© 2021 г. А.Ю. ПОПКОВ, канд. техн. наук (apopkov@isa.ru)
(Федеральный исследовательский центр "Информатика и управление" РАН, Москва)

РАНДОМИЗИРОВАННОЕ МАШИННОЕ ОБУЧЕНИЕ НЕЛИНЕЙНЫХ МОДЕЛЕЙ С ПРИМЕНЕНИЕМ К ПРОГНОЗИРОВАНИЮ РАЗВИТИЯ ЭПИДЕМИЧЕСКОГО ПРОЦЕССА[1]

Развивается дискретный подход в теории рандомизированного машинного обучения, ориентированный на применение к нелинейным моделям. Формулируется задача энтропийного оценивания распределений вероятностей и шумов измерений для дискретных нелинейных моделей. Рассматриваются вопросы, связанные с применением таких моделей к задачам прогнозирования, в частности проблеме генерации энтропийно-оптимальных распределений. Демонстрация предложенных методов проводится на решении задачи прогнозирования общего количества инфицированных SARS-CoV-2 в Германии в 2020 г.

Ключевые слова: рандомизированное машинное обучение, энтропия, энтропийное оценивание, прогнозирование, рандомизированное прогнозирование, COVID-19, SARS-CoV-2.

DOI: 10.31857/S0005231021060064

1. Введение

Большое количество различных процессов, наблюдаемых в разных областях человеческой деятельности, природе и т.п., не могут быть эффективно описаны линейными математическими моделями. В этой связи разработка и развитие общих подходов нелинейного моделирования являются актуальными задачами. Однако необходимо отметить, что разработка и применение нелинейных моделей для конкретных задач сопряжены с определенными трудностями, связанными как с их обучением с использованием реальных данных, так и с выбором структуры модели.

Машинное обучение как раздел прикладной науки интегрирует в себе большое количество методов и подходов, накопленных в различных научных дисциплинах [1, 2], большой вклад в которые был сделан в таких направлениях, как теория вероятностей и математическая статистика [3, 4]. Методы машинного обучения успешно применяются к различным задачам, в частности к задачам классификации и регрессии, которые относятся к задачам обучения с учителем [1], основной особенностью которых является идея настройки (обучения) параметров требуемой модели с использованием реальных данных. Настроенную (обученную) модель предполагается использовать

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 20-07-00683).

для прогнозирования, т.е. получать от нее ответ, предъявляя ей входные данные, не участвующие ранее в обучении.

Этот подход является эффективным и известен, по крайней мере, с 1960-х гг. [5–8]. Однако большинство подходов, разработанных в данной области, ориентировано на использование линейных моделей. В частности, одним из эффективных подходов к решению задач классификации является линейная классификация, состоящая в поиске линейной разделяющей гиперплоскости и применяемая, например, в широко используемом методе опорных векторов [9]. Задачи восстановления регрессии давно и широко применяются в эконометрике [3], и также большинство из них в этой области ориентировано на восстановление линейной регрессии. Основные причины такого положения состоят в основном в том, что, во-первых, линейные модели легче исследовать и интерпретировать, во-вторых, численные и аналитические решения линейных задач возможно получать либо абсолютно точно (аналитически), либо с высокой точностью численно и, в-третьих, многие практические задачи часто можно свести к линейным постановкам, а следовательно, получать более качественное их решение с учетом отмеченных ранее свойств.

В то же время в некоторых прикладных задачах классификации и регрессии проявляются различные нелинейные эффекты, которые необходимо каким-то образом решать. Эффективным подходом к этой проблеме является, например, ядерный подход, состоящий в нелинейном переходе в пространство высокой размерности с последующим применением в нем уже линейного метода для решения задачи классификации или регрессии [1]. Этот подход демонстрирует свою эффективность во многих прикладных задачах и приводит к появлению различных «ядерных» версий известных линейных методов. Тем не менее часто вопрос о выборе модели (точнее, ядерной функции) продолжает возникать при применении этих методов на практике. Кроме этого, всякий переход в пространство высокой размерности при отсутствии большого количества данных в этом пространстве неизбежно приводит к нежелательным эффектам таким, например, как переобучение, а также ряду других.

Другим подходом борьбы с нелинейными эффектами предлагают методы, в которых не выделяется модель в явном виде, например методы, основанные на деревьях решений [10], нейросетевые модели [1, 11, 12] и ряд других.

Таким образом, с одной стороны, наблюдается ситуация, когда в практических задачах анализа данных явным образом наблюдается наличие “нелинейности”, требующей применения нелинейных моделей, с другой стороны, существующие методы, части используемые на практике, недостаточно развиты для эффективного решения задачи в нелинейном случае и требуют переформулирования или какой-либо адаптации задачи к их применению.

В настоящей работе предлагается универсальный подход к работе с нелинейными моделями в задачах анализа данных, в частности в задачах обучения регрессионных моделей. Этот подход основан на теории рандомизированного машинного обучения [13], основная идея которой состоит в искусственной рандомизации параметров модели, что позволяет перейти от модели с детерминированными параметрами к модели со случайными параметрами и определять в результате обучения не их точечные оценки, а их распреде-

ления. Распределения определяются таким образом, чтобы они доставляли максимум энтропийному функционалу при условии баланса со средним выходом модели.

Основным достоинством предлагаемого подхода является независимость от реальных характеристик используемых данных. Для корректного применения метода не требуется подтверждения или предположения о нормальности данных (или иных их вероятностных свойствах), а полученные в результате обучения распределения вычислены в условиях максимальной энтропии, таким образом отражая наиболее "плохой" сценарий развития исследуемого процесса, соответствующий его максимальной неопределенности. Данные свойства метода энтропийного оценивания восходят к работам Больцмана [14], Джейнса [15, 16], Шеннона [17]. Еще одной важной особенностью метода является получение энтропийно-оптимальных распределений шумов, содержащихся в данных, вместе с оптимальными распределениями параметров. Это свойство существенно отличает метод от классических подходов, в которых делаются различные предположения о характеристиках шумов и данных.

Оценки характеристик модели используются для прогнозирования моделируемого процесса. Стандартный подход к прогнозированию состоит в применении модели с точечными оценками параметров, полученных оцениванием по реальным данным, для неизвестных ("будущих") точек данных [18–20].

С учетом того, что подобрать модель точно под данные не представляется возможным, вводится предположение о стохастической природе данных, точнее о том, что механизм порождения наблюдаемых данных, который неизвестен, содержит стохастическую компоненту. Следствием этого предположения становится то, что в наблюдаемых данных есть как детерминированная, так и стохастическая компонента, вероятностные характеристики которой неизвестны. Фактически, на моделирование этой стохастической компоненты и направлен весь разработанный к настоящему времени математический аппарат, применяемый в данной области. Существенную часть этого подхода составляет предположение о нормальности случайных компонент данных, из которого становится возможным установить свойства получаемых оценок параметров моделей.

Например, известный и широко применяемый метод максимума правдоподобия основан на идее максимизации распределения параметров модели при условии наблюдаемых данных: необходимо определить такие значения параметров, при которых будет достигаться максимум функции правдоподобия данных. Это означает, что вероятность того, что при найденных значениях параметров будут наблюдаться существующие данные, максимальна. Предположение о нормальности (или о наличии другого известного закона распределения) случайных компонент данных является основой этого метода, без этого предположения будет невозможно получить функцию правдоподобия в аналитическом виде [3].

Для вычисления оценок, максимизирующих правдоподобие, часто применяется метод наименьших квадратов (МНК), однако получаемые им оценки соответствуют оценкам максимального правдоподобия только для линейных

моделей. В нелинейном случае определение свойств этих оценок связано с существенными трудностями.

Предположение о нормальности стохастических компонент моделей, а значит и данных, с которыми ассоциирована модель, очевидно, не всегда является вполне корректным, так как не всегда удается выяснить или доказать требуемые факты о вероятностных характеристиках данных по имеющимся в наличии данным.

Энтропийно-оптимальные распределения, полученные на этапе обучения модели, могут быть использованы для прогнозирования несколькими способами, в частности, они могут быть сгенерированы для получения ансамбля выходов модели с последующим его анализом.

Работа посвящена развитию метода энтропийно-рандомизированного обучения и прогнозирования для нелинейным моделей с дискретными параметрами. Переход от непрерывных к дискретным моделям позволяет преодолеть трудности использования непрерывных моделей в условиях большого количества переменных, которые приводят к проблеме вычисления многомерных интегралов, произвести которое точно (аналитически) невозможно, а численное решение сопряжено с существенными вычислительными трудностями.

Предлагаемый в работе подход демонстрируется на примере задачи прогнозирования общего количества инфицированных в результате развития эпидемии COVID-19 в Германии. Проводится сравнение предлагаемого подхода с нелинейным методом наименьших квадратов [18, 21].

2. Нелинейная дискретная рандомизированная модель

Рассмотрим объект с n скалярными входами x_i, $i = \overline{1,n}$ и выходом $\hat{y}$, преобразование которых описывается в общем случае нелинейной функцией

$$\hat{y} = \Phi(\mathbf{x}, \mathbf{a}), \tag{2.1}$$

где $\mathbf{x} = (x_1, \ldots, x_n)$ — вектор входов, $\mathbf{a} = (a_1, \ldots, a_d)$ — вектор параметров модели.

Выход модели измеряется с некоторым шумом ξ, действующим аддитивно на выход, приводя к модели следующего вида:

$$v = \hat{y} + \xi = \Phi(\mathbf{x}, \mathbf{a}) + \xi. \tag{2.2}$$

Предположим, что значения каждого параметра сосредоточены на интервале $\mathcal{A}_k = [a_k^-, a_k^+]$, $k = \overline{1,d}$, и выход модели измеряется с шумом ξ_j, значения которого сосредоточены на интервале $\Xi_j = [\xi_j^-, \xi_j^+]$ для каждого заданного входа $\mathbf{x}_j$, $j = \overline{1,m}$.

Параметры a_k реализуются дискретной случайной величиной с M значениями на интервале $\mathcal{A}_k$, приводя к следующим распределениям:

$$a_{k\ell} \in A_k, \quad p_{k\ell} \in [0,1], \quad k = \overline{1,d}, \quad \ell = \overline{1,M}, \tag{2.3}$$

где $a_{k\ell}$ — значения случайной величины, а $p_{k\ell}$ — вероятности их реализации.

Шумы измерений выхода реализуются дискретной случайной величиной ξ_j с L значениями на интервале Ξ_j для каждого входа x_j. Измерения выхода производятся независимо друг от друга, таким образом при m измерениях приводя к следующим распределениям:

$$\xi_{jh} \in \Xi_j, \quad q_{jh} \in [0,1], \quad j = \overline{1,m}, \quad h = \overline{1,L}, \tag{2.4}$$

где ξ_{jh} — значения случайной величины, а q_{jh} — вероятности их реализации.

С учетом m измерений получаем итоговый вид модели (2.2):

$$\mathbf{v} = \hat{\mathbf{y}} + \boldsymbol{\xi} = \Phi(\mathbf{x}_j, \mathbf{a}) + \boldsymbol{\xi}, \quad j = \overline{1,m}, \tag{2.5}$$

где $\mathbf{v} = (v_1, \ldots, v_m)$ — вектор измеренного выхода модели, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ — вектор шумов, $\hat{\mathbf{y}} = (y_1, \ldots, y_m)$ — вектор выхода модели.

Распределения параметров и шумов измерений модели подлежат оцениванию с использованием реальных данных об измерениях выхода объекта, модель которого описывается (2.5).

3. Обучение модели с использованием реальных данных

Рандомизированное машинное обучение базируется на принципах энтропийного оценивания параметров модели и шумов измерений ее выхода. Энтропийно-оптимальные распределения отражают наиболее неопределенную ситуацию, что в условиях полного отсутствия информации о реальных характеристиках является единственным доступным в этих условиях решением [13, 22, 23].

Для вычисления оптимальных распределений требуется решить задачу условной максимизации энтропии распределений параметров и шумов измерений при условиях нормировки соответствующих распределений и выполнении условий на баланс среднего выхода модели с измерением выхода объекта. Задача формулируется следующим образом:

$$H(P,Q) = -\sum_{k=1}^{d}\sum_{\ell=1}^{M} p_{k\ell} \ln p_{k\ell} - \sum_{j=1}^{m}\sum_{h=1}^{L} q_{jh} \ln q_{jh} \to \max_{P,Q}, \tag{3.1}$$

где P и Q — распределения параметров и шумов (2.3) и (2.4), при условиях:

$$\sum_{\ell=1}^{M} p_{k\ell} = 1, \quad \sum_{h=1}^{L} q_{jh} = 1, \quad k = \overline{1,d}, \quad j = \overline{1,m}, \tag{3.2}$$

$$\mathbb{E}[\mathbf{v}] = \mathbb{E}[\Phi(\mathbf{x}_j, \mathbf{a}) + \boldsymbol{\xi}] = \mathbf{y}, \tag{3.3}$$

где $\mathbf{y} = (y_1, \ldots, y_m)$ — вектор измерений выхода объекта (реальные данные выхода).

Условие (3.3) определяет баланс среднего выхода модели с реальными данными выхода:

$$\mathbb{E}[\mathbf{v}_j] = \mathbb{E}[\Phi(\mathbf{x}_j, \mathbf{a}) + \xi_j] = \mathbb{E}[\Phi(\mathbf{x}_j, \mathbf{a})] + \mathbb{E}[\xi_j] =$$
$$= \sum_{\substack{\ell_k=1 \\ k=\overline{1,d}}}^{M} \Phi(\mathbf{x}_j, a_{1\ell_1}, \ldots, a_{d\ell_d}) p_{1\ell_1} \cdots p_{d\ell_d} + \sum_{h=1}^{L} \xi_{jh} q_{jh} =$$
$$= \bar{\Phi}(\mathbf{x}_j) + \sum_{h=1}^{L} \xi_{jh} q_{jh} = y_j. \tag{3.4}$$

Сумма в выражении для $\bar{\Phi}$ содержит M^d членов, суммирование осуществляется для всех комбинаций значений случайных величин $a_{k\ell}$. Решение задачи (3.1)–(3.3), подробно рассмотренное в Приложении, дает энтропийно-оптимальные распределения параметров и шумов измерений, что и является конечной целью обучения модели с использованием реальных данных.

4. Рандомизированное прогнозирование

В результате обучения модель оказывается снабжена энтропийно-оптимальными оценками распределений параметров и измерительных шумов, формируя таким образом *рандомизированную предсказательную модель* (РПМ). Такая модель определяет специальную методику прогнозирования — *рандомизированное прогнозирование*, элементы которого применялись для некоторых прикладных задач [24–26].

Рандомизированное прогнозирование базируется на генерации энтропийно-оптимальных распределений параметров (П.4) и измерительных шумов (П.5) с последующим построением ансамбля выхода модели для новых, не известных при обучении, входов модели.

Рассмотрим набор входов РПМ, для которых требуется построить прогноз, который может быть представлен в виде блочного вектора или матрицы, столбцами которой являются указанные входы:

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_s\} = \begin{bmatrix} x_{11} & \ldots & x_{1s} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{ns} \end{bmatrix}.$$

Пусть имеется выборка параметров из распределения P объема S. Тогда ансамбль выхода модели для одного входа $\mathbf{x}$ формируется согласно (2.1) и имеет вид

$$\hat{\mathcal{Y}} = \{\hat{y}_i = \Phi(\mathbf{x}, \mathbf{a}_i)\}, \quad i = \overline{1, S},$$

где $\mathbf{a}_i$ — реализация параметров с распределением P. Ансамбль содержит S траекторий.

Теперь для каждого входа $\mathbf{x}_j$, $j = \overline{1,s}$ и каждой реализации параметров $\mathbf{a}_i$, $i = \overline{1,S}$ рассмотрим выборку шумов из распределения q объема U и сформируем итоговый ансамбль выхода модели согласно (2.2):

$$\mathcal{V} = \{\mathbf{v}_j = \Phi(\mathbf{x}_j, \mathbf{a}_i) + \boldsymbol{\xi}_j\}, \quad i = \overline{1,S}, \quad j = \overline{1,s},$$

где $\boldsymbol{\xi}_j = (\xi_{j1}, \ldots, \xi_{jU})$ — вектор реализаций шумов для j-го входа, $\mathbf{v}_j$ — вектор измеренного выхода модели для j-го входа.

Таким образом, при прогнозировании для каждого входа и каждой реализации параметров модели, генерируется шум в количестве U реализаций. В итоге ансамбль $\mathcal{V}$ состоит из $W = SU$ траекторий, которые все вместе могут быть представлены блочным вектором или матрицей со строками, соответствующими прогнозируемому выходу модели для каждого входа

$$\mathcal{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_W] = \begin{bmatrix} v_{11} & \ldots & v_{1s} \\ \vdots & \ddots & \vdots \\ v_{W1} & \ldots & v_{Ws} \end{bmatrix}.$$

Для построения итоговой прогнозной траектории моделируемого процесса по ансамблю $\mathcal{V}$ могут быть вычислены средняя и медианные траектории, область стандартного отклонения, а также другие выборочные вероятностные характеристики.

Как видно из выражений ансамблей, для их формирования необходимо иметь распределение шума для каждого входа. Распределения, полученные при обучении, не могут быть напрямую использованы для произвольного количества прогнозных входов, так как получены из известных на этом этапе данных, а количество и характеристики данных при прогнозировании заранее не известны. Выходов из этой ситуации может быть несколько.

Первый состоит в применении в качестве прогнозного распределения шума q распределения, определяемого выражением (П.5) для среднего значения параметра λ (множителей Лагранжа).

Второй подход состоит в использовании в качестве прогнозного распределение шума для одного из входов, используемых при обучении, например последнего, если постановка задачи допускает порядок входов в наборе. Идея этого способа состоит в следующем: если исходить из того, что измерения в каждой точке последовательно расположенных данных производятся одним "устройством", то логично ожидать некоторой стабилизации характеристик этого измерительного устройства, которое достигается к последнему измерению из последовательности.

Третий подход основан на том предположении, что в результате энтропийного оценивания одновременно и параметров, и шумов можно рассматривать чистый выход модели без шума. Таким образом, можно говорить, что энтропийное оценивание осуществляет фильтрацию. В этом случае применение модели должно осуществляться в чистом виде, без шума.

Важной проблемой в применении энтропийно-рандомизированного подхода к прогнозированию является генерация оптимальных распределений параметров и шумов измерений, полученных при оценивании (обучении) модели.

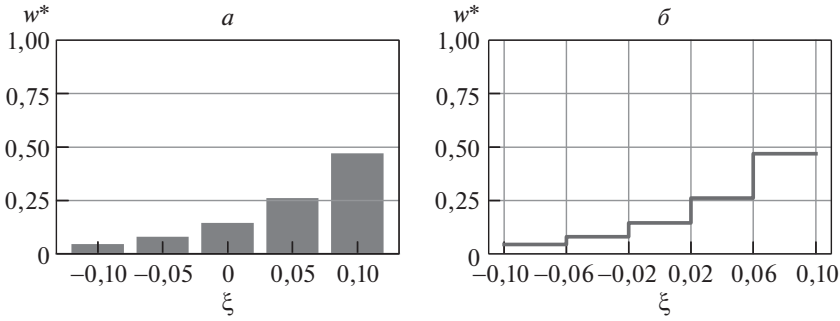

Рис. 1. Дискретное и кусочно-постоянное непрерывное распределение.

Для решения этой проблемы можно предложить два основных подхода, использующих генератор равномерно распределенных случайных чисел.

Первый подход состоит в генерации дискретного распределения. Для этого применяется стандартный подход, состоящий в случайном выборе значения случайной величины и затем соответствующей ему вероятности. В этом случае, очевидно, реализации, сделанные таким образом, будут представлять собой набор значений соответствующей случайной величины, многие из которых будут повторяться.

Второй подход основан на идее представления дискретного распределения в качестве кусочно-постоянной аппроксимации некоторого непрерывного распределения на соответствующем интервале. Для этого интервал значений соответствующей случайной величины разбивается на $L + 1$ подынтервалов (где L — количество значений дискретной случайной величины), левые и правые границы конечных подынтервалов соответствуют левым и правым границам интервала распределения. Внутри каждого подынтервала генерация происходит равномерно. В результате реализации такого подхода можно получить существенно больше различных значений соответствующих случайных величин. Пример построения непрерывных распределений для случайной величины $w^*(\xi)$, где $\xi \in [-0{,}1, 0{,}1]$ и $L = 5$, представлен на рис. 1.

5. Прогнозирование роста числа инфицированных COVID-19 в Германии

Предлагаемый в работе подход применяется для моделирования динамики развития эпидемии COVID-19 в Германии на основе данных Университета Джонса Хопкинса [27] начиная с сорокового дня эпидемии (8 марта 2020 г.), когда впервые общее количество инфицированных превысило 1000 человек.

Данные о развитии эпидемии (см. рис. 2) свидетельствуют о том, что сначала инфекция активно распространяется в популяции и наблюдается ее экспоненциальный рост. Далее наблюдается снижение числа зараженных, вероятно, вследствие ограничительных мер или увеличения количества иммунных членов популяции. При этом, как и в большинстве живых систем при рассмотрении их на относительно коротком промежутке времени, можно предположить, что объем популяции не меняется (например, можно пренебречь миграцией, воспроизводством и смертностью), а значит, существует ограниче-

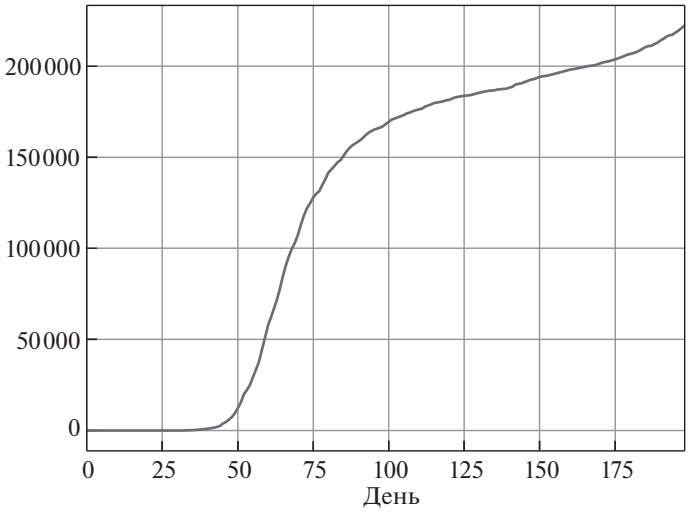

Рис. 2. Общее количество инфицированных в Германии по дням начиная с 8.03.2020.

ние на количество инфицированных. Динамика инфицированных членов популяции N в такой системе может быть описана следующим уравнением [28]:

$$\frac{dN}{dt} = \lambda N \left(1 - \frac{N}{K}\right), \tag{5.1}$$

где λ — скорость роста инфицированных, N — количество инфицированных, K — объем популяции. Решением этого уравнения является кривая Ферхюльста [29, 30]

$$N(t) = \frac{K}{1 + Be^{-\lambda t}}, \quad B = \frac{K - N_0}{N_0}, \tag{5.2}$$

где N_0 — количество инфицированных в популяции в начальный момент времени [30].

Модель вида (5.2) активно использовалась в начале 2020 г. для предсказания общего количества заболевших [31–37] и показала свою эффективность на начальном этапе развития эпидемии. В этой связи представляется обоснованным использовать аналогичную модель для применения энтропийно-рандомизированного подхода к прогнозированию общего количества инфицированных. В качестве такой модели будем использовать трехпараметрическую логистическую модель роста (Logistic Growth Model, LGM), которая определяет преобразование скалярного входа x в выход $\hat{y}$ с использованием логистической нелинейной функции

$$\hat{y} = \Phi(x, \mathbf{a}) = \frac{a_3}{1 + a_1 e^{-a_2 x}}, \tag{5.3}$$

где $\mathbf{a} = (a_1, a_2, a_3)$ — вектор параметров модели. Данная модель является обобщением модели (5.2) и рассматривается здесь как абстрактная модель с параметрами без использования дополнительных связей между ними.

В контексте рассматриваемой задачи входом является порядковый номер (или индекс) дня, а выходом — накопленное (общее) количество инфицированных. Вход и выход являются целыми, однако в вычислениях целые числа преобразуются в числа с плавающей точкой.

Рандомизированная модель, выход которой искажен аддитивным шумом, а параметры и шумы реализуются дискретными случайными величинами со значениями из соответствующих интервалов, имеет вид

$$v = \hat{y} + \xi = \Phi(x, \mathbf{a}) + \xi,$$

$$a_{k\ell} \in A_k, \quad p_{k\ell} \in [0,1], \quad k = \overline{1,d}, \quad \ell = \overline{1,M},$$
$$\xi_{jh} \in \Xi_j, \quad q_{jh} \in [0,1], \quad j = \overline{1,m}, \quad h = \overline{1,L},$$

где $a_{k\ell}$, ξ_{jh} — значения случайных величин, реализующих параметры и шумы, а $p_{k\ell}$, q_{jh} — вероятности их реализации, m — количество точек данных, $d = 3$.

Для обучения (оценивания) предсказательной модели использовались данные за несколько дней (интервал обучения $\mathcal{T}_{est}$) начиная с 8 марта 2020 г. (40-й день эпидемии в Германии), когда впервые было зафиксировано общее количество инфицированных свыше 1000 человек.

Прогнозирование производится на следующие за интервалом оценивания дни вплоть до 120-го дня эпидемии (интервал прогнозирования $\mathcal{T}_{pred}$).

Полученные рандомизированные прогнозы сравнивались с подгонкой кривой по модели (5.3) с помощью нелинейного метода наименьших квадратов, реализованного функцией `curve_fit` библиотеки `scipy` на платформе Python 3.7.

После оценивания модели проводилась ее реализация на интервале оценивания (тестирование) и на интервале прогнозирования с вычислением следующих метрик качества для истинных (реальных, true) значений y и предсказанных (модельных, predicted) значений $\hat{y}$:

1) коэффициент детерминации R^2, определяемый формулой

$$R^2(y, \hat{y}) = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2},$$

где $\bar{y}$ — среднее значение по реальным данным позволяет оценить качество приближения кривой (goodness of fit (GoF)) и предсказательную способность модели через долю объясненной дисперсии. Максимум этого индикатора равен 1, чем его значение ближе к единице, тем выше качество модели;

2) средне-квадратичная ошибка (Mean Squared Error (MSE)), определяемая формулой

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n-1}(y_i - \hat{y}_i)^2,$$

показывает ожидаемую (среднюю) квадратичную ошибку;

3) Norm Error (NE), определяемую формулой

$$\mathrm{NE}(y,\hat{y}) = \frac{\|y-\hat{y}\|}{\|y\|+\|\hat{y}\|} = \frac{\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum\limits_{i=1}^{n}y_i^2+\sum\limits_{i=1}^{n}\hat{y}_i^2};$$

4) Rooted Norm Error (RNE), определяемую формулой

$$\mathrm{RNE}(y,\hat{y}) = \frac{\sqrt{\|y-\hat{y}\|}}{\sqrt{\|y\|}+\sqrt{\|\hat{y}\|}} = \frac{\sqrt{\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2}}{\sqrt{\sum\limits_{i=1}^{n}y_i^2}+\sqrt{\sum\limits_{i=1}^{n}\hat{y}_i^2}}.$$

Согласно теории метода энтропийного оценивания определяемые в результате оценивания распределения параметров и шумов являются интервальными. Таким образом, для применения метода необходимо задать эти интервалы. В данной работе интервалы для параметров задавались на основе оптимальных значений, полученных при оценивании методом наименьших квадратов. Границы интервалов устанавливались в пределах 20% от этих значений.

В экспериментах использовались данные, масштабированные на отрезок $[0,1]$ на интервале оценивания.

Оценивание, тестирование и прогнозирование производилось для нескольких конфигураций:

- без шума;
- с шумом в пределах 10%;
- с шумом в пределах 30%.

При прогнозировании использовалось распределение шума, полученное для последней точки на интервале оценивания. Тестирование модели проводилось в конфигурации, соответствующей прогнозированию.

На рисунках изображены следующие результаты моделирования (траектории):

- метод наименьших квадратов (пунктирная линия с меткой ols);
- реальные данные (пунктирная линия с меткой real);
- рандомизированное прогнозирование со средними по распределению значениями параметров модели (линия с меткой mean_params);
- рандомизированное прогнозирование со средним по ансамблю (линия с меткой mean);
- рандомизированное прогнозирование с медианой по ансамблю (линия с меткой med).

Светло-серым цветом отмечены траектории, составляющие полученный ансамбль, темно-серым цветом — область стандартного отклонения по ансамблю. Все эксперименты производились для выборки из распределения параметров модели объемом 1000 и выборки из распределениям шумов объемом 100 для каждого значения параметра. Генерация распределений шума

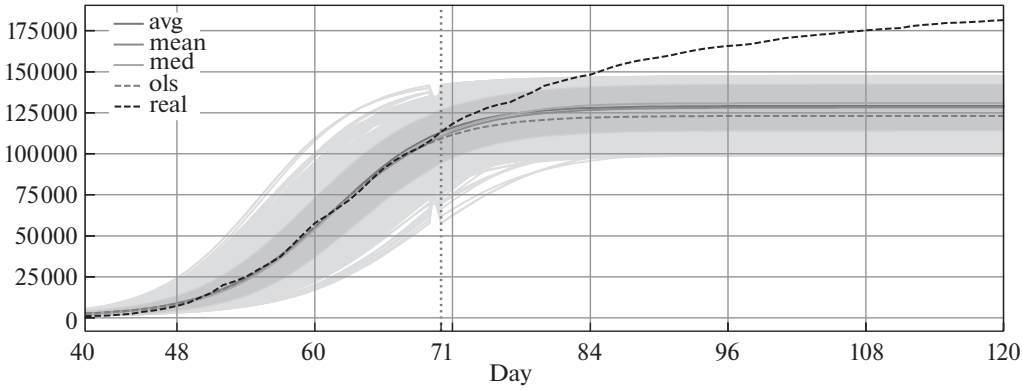

Рис. 3. Прогнозирование без шума (NN) для $\mathcal{T}_{est} = [40, 70]$ и $\mathcal{T}_{pred} = [71, 120]$.

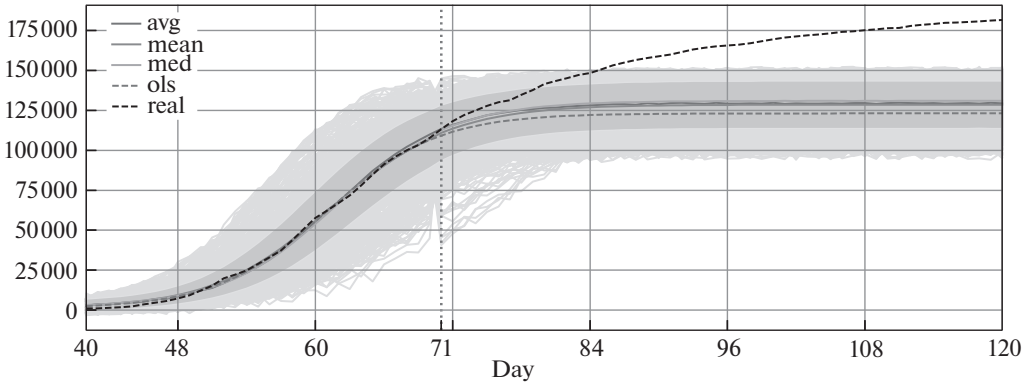

Рис. 4. Прогнозирование с шумом в пределах 10% (N1) для $\mathcal{T}_{est} = [40, 70]$ и $\mathcal{T}_{pred} = [71, 120]$.

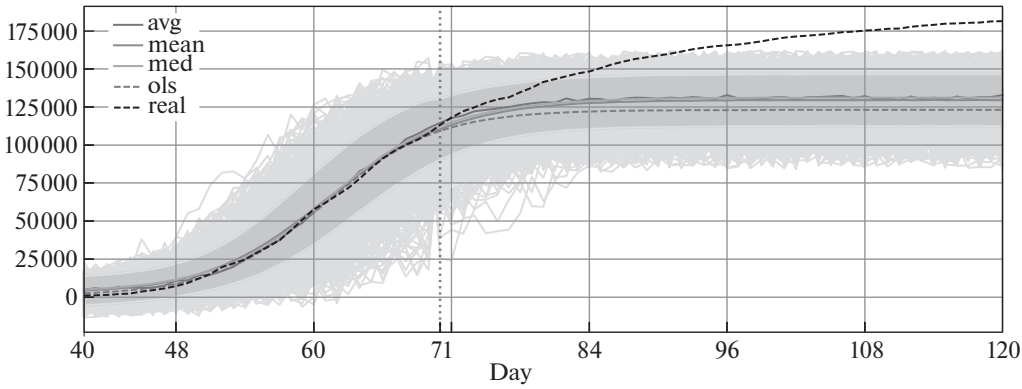

Рис. 5. Прогнозирование с шумом в пределах 30% (N3) для $\mathcal{T}_{est} = [40, 70]$ и $\mathcal{T}_{pred} = [71, 120]$.

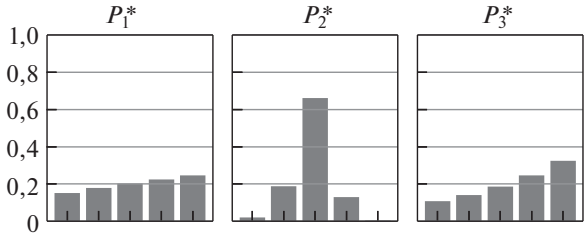

Рис. 6. Энтропийно-оптимальное распределение параметров P^*.

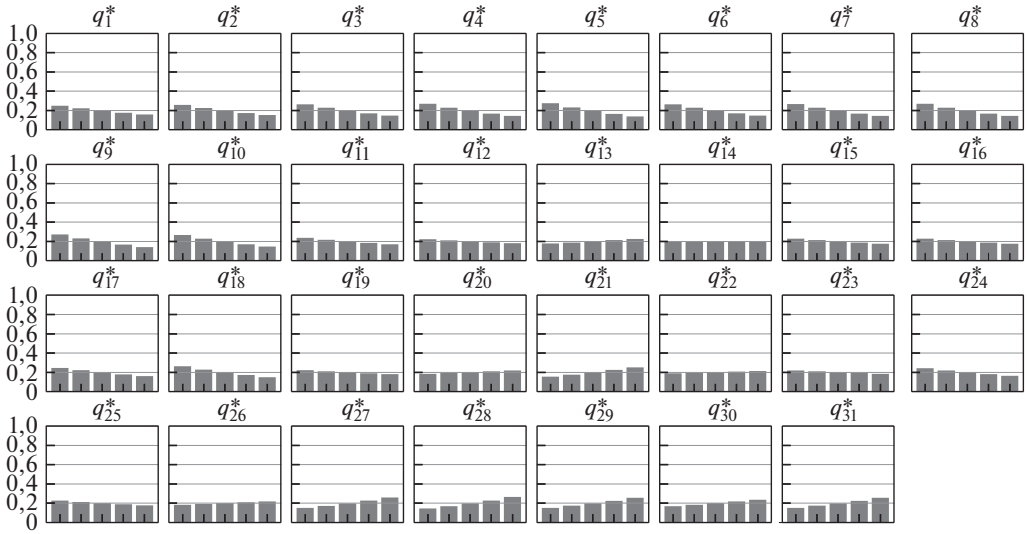

Рис. 7. Энтропийно-оптимальные распределения шумов Q^*.

проводилась для каждой точки соответствующего интервала (тестирования и прогнозирования). Таким образом, полученный ансамбль состоял из 10^5 траекторий. Вертикальная красная пунктирная линия нанесена в точке начала интервала прогнозирования. Эксперименты проводились на платформе Python 3.7 в среде Windows 10.

На рис. 3–5 приведены результаты реализации рандомизированной предсказательной модели на интервалах $\mathcal{T}_{est} = [40, 70]$ и $\mathcal{T}_{pred} = [71, 120]$ для трех вариантов прогнозирования: без шума (NN), с шумом 10% (N1) и с шумом 30% (N3).

На рис. 6–7 изображены энтропийно-оптимальные распределения параметров и шумов, полученные в результате обучения модели на интервале $\mathcal{T} = [40, 70]$ с шумом 30%.

На рис. 8–9 приведены результаты реализации рандомизированной предсказательной модели на интервалах $\mathcal{T}_{est} = [40, 76]$, $\mathcal{T}_{est} = [40, 79]$ и соответствующих интервалах прогнозирования.

Важно отметить, что на 77-й день произошел небольшой рост количества заболевших, что видно на графике. Однако при обучении модели этих данных

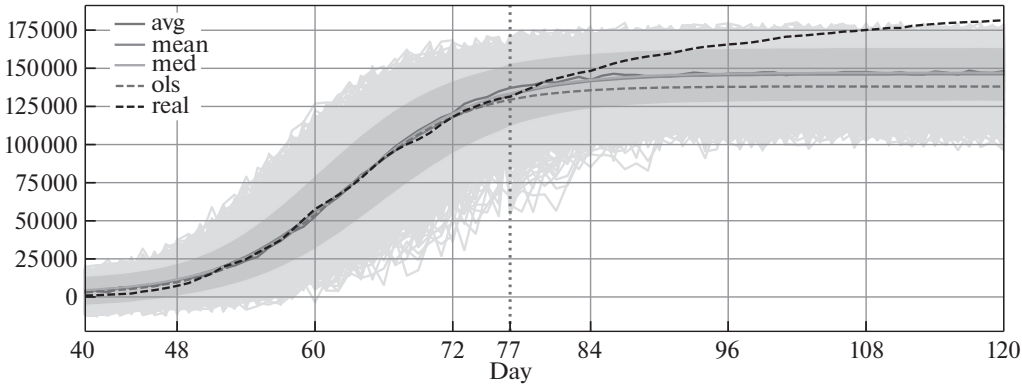

Рис. 8. Прогнозирование с шумом в пределах 30% (N3) для $\mathcal{T}_{est} = [40, 76]$ и $\mathcal{T}_{pred} = [77, 120]$.

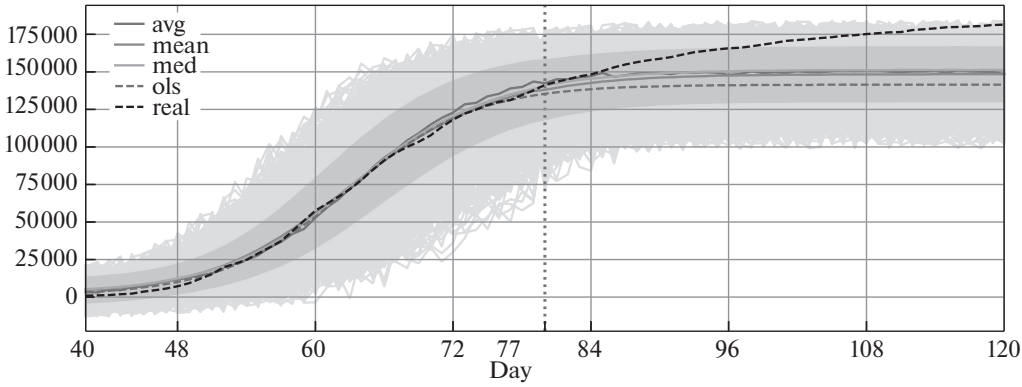

Рис. 9. Прогнозирование с шумом в пределах 30% (N3) для $\mathcal{T}_{est} = [40, 79]$ и $\mathcal{T}_{pred} = [80, 120]$.

еще не было. На 79-й день по-прежнему происходил рост количества заболевших, который начался ранее, но в этом случае уже была возможность учесть эти данные при обучении. На графиках видно, что во всех случаях рандомизированная модель предоставляет более реалистичный прогноз по сравнению с классической моделью.

Оценки параметров, полученные методом наименьших квадратов, и интервалы параметров рандомизированной модели указаны в табл. 1–2. В конфигурациях с шумом интервалы шумов задаются как $\Xi_j = [-0{,}1, 0{,}1]$ для N1 и $\Xi_j = [-0{,}3, 0{,}3]$ для N3 соответственно.

Значения показателей качества моделирования при тестировании на трех разных интервалах для варианта модели с шумом 30% указаны в табл. 3.

Анализируя полученные результаты, можно отметить, что стандартная методика прогнозирования, связанная с использованием подгонки кривой под данные методом наименьших квадратов, хоть и обладает определенной эф-

Таблица 1. Оценки параметров, полученные методом наименьших квадратов

$\mathcal{T}_{test}$	a_1	a_2	a_3
$[40, 70]$	297749	0,2065	1,1448
$[40, 76]$	1086539	0,1856	1,0624
$[40, 79]$	83517	0,1803	1,0278

Таблица 2. Конфигурации параметров рандомизированной модели

$\mathcal{T}_{pred}$	A_1	A_2	A_3
$[40, 70]$	$[238199, 357299]$	$[0{,}1652, 0{,}2478]$	$[0{,}9158, 1{,}3738]$
$[40, 76]$	$[86923, 130384]$	$[0{,}1485, 0{,}2227]$	$[0{,}8499, 1{,}2749]$
$[40, 79]$	$[66813, 100220]$	$[0{,}1443, 0{,}2164]$	$[0{,}8223, 1{,}2334]$

Таблица 3. Метрики качества на интервале оценивания

	R^2	MSE	NE	RNE
$\mathcal{T}_{test} = [40, 70]$				
ols	0,9984	0,0002	0,0004	0,0135
mean_params	0,9899	0,0011	0,0022	0,0335
mean	**0,9997**	0,0000	0,0001	0,0058
med	0,9980	0,0002	0,0004	0,0149
$\mathcal{T}_{test} = [40, 76]$				
ols	0,9982	0,0002	0,0004	0,0139
mean_params	0,9903	0,0012	0,0020	0,0314
mean	**0,9994**	0,0001	0,0001	0,0078
med	**0,9985**	0,0002	0,0003	0,0127
$\mathcal{T}_{test} = [40, 79]$				
ols	0,9982	0,0002	0,0004	0,0133
mean_params	0,9903	0,0012	0,0019	0,0306
mean	**0,9997**	0,0000	0,0001	0,0054
med	**0,9986**	0,0002	0,0003	0,0119

фективностью, не всегда способна качественно решить задачу построения корректного прогноза.

Из-за специфики рассматриваемой здесь эпидемии COVID-19 во всем мире наблюдается существенное искажение данных, связанных с ней. В этой связи представляется актуальной задача прогнозирования количества инфицированных с некоторым превышением. Из полученных результатов видно, что прогнозирование с помощью логистической модели, оцененной с использованием тех данных, которые были на момент прогноза, существенно недооценивает реальные данные. В то же время прогнозы, получаемые с использованием предлагаемого в работе подхода, показывают превышение прогнозных значений по сравнению с МНК. Необходимо также отметить, что использование шума в модели, который оценивается поточечно и потом используется при прогнозировании, позволяет построить более реалистичный прогноз.

6. Заключение

В работе развит метод рандомизированного машинного обучения и прогнозирования, основанный на использовании дискретных случайных величин, что приводит к задачам, более адаптированным к численному решению с использованием современной вычислительной техники. Демонстрация предложенного метода проведена на задаче прогнозирования общего количества инфицированных COVID-19 в Германии. Полученные результаты свидетельствуют о работоспособности и эффективности метода и его численной реализации, что определяется меньшей ошибкой при прогнозировании по сравнению со стандартной методикой, основанной на методе наименьших квадратов. Необходимо также отметить, что построенная рандомизированная модель показала хороший результат на интервале обучения, однако на интервале прогноза погрешность по сравнению с реальными данными составила существенную величину. Это связано, по всей видимости, с тем, что логистическая модель эффективна для прогноза не на всех этапах развития эпидемии, в частности, для обеспечения приемлемого уровня качества прогноза необходимо наличие признаков замедления эпидемии на интервале обучения, а также наличие реального затухания эпидемии на интервале прогноза. При проведении экспериментов для обучения использовались данные на этапе начала и активного развития эпидемии, что объясняет большую ошибку при прогнозировании.

ПРИЛОЖЕНИЕ

Рассмотрим решение задачи (3.1)–(3.3), которое осуществляется методом множителей Лагранжа. Функция Лагранжа будет иметь вид

$$L(P,Q,\alpha,\beta,\lambda) = -H(P,Q)+ \\ + \sum_{k=1}^{d} \alpha_k \left(\sum_{\ell=1}^{M} p_{k\ell} - 1 \right) + \sum_{j=1}^{m} \beta_j \left(\sum_{h=1}^{L} q_{jh} - 1 \right) + \\ + \sum_{j=1}^{m} \lambda_j \left(\bar{\Phi}(\mathbf{x}_j) + \sum_{h=1}^{L} \xi_{jh} q_{jh} - y_j \right).$$

Для поиска экстремума функции Лагранжа вычислим производные по прямым переменным P и Q:

$$\frac{\partial L}{\partial P} = \frac{\partial L}{\partial p_{k\ell}} = \ln p_{k\ell} + 1 + \alpha_k + \sum_{j=1}^{m} \lambda_j \frac{\partial \bar{\Phi}_j}{\partial p_{k\ell}},$$

$$\frac{\partial L}{\partial Q} = \frac{\partial L}{\partial q_{jh}} = \ln q_{jh} + 1 + \beta_j + \sum_{j=1}^{m} \lambda_j \xi_{jh},$$

$$k = \overline{1,d}, \quad j = \overline{1,m}, \quad \ell = \overline{1,M}, \quad h = \overline{1,L},$$

где $\bar{\Phi}_j = \bar{\Phi}(\mathbf{x}_j)$ и производная среднего значения модели по $p_{k\ell}$ определяется выражением

$$\frac{\partial \bar{\Phi}_j}{\partial p_{k\ell}} = \sum_{\substack{\ell_s=1 \\ s=\overline{1,d}}}^{M} \Phi(x_j, a_{1\ell_1}, \ldots, a_{d\ell_d}) \prod_{r \neq k} p_{r\ell_r}. \tag{П.1}$$

Сумма в выражении для производной $\frac{\partial \bar{\Phi}}{\partial p_{k\ell}}$ содержит $M(d-1)$ членов.

Приравнивая к нулю производные функции Лагранжа по прямым переменным, получим выражения оптимальных распределений вероятностей параметров и шумов от множителей Лагранжа:

$$p^*_{k\ell}(\alpha, \lambda) = \exp\left(-1 - \alpha_k - \sum_{j=1}^{m} \lambda_j \frac{\partial \bar{\Phi}_j}{\partial p_{k\ell}}\right),$$
$$q^*_{jh}(\beta, \lambda) = \exp\left(-1 - \beta_j - \lambda_j \xi_{jh}\right).$$

Преобразуем эти выражения следующим образом:

$$p^*_{k\ell}(\alpha, \lambda) = \exp\left(-\left(1 + \alpha_k\right)\right) \exp\left(-\sum_{j=1}^{m} \lambda_j \frac{\partial \bar{\Phi}_j}{\partial p_{k\ell}}\right), \tag{П.2}$$

$$q^*_{jh}(\beta, \lambda) = \exp\left(-\left(1 + \beta_j\right)\right) \exp\left(-\lambda_j \xi_{jh}\right), \tag{П.3}$$

и, подставляя их в условия нормировки (3.2), получим выражения:

$$\exp\left(1 + \alpha_k\right) = \exp\left(-\sum_{\ell=1}^{M} \lambda_j \frac{\partial \bar{\Phi}_j}{\partial p_{k\ell}}\right),$$
$$\exp\left(1 + \beta_j\right) = \exp\left(-\lambda_j \xi_{jh}\right).$$

Подставим эти выражения обратно в (П.2)–(П.3), исключив таким образом множители α и β, получим финальные выражения энтропийно-оптимальных распределений параметров и шумов, зависящих от множителей λ:

$$p^*_{k\ell}(\lambda) = \frac{\exp\left(-\sum\limits_{j=1}^{m} \lambda_j \frac{\partial \bar{\Phi}_j}{\partial p_{k\ell}}\right)}{\sum\limits_{\ell=1}^{M} \exp\left(-\sum\limits_{j=1}^{m} \lambda_j \frac{\partial \bar{\Phi}_j}{\partial p_{k\ell}}\right)}, \quad k = \overline{1,d},\ \ell = \overline{1,M}, \tag{П.4}$$

$$q^*_{jh}(\lambda) = \frac{\exp\left(-\lambda_j \xi_{jh}\right)}{\sum\limits_{h=1}^{L} \exp\left(-\lambda_j \xi_{jh}\right)}, \quad j = \overline{1,m},\ h = \overline{1,L}. \tag{П.5}$$

Множители λ определяются решением системы уравнений, получающихся подстановкой выражений (П.4)–(П.5) в балансовые соотношения (3.3):

$$(\text{П.6}) \qquad \sum_{\substack{\ell_k=1\\k=\overline{1,d}}}^{M} \Phi(\mathbf{x}_j, a_{1\ell_1}, \ldots, a_{d\ell_d}) \prod_{\substack{\ell_s=1\\s=\overline{1,d}}}^{M} p^*_{s\ell_s}(\lambda) + \sum_{h=1}^{L} \xi_{jh} q^*_{jh}(\lambda) = y_j, \quad j = \overline{1,m}.$$

Таким образом, решив систему (П.6), получим энтропийно-оптимальные распределения параметров и шумов измерений, что и является конечной целью обучения модели с использованием реальных данных.

Необходимо отметить, что для решения этой системы на практике необходимо привлекать какой-нибудь численный метод, так как ее решение аналитически сопряжено с существенными трудностями.

Вычисление левой части системы потребует вычисления среднего значения случайной функции $\bar{\Phi}$, а также ее производной по распределению P, определяемых выражениями (3.4) и (П.1). Суммирования в этих выражениях должны производиться по всем комбинациям индексов, таким образом количество операций суммирования растет как степень от количества параметров d.

СПИСОК ЛИТЕРАТУРЫ

1. *Bishop C.M.* Pattern Recognition and Machine Learning. Springer, Series: Information Theory and Statistics, 2006.
2. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer, 2001.
3. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: Юнити, 1998.
4. *Мерков А.Б.* Распознавание образов. Введение в методы статистического обучения. М.: URSS, 2010.
5. *Аркадьев А.Г., Браверман Э.М.* Обучение машины распознаванию образов. М.: Наука, 1964.
6. *Цыпкин Я.З.* Основы теории обучающихся систем. М.: Наука, 1970.
7. *Вапник В.Н., Червоненкис А.Я.* Восстановление зависимостей по эмпирическим данным, М.: Наука, 1979.
8. *Вапник В.Н. Червоненкис А.Я.* Теория распознавания образов. М.: Наука, 1974.
9. *Cristianini N., Shawe-Taylor J.* An introduction to support vector machines, 2000.
10. *Breiman J.H., Friedman R., Olshen A., Stone C.J.* Classification and regression trees. 1984.
11. *Rosenblatt F.* The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory, 1957.
12. *Rumelhart D.E., Williams R.J., Hinton G.* Learning representations by back-propagating errors // Nature. 1986. V. 323. No. 6088. P. 533–538.
13. *Попков Ю.С., Попков А.Ю., Дубнов Ю.А.* Рандомизированное машинное обучение при ограниченных наборах данных: от эмпирической вероятности к энтропийной рандомизации. М.: ЛЕНАНД, 2019.

14. *Больцман Л.* О связи между вторым началом механической теории теплоты и теорией вероятностей в теоремах о тепловом равновесии / Больцман Л.Э. Избранные труды. под ред. Шлак Л.С. Классики науки. М.: Наука, 1984.

15. *Jaynes E.T.* Information theory and statistical mechanics // Physical review. 1957. V. 106. No. 4. P. 620–630.

16. *Jaynes E.T.* Probability theory: the logic of science. Cambridge university press, 2003.

17. *Shannon C.E.* Communication theory of secrecy systems // Bell Labs Technical J. 1949. V. 28. No. 4. P. 656–715.

18. *Diebold F.* Elements of Forecasting. Thomson, South-Western, Ohio, US, 4th edition, 2007.

19. *Gneiting T., Katzfuss M.* Probabilistic forecasting // Annual Review of Statistics and Its Application, 2014. No. 1. P. 125–151.

20. *Hong T., Fan S.* Probabilistic electric load forecasting: A tutorial review // Int. J. Forecasting. 2016. V. 32. No. 3. P. 914–938.

21. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989.

22. *Golan A., Judge G., Miller D.* Maximum Entropy Econometrics: Robust Estimation with Limited Data. N.Y.: John Wiley & Sons, 1996.

23. *Golan A. et al.* Information and Entropy Econometrics — A Review and Synthesis // Foundations and Trends in Economet. 2008. V. 2. No. 1–2. P. 1–145.

24. *Popkov Y.S., Volkovich Z., Dubnov Y.A., Avros R., Ravve E.* Entropy 2-soft classification of objects // Entropy. 2017. V. 19. No. 4. P. 178.

25. *Попков Ю.С., Попков А.Ю., Дубнов Ү.А.* Элементы рандомизированного прогнозирования и его применение для предсказания суточной электрической нагрузки энергетической системы // АиТ. 2020. № 7. С. 148–172.

    Popkov Y.S., Popkov A.Y., Dubnov Y.A. Elements of Randomized Forecasting and Its Application to Daily Electrical Load Prediction in a Regional Power System // Autom. Remote Control. 2020. V. 81. P. 1286–1306.

26. *Popkov Y.S., Popkov A.Y., Dubnov Y.A., Solomatine D.* Entropy-randomized forecasting of stochastic dynamic regression models // Mathematics. 2020. V. 8. No. 7. P. 1119.

27. *Dong E., Du H., Gardner L.* An interactive web-based dashboard to track covid-19 in real time // The Lancet Infectious Diseases. 2020. V. 20. No. 5. P. 533–534.

28. *van den Driessche P.* Mathematical Epidemiology. In F. Brauer, P. van den Driessche & J. Wu (Eds.), Lecture Notes in Mathematics, 2008, `https://doi.org/10.1007/978-3-540-78911-6`.

29. *Verhulst P.-F.* Notice sur la loi que la population suit dans son accroissement // Corresp. Math. Phys., 1893. No. 10. P. 113–126.

30. *Singer H.M.* The COVID-19 pandemic: growth patterns, power law scaling, and saturation // Physical Biology. 2020. Vol. 17. No. 5. P. 055001. `https://doi.org/10.1088/1478-3975/ab9bf5`

31. *Kumar J., Hembram K.P.S.S.* Epidemiological study of novel coronavirus (COVID-19) // ArXiv. 2020. `http://arxiv.org/abs/2003.11376`

32. *Yang W., Zhang D., Peng L., Zhuge C., Hong L.* Rational evaluation of various epidemic models based on the COVID-19 data of China // ArXiv. 2020.

    `http://arxiv.org/abs/2003.05666`

33. *Tatrai D., Varallyay Z.* COVID-19 epidemic outcome predictions based on logistic fitting and estimation of its reliability // ArXiv. 2020. http://arxiv.org/abs/2003.14160

34. *Morais A.F.* Logistic approximations used to describe new outbreaks in the 2020 COVID-19 pandemic. ArXiv. 2020. http://arxiv.org/abs/2003.11149

35. *Shen C.Y.* Logistic growth modelling of COVID-19 proliferation in China and its international implications // Int. J. Infectious Diseases. 2020. Vol. 96. P. 582–589. https://doi.org/10.1016/j.ijid.2020.04.085

36. *Wang P., Zheng X., Li J., Zhu B.* Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics // Chaos Solitons & Fractals. 2020. Vol. 139. P. 110058. https://doi.org/10.1016/j.chaos.2020.110058

37. *Chen D.-G., Chen X., Chen J. K.* Reconstructing and forecasting the COVID-19 epidemic in the United States using a 5-parameter logistic growth model // Global Health Research and Policy. 2020. Vol. 5. No. 1. P. 25. https://doi.org/10.1186/s41256-020-00152-5

Статья представлена к публикации членом редколлегии А.И. Михальским.

Поступила в редакцию 15.10.2020

После доработки 12.01.2020

Принята к публикации 15.01.2021