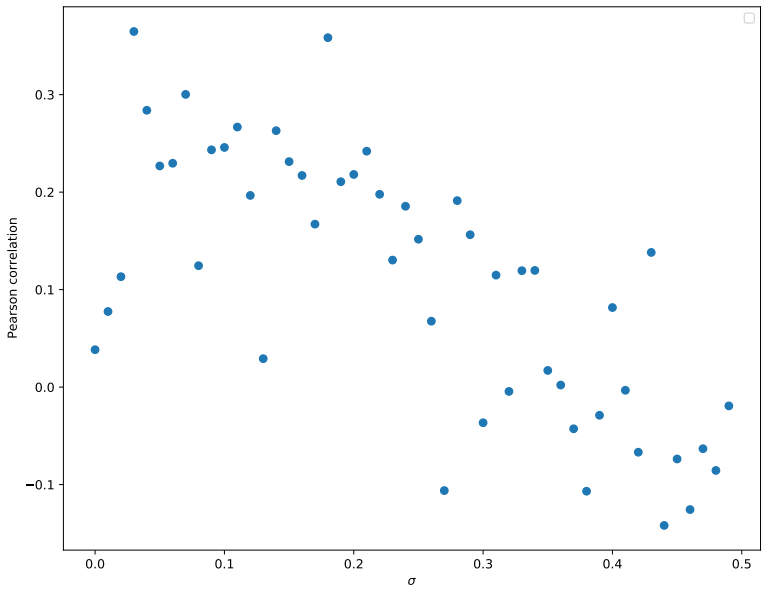


Рис. 4 График зависимости расстояние между локальными моделями от параметра шума σ для выборок Boston housing и Servo.

На графике видно, что при увеличении параметра шума σ есть тенденция к уменьшению скалярного произведения между моделями. Это означает, с увеличением шума локальные модели отдаляются друг от друга.

Исследуем качество аппроксимации при помощи ансамбля двух линейных локальных моделей. Ошибка аппроксимации вычисляется по формуле

$$error = \sum_{k=1}^{2} \sum_{i=1}^{N} \pi_k(\mathbf{x}_i, \mathbf{w}^k)(y_i^{pred_k} - y_i^{real})^2, \tag{3.7}$$

где $N = 673$ — количество объектов в общей выборке, $y_i^{pred_k}$ — предсказанное значение для i-го объекта k-ой моделью, а y_i^{real} — значение целевой переменной i-го объекта. График зависимости ошибки аппроксимации от параметра шума представлен на рисунке 5:

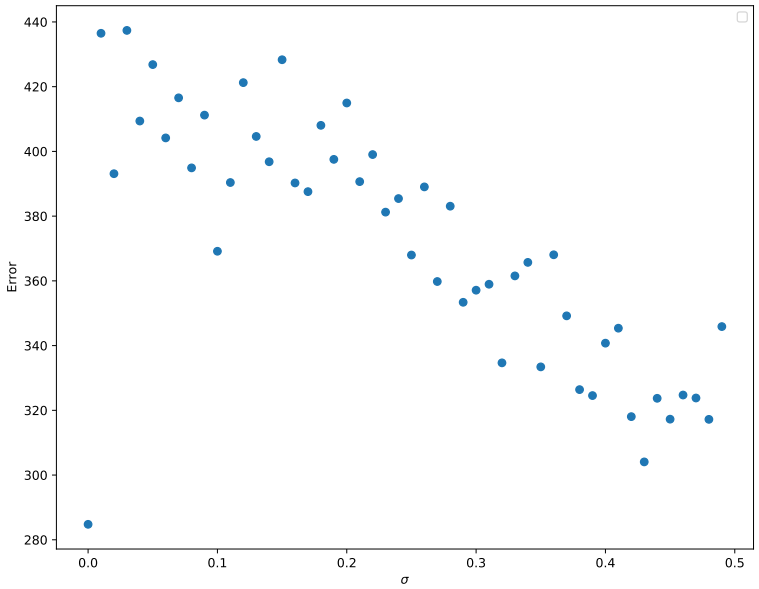


Рис. 5 График зависимости ошибки аппроксимации от параметра шума σ для выборок Boston housing и Servo.

Литература

[1] M. Serdar Yumlu, Fikret S. Gurgen, and Nesrin Okay. Financial time series prediction using mixture of experts. In *Computer and Information Sciences - ISCIS 2003*, pages 553–560. Springer Berlin Heidelberg, 2003.

[2] Yiu-ming Cheung, Wai Leung, and Lei Xu. Application of mixture of experts model to financial time series forecasting. October 1995.

[3] Andreas S. Weigend and Shanming Shi. Predicting daily probability distributions of s&p500 returns. *Journal of Forecasting*, 19(4):375–392, 2000.

[4] S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, aug 2012.

[5] Reza Ebrahimpour, Mohammad Moradian, Alireza Esmkhani, and Farzad Jafarlou. Recognition of persian handwritten digits using characterization loci and mixture of experts. *JDCTA*, 3:42–46, January 2009.

[6] Andrew Estabrooks and Nathalie Japkowicz. A mixture-of-experts framework for text classification. In *Proceedings of the 2001 workshop on Computational Natural Language Learning*. Association for Computational Linguistics, 2001.

[7] S. Iman Mossavat, Oliver Amft, Bert de Vries, Petko N. Petkov, and W. Bastiaan Kleijn. A bayesian hierarchical mixture of experts approach to estimate speech quality. In *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, jun 2010.

[8] Fengchun Peng, Robert A. Jacobs, and Martin A. Tanner. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91(435):953–960, sep 1996.

[9] Cristian Sminchisescu, Atul Kanaujia, and Dimitris N. Metaxas. BM^3e : Discriminative density propagation for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):2030–2044, nov 2007.

[10] Ronan Collobert, Samy Bengio, and Yoshua Bengio. A parallel mixture of SVMs for very large scale problems. *Neural Computation*, 14(5):1105–1114, may 2002.

[11] Volker Tresp. Mixtures of gaussian processes. In *Advances in Neural Information Processing Systems 13*, pages 654–660. MIT Press, 2001.

[12] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.

[13] Michael I. Jordan and Robert A. Jacobs. Hierarchies of adaptive experts. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 985–992. Morgan-Kaufmann, 1992.

[14] Carl Edward Rasmussen and Zoubin Ghahramani. Infinite mixtures of gaussian process experts. In *In Advances in Neural Information Processing Systems 14*, pages 881–888. MIT Press, 2001.

[15] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Lifelong learning with a network of experts.

[16] Ekaterina Garmash and Christof Monz. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[17] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* SPRINGER NATURE, 2011.

[18] V. V. Strijov A. V. Grabovoy. Prior distribution choices for a mixture of experts. *Machine learning and data analysis*, 2020.

Поступила в редакцию

На графике видно, что с ростом параметра шума σ качество аппроксимации уменьшается.