Оптимизация, системный анализ и исследование операций

© 2022 г. А.В. ГРАБОВОЙ (grabovoy.av@phystech.edu)
(Московский физико-технический институт),
В.В. СТРИЖОВ, д-р физ.-мат. наук (strijov@phystech.edu)
(Московский физико-технический институт;
Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, Москва)

ВЕРОЯТНОСТНАЯ ИНТЕРПРЕТАЦИЯ ЗАДАЧИ ДИСТИЛЛЯЦИИ[1]

Статья посвящена методам понижения сложности аппроксимирующих моделей. Предлагается вероятностное обоснование методов дистилляции и привилегированного обучения. Приведены общие выводы для произвольной параметрической функции с наперед заданной структурой. Показано теоретическое обоснование для частных случаев: линейной и логистической регрессии. Проводится анализ рассмотренных моделей в вычислительном эксперименте на синтетических выборках и реальных данных. В качестве реальных данных рассматриваются выборки FashionMNIST и Twitter Sentiment Analysis.

Ключевые слова: выбор модели, байесовский вывод, дистилляция модели, привилегированное обучение.

DOI: 10.31857/S0005231022010093

1. Введение

Увеличение точности аппроксимации в задачах машинного обучения увеличивает сложность моделей и снижает их интерпретируемость. Примерами являются трансформеры [1], BERT [2], ResNet [3] и ансамбли этих моделей.

При построении модели оптимизируются два критерия: сложность и точность аппроксимации модели. Сложность определяет время, которое модель требует для принятия решения, и интерпретируемость модели. Модель меньшей сложности является более предпочтительной [4]. С учетом снижения сложности требуется сохранить приемлемой точность аппроксимации. В данной статье рассматривается метод *дистилляции* модели, предназначенный

[1] Настоящая статья содержит результаты проекта Математические методы интеллектуального анализа больших данных, выполняемого в рамках реализации Программы Центра компетенций Национальной технологической инициативы "Центр хранения и анализа больших данных", поддерживаемого Министерством науки и высшего образования Российской Федерации по Договору МГУ им. М.В. Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 11.12.2018 № 13/1251/2018. Работа выполнена при поддержке Российского фонда фундаментальных исследований (проекты 19-07-01155, 19-07-00875, 19-07-00885).

для снижения сложности при сохранении точности моделей. Этот метод строит новые модели на основе ранее обученных моделей.

Определение 1. Дистилляция модели — снижение ее сложности путем выбора из множества более простых моделей с использованием ответов более сложной модели.

Основным подходом дистилляции модели учителя в модель ученика является метод, основанный на использовании ответов модели учителя при оптимизации модели ученика [5–10]. В первых публикациях по генерации псевдометок предлагается пополнить множество объектов редких классов с помощью предобученной модели [6]. Это искусственно увеличивает объем обучающей выборки. В [5] предложен метод, в рамках которого моделью учителя генерируются новые метки объектов. Эти метки соответствуют вероятностям классов с некоторым параметром температуры, который позволяет увеличивать или уменьшать дисперсию в полученных ответах учителя. В [5] проведен ряд экспериментов по дистилляции моделей для разных задач машинного обучения: эксперимент на выборке MNIST [11], в котором нейросеть избыточной сложности была дистиллирована в нейросеть меньшей сложности, и эксперимент по распознаванию речи, в котором ансамбль моделей был дистиллирован в одну модель. Также в [5] был проведен эксперимент по обучению экспертных моделей на основе одной большой модели. В [8] предложено добавить к новым вероятностным меткам, введенным Дж. Хинтоном, метки классов, которые соответствуют предсказанному классу модели учителя. Различные подходы к дистилляции рассматривают значение на промежуточных слоях модели учителя [12–14]. В [12, 14] обучение происходит при помощи введения дополнительных матриц, которые выравнивают размеры промежуточных слоев модели учителя и ученика. В [13] предложен метод передачи селективности нейронов, основанный на минимизации максимального среднего отклонения между выходами всех слоев модели учителя и ученика.

Определение 2. Привилегированная информация — множество признаков, которые доступны только в момент выбора модели, но не в момент тестирования.

В [15] В.Н. Вапником введено понятие *привилегированной информации.* В [7] метод дистилляции [5] используется вместе с привилегированным обучением [15]. В предложенном методе на первом этапе обучается модель *учителя* в пространстве привилегированной информации, после чего обучается модель *ученика* в исходном признаковом пространстве, используя *дистилляцию* [5]. Для обучения строится функция ошибки специального вида. Эта функция состоит из нескольких слагаемых, включая ошибки учителя, ученика и регуляризирующие элементы. Первые варианты подобной функции ошибки были предложены А.Г. Ивахненко [16].

Определение 3. Учитель — фиксируемая модель, ответы которой используются при выборе модели ученика.

Определение 4. Ученик — модель, которая выбирается согласно заданному критерию.

Данная статья посвящена вероятностной интерпретации методов дистилляции, предложенных Дж. Хинтоном [5] и В.Н. Вапником [15]. В рамках вероятностного подхода предлагаются анализ и обобщение функции ошибки [5, 7]. Рассматриваются задачи классификации и регрессии в [16]. В ходе вычислительного эксперимента обучается модель ученика с использованием модели учителя и без использования модели учителя. Рассмотрены выборки задач классификации изображений FashionMNIST [17] и классификации текстов Twitter Sentiment Analysis [18]. Выборка FashionMNIST включена в эксперимент вместо выборки MNIST, так как последняя имеет приемлемое качество аппроксимации даже для линейного классификатора. Вычислительный эксперимент рассматривает различные модели: линейную модель, полносвязную нейронную сеть, сверточную нейронную сеть [19], модель Bi-LSTM [20] и модель BERT [2].

2. Постановка задачи обучения с учителем

Заданы множество объектов $\mathbf{\Omega}$ и множество целевых переменных $\mathbb{Y}$. Множество $\mathbb{Y} = \{1, \ldots, K\}$ для задачи классификации, где K — число классов, множество $\mathbb{Y} = \mathbb{R}$ для задачи регрессии. Для каждого объекта из $\omega_i \in \mathbf{\Omega}$ задана целевая переменная $\mathbf{y}_i = \mathbf{y}(\omega_i)$. Множество целевых переменных для всех объектов обозначим $\mathbf{Y}$. Для множества $\mathbf{\Omega}$ задано отображение в признаковое пространство $\mathbb{R}^n$:

$$\varphi : \mathbf{\Omega} \to \mathbb{R}^n, \quad |\mathbf{\Omega}| = m,$$

где n — размерность признакового пространства, а m — число объектов в множестве $\mathbf{\Omega}$. Отображение φ отображает объект $\omega_i \in \mathbf{\Omega}$ в соответствующий ему вектор признаков $\mathbf{x}_i = \varphi(\omega_i)$. Пусть для объектов $\mathbf{\Omega}^* \subset \mathbf{\Omega}$ задана привилегированная информация

$$\varphi^* : \mathbf{\Omega}^* \to \mathbb{R}^{n^*}, \quad |\mathbf{\Omega}^*| = m^*,$$

где $m^* \leqslant m$ — число объектов с привилегированной информацией, n^* — число признаков в пространстве привилегированной информации. Отображение φ^* отображает объект $\omega_i \in \mathbf{\Omega^*}$ в соответствующий ему вектор признаков $\mathbf{x}_i^* = \varphi^*(\omega_i)$.

Множество индексов объектов, для которых известна привилегированная информация, обозначим

$$\mathcal{I} = \Big\{ 1 \leqslant i \leqslant m | \text{ для } i\text{-го объекта задана привилегированная информация} \Big\},$$

а множество индексов объектов, для которых неизвестна привилегированная информация, обозначим $\bar{\mathcal{I}} = \{1, \ldots, m\} \setminus \mathcal{I}$.

Пусть на множестве привилегированных признаков задана функция учителя

$$\mathbf{f}\big(\mathbf{x}^*\big) : \mathbb{R}^{n^*} \to \mathbb{Y}^*,$$

где для задачи регрессии $\mathbb{Y}^* = \mathbb{R}^1$, а для задачи классификации $\mathbb{Y}^*$ является единичным симплексом $\mathcal{S}_K$ в пространстве размерности K. Модель учителя $\mathbf{f}$ ставит объекты $\mathbf{X}^*$ в соответствие объектам $\mathbf{S}$, т.е. $\mathbf{f}\big(\mathbf{x}_i^*\big) = \mathbf{s}_i$.

Требуется выбрать модель ученика $\mathbf{g}\big(\mathbf{x}\big)$ из множества

$$\mathfrak{G} = \Big\{\mathbf{g} | \mathbf{g} : \mathbb{R}^n \to \mathbb{Y}^*\Big\}. \tag{1}$$

Например, для задачи классификации множество $\mathfrak{G}$ — обобщенно-линейные модели

$$\mathfrak{G}_{\text{lin,cl}} = \Big\{\mathbf{g} | \mathbf{g}\big(\mathbf{W}, \mathbf{x}\big) = \mathbf{softmax}\big(\mathbf{W}\mathbf{x}\big), \quad \mathbf{W} \in \mathbb{R}^{n\times K}\Big\}.$$

3. Постановка задачи Хинтона и Вапника

Рассмотрим описание метода, предложенного в публикациях [5, 7], в которых предполагается, что для всех данных доступна привилегированная информация $\mathcal{I} = \{1, 2, \ldots, m\}$. В [5] решается задача классификации:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \qquad \mathbf{x}_i \in \mathbb{R}^n, \quad y_i \in \mathbb{Y} = \{1, \ldots, K\},$$

где y_i — класс объекта. Обозначим через $\mathbf{y}_i$ вектор вероятности класса объекта $\mathbf{x}_i$.

В постановке Хинтона рассматривается параметрическое семейство функций:

$$\mathfrak{G}_{\text{cl}} = \Big\{\mathbf{g} | \mathbf{g} = \text{softmax}\big(\mathbf{z}\big(\mathbf{x}\big)/T\big), \quad \mathbf{z} : \mathbb{R}^n \to \mathbb{R}^K\Big\}, \tag{2}$$

где $\mathbf{z}$ — дифференцируемая параметрическая функция заданной структуры, T — параметр температуры. В качестве модели учителя $\mathbf{f}$ рассматривается функция из множества $\mathfrak{F}_{\text{cl}}$:

$$\mathfrak{F}_{\text{cl}} = \Big\{\mathbf{f} | \mathbf{f} = \text{softmax}\big(\mathbf{v}\big(\mathbf{x}\big)/T\big), \quad \mathbf{v} : \mathbb{R}^n \to \mathbb{R}^K\Big\}, \tag{3}$$

где $\mathbf{v}$ — дифференцируемая параметрическая функция, модель заданной структуры, T — параметр температуры. Свойства параметра температуры T:

1) при $T \to 0$ получаем вектор, в котором один из классов имеет высокую вероятность;
2) при $T \to \infty$ получаем равновероятные классы.

Функция потерь $\mathcal{L}$, в которой учитывается перенос информации от модели учителя $\mathbf{f}$ к модели ученика $\mathbf{g}$, имеет вид:

$$\mathcal{L}_{st}(\mathbf{g}) = \underbrace{-\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log \mathbf{g}(\mathbf{x}_i)\big|_{T=1}}_{\text{исходная функция потерь}} - \tag{4}$$
$$\underbrace{-\sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{f}(\mathbf{x}_i)\big|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)\big|_{T=T_0}}_{\text{слагаемое дистилляции}},$$

где $\cdot\big|_{T=t}$ обозначает, что параметр температуры T в предыдущей функции равняется t.

Получаем оптимизационную задачу

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g}\in\mathfrak{G}_{\text{cl}}} \mathcal{L}_{st}(\mathbf{g}). \tag{5}$$

Публикация [7] обобщает метод, предложенный в [5]. Решение задачи оптимизации (5) зависит только от вектора ответов модели учителя $\mathbf{f}$. Следовательно, признаковые пространства учителя и ученика могут различаться. В этом случае получаем постановку задачи:

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{x}_i^*, y_i)\}_{i=1}^m, \qquad \mathbf{x}_i \in \mathbb{R}^n, \quad \mathbf{x}_i^* \in \mathbb{R}^{n^*}, \quad y_i \in \{1, \ldots, K\},$$

где $\mathbf{x}_i$ — информация, доступная на этапах обучения и контроля, а $\mathbf{x}_i^*$ — информация, доступная только на этапе обучения. Модель учителя принадлежит множеству моделей $\mathfrak{F}_{cl}^*$:

$$\mathfrak{F}_{\text{cl}}^* = \left\{\mathbf{f} | \mathbf{f} = \text{softmax}\big(\mathbf{v}^*(\mathbf{x}^*)/T\big), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \to \mathbb{R}^K\right\}, \tag{6}$$

где $\mathbf{v}^*$ — дифференцируемая параметрическая функция заданной структуры, T — параметр температуры. Множество моделей $\mathfrak{F}_{\text{cl}}^*$ отличается от множества моделей $\mathfrak{F}_{\text{cl}}$ из выражения (3). В множестве $\mathfrak{F}_{\text{cl}}$ модели используют пространство исходных признаков, а в множестве $\mathfrak{F}_{\text{cl}}^*$ модели используют пространство привилегированных признаков. Функция потерь (4) в случае модели учителя $\mathbf{f} \in \mathfrak{F}_{\text{cl}}^*$ переписывается в виде:

$$\mathcal{L}_{\text{st}}(\mathbf{g}) = -\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log \mathbf{g}(\mathbf{x}_i)\big|_{T=1} - \sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{f}(\mathbf{x}_i^*)\big|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)\big|_{T=T_0}. \tag{7}$$

Требуется построить модель, которая использует привилегированную информацию $\mathbf{x}_i^*$ при обучении. Для этого рассмотрим двухэтапную модель обучения, предложенную в [7]:

1) выбираем оптимальную модель учителя $\mathbf{f} \in \mathfrak{F}_{\text{cl}}^*$;

2) выбираем оптимальную модель ученика $\mathbf{g} \in \mathfrak{G}_{\text{cl}}$, используя дистилляцию [5].

Модель ученика минимизирует (7). Модель учителя минимизирует кросс-энтропийную функцию ошибки

$$\mathcal{L}_{\text{th}}(\mathbf{f}) = -\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log \mathbf{f}(\mathbf{x}_i^*).$$

4. Постановка задачи: вероятностный подход

4.1. Метод максимального правдоподобия

Задано распределение целевой переменной $p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g})$. Для поиска $\hat{\mathbf{g}}$ воспользуемся методом максимального правдоподобия. В качестве $\hat{\mathbf{g}}$ выбирается функция, которая максимизирует правдоподобие модели:

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathfrak{G}} \prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}), \tag{8}$$

где множество $\mathfrak{G}$ задается в (1).

4.2. Подход дистилляции модели учителя в модель ученика

Рассмотрим вероятностную постановку, в которой выполнены ограничения:

1) задано распределение целевой переменной $p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g})$;
2) задано совместное распределение целевой переменной и ответов модели учителя $p(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g})$;
3) для всех $\omega \in \mathbf{\Omega}^*$ элементы $\mathbf{y}(\omega)$ и $\mathbf{s}(\omega)$ являются зависимыми величинами, так как ответы учителя должны коррелировать с истинными ответами;
4) если $|\mathbf{\Omega}^*| = 0$, то решение должно соответствовать решению (8).

Рассмотрим совместное правдоподобие истинных меток и меток учителя:

$$p(\mathbf{Y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}) = \prod_{i\notin\mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) \prod_{i\in\mathcal{I}} p(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}). \tag{9}$$

Перепишем $p(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g})$ по формуле условной вероятности:

$$p(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}) = p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) p(\mathbf{s}_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}). \tag{10}$$

Подставляя выражения (10) в (9), получим

$$p(\mathbf{Y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}) = \prod_{i\notin\mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) \prod_{i\in\mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) \prod_{i\in\mathcal{I}} p(\mathbf{s}_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}).$$

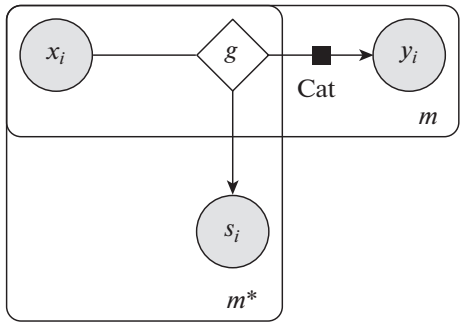

Рис. 1. Вероятностная модель в формате плоских нотаций.

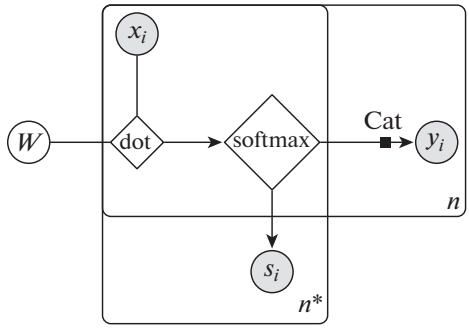

Рис. 2. Вероятностная модель, используемая в синтетическом эксперименте.

Заметим, что $\mathbf{y}_i$ и $\mathbf{s}_i$ зависимы только через переменную $\mathbf{x}_i$, тогда $p\big(\mathbf{s}_i|\mathbf{y}_i,\mathbf{x}_i,\mathbf{g}\big) = p\big(\mathbf{s}_i|\mathbf{x}_i,\mathbf{g}\big)$. Получаем совместное правдоподобие

$$p\big(\mathbf{Y},\mathbf{S}|\mathbf{X},\mathbf{g},\mathcal{I}\big) = \prod_{i\notin\mathcal{I}} p\big(\mathbf{y}_i|\mathbf{x}_i,\mathbf{g}\big) \prod_{i\in\mathcal{I}} p\big(\mathbf{y}_i|\mathbf{x}_i,\mathbf{g}\big) \prod_{i\in\mathcal{I}} p\big(\mathbf{s}_i|\mathbf{x}_i,\mathbf{g}\big). \tag{11}$$

Используя (11), получаем оптимизационную задачу для поиска $\hat{\mathbf{g}}$

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathcal{G}} \prod_{i\notin\mathcal{I}} p\big(\mathbf{y}_i|\mathbf{x}_i,\mathbf{g}\big) \prod_{i\in\mathcal{I}} p\big(\mathbf{y}_i|\mathbf{x}_i,\mathbf{g}\big) \prod_{i\in\mathcal{I}} p\big(\mathbf{s}_i|\mathbf{x}_i,\mathbf{g}\big). \tag{12}$$

Для удобства будем минимизировать логарифм выражения. Тогда из (12) получаем, что

$$\begin{aligned}\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathcal{G}} \sum_{i\notin\mathcal{I}} \log p\big(\mathbf{y}_i|\mathbf{x}_i,\mathbf{g}\big) + (1-\lambda) \sum_{i\in\mathcal{I}} \log p\big(\mathbf{y}_i|\mathbf{x}_i,\mathbf{g}\big) + \\ + \lambda \sum_{i\in\mathcal{I}} \log p\big(\mathbf{s}_i|\mathbf{x}_i,\mathbf{g}\big),\end{aligned} \tag{13}$$

где параметр $\lambda \in [0,1]$ введен для взвешивания ошибок на истинных ответах и ошибок ответов учителя.

На рис. 1 показан вид вероятностной модели в графовой нотации для произвольной функции $\mathbf{g}$. Для каждой реализации $\mathbf{g}$ соответствующий блок требует уточнения. На рис. 2 показана более подробная реализация в случае, когда $\mathbf{g}$ — линейная модель.

5. Обучение с учителем для задачи классификации и регрессии

5.1. Случай классификации

Для задачи многоклассовой классификации рассматриваются вероятностные *предположения*:

1) *рассматривается функция учителя* $\mathbf{f} \in \mathfrak{F}^*_{\text{cl}}$ (6);
2) *рассматривается функция ученика* $\mathbf{g} \in \mathfrak{G}_{\text{cl}}$ (2);
3) *для истинных меток рассматривается категориальное распределение* $p\big(y|\mathbf{x},\mathbf{g}\big) = \text{Cat}\big(\mathbf{g}\big(\mathbf{x}\big)\big)$, *где* $\mathbf{g}\big(\mathbf{x}\big)$ *задает вероятность каждого класса*;
4) *для меток учителя введем плотность распределения*

$$p\big(\mathbf{s}|\mathbf{x},\mathbf{g}\big) = C\prod_{k=1}^{K} g_k\big(\mathbf{x}\big)^{s^k}, \tag{14}$$

где g^k *— вероятность класса* k, *которую предсказывает модель ученика, а* s^k *— вероятность класса* k, *которую предсказывает модель учителя.*

Теорема 1. Пусть вероятность каждого класса отделима от нуля и единицы, т.е. для всех k *выполняется условие*

$$1 > 1 - \varepsilon > g_k\big(\mathbf{x}\big) > \varepsilon > 0.$$

Тогда при

$$C = (-1)^K\,\frac{K^{K/2}}{2^{K(K-1)/2}}\prod_{k=1}^{K} g_k\big(\mathbf{x}\big)\log g_k\big(\mathbf{x}\big)$$

функция $p\big(\mathbf{s}|\mathbf{x},\mathbf{g}\big)$, *определенная в* (14), *является плотностью распределения.*

Доказательство. Во-первых, покажем, что для произвольного вектора ответов $\mathbf{s} \in \mathcal{S}_K$ выполняется $p\big(\mathbf{s}|\mathbf{x},\mathbf{g}\big) \geqslant 0$. Заметим, что для всех k выполняется

$$\log g_k\big(\mathbf{x}\big) < 0,$$

тогда

$$C = \underbrace{\frac{K^{K/2}}{2^{K(K-1)/2}}}_{>0}\prod_{k=1}^{K}\underbrace{g_k\big(\mathbf{x}\big)}_{>\varepsilon}\underbrace{\big(-\log g_k\big(\mathbf{x}\big)\big)}_{>0} > 0.$$

Так как $g_k(\mathbf{x}) > 0$ и $C > 0$, получаем, что $p(\mathbf{s}|\mathbf{x}, \mathbf{g}) \geqslant 0$. Во-вторых, покажем, что интеграл по всему пространству ответов $\mathcal{S}_K$ является конечным:

$$\int\limits_{\mathcal{S}_K} p(\mathbf{s}|\mathbf{x}, \mathbf{g})\, ds = \int\limits_{\mathcal{S}_K} \prod_{k=1}^{K} g_k(\mathbf{x})^{s^k}\, ds = \prod_{k=1}^{K} \int\limits_{\mathcal{S}_K} g_k(\mathbf{x})^{s^k}\, ds = \tag{15}$$

$$= \prod_{k=1}^{K} \int\limits_{0}^{1} \frac{r^{K-1}\sqrt{K}}{(K-1)!\sqrt{2^{K-1}}}\, g_k(\mathbf{x})^{r}\, dr =$$

$$= \prod_{k=1}^{K} \underbrace{\frac{\sqrt{K}}{(K-1)!\sqrt{2^{K-1}}}}_{D} \int\limits_{0}^{1} r^{K-1} g_k(\mathbf{x})^{r}\, dr =$$

$$= D^K \prod_{k=1}^{K} \int\limits_{0}^{1} r^{K-1} \exp\bigl(r \log g_k(\mathbf{x})\bigr)\, dr =$$

$$= (-D)^K \prod_{k=1}^{K} \log g_k(\mathbf{x}) \left(\Gamma(K) - \Gamma\bigl(K, -\log g_k(\mathbf{x})\bigr)\right) =$$

$$= (-D)^K\, (K-1)!^K \prod_{k=1}^{K} \log g_k(\mathbf{x}) \left(1 - g_k(\mathbf{x}) \exp_{K-1}\bigl(-\log g_k(\mathbf{x})\bigr) + g_k(\mathbf{x})\right) =$$

$$= \frac{\left(-\sqrt{K}\right)^K}{2^{K(K-1)/2}} \prod_{k=1}^{K} \log g_k(\mathbf{x}) \left(1 - g_k(\mathbf{x}) \exp_{K-1}\bigl(-\log g_k(\mathbf{x})\bigr) + g_k(\mathbf{x})\right) < \infty,$$

где $\Gamma(K)$ является гамма-функцией, $\Gamma\bigl(K, -\log g_k(\mathbf{x})\bigr)$ является неполной гамма функцией, $\exp_n(x)$ является суммой Тейлора из первых n слагаемых. В рамках приближенных расчетов будем считать, что $\exp_n(x) \approx \exp(x)$, тогда с учетом (15) получаем

$$C(\mathbf{g}, \mathbf{x}) = \int\limits_{\mathcal{S}_K} p(\mathbf{s}|\mathbf{x}, \mathbf{g})\, ds \approx (-1)^K\, \frac{K^{K/2}}{2^{K(K-1)/2}} \prod_{k=1}^{K} g_k(\mathbf{x}) \log g_k(\mathbf{x}). \tag{16}$$

Полученное выражение (16) заканчивает доказательство теоремы 1.

Из теоремы 1 следует, что плотность, введенная для меток учителя, является плотностью распределения. Поэтому можно воспользоваться выражением (13). Используя предположения 1–4 и подставляя в (13), получаем

оптимизационную задачу:

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathcal{G}} \sum_{i\notin\mathcal{I}}\sum_{k=1}^{K} y_i^k \log g_k\big(\mathbf{x}_i\big)\big|_{T=1} +$$

$$+(1-\lambda)\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} y_i^k \log g_k\big(\mathbf{x}_i\big)\big|_{T=1} + \lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} s_{i,k} \log g_k\big(\mathbf{x}_i\big)\big|_{T=T_0} + \tag{17}$$

$$+\lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K}\left(\log g_k\big(\mathbf{x}_i\big)\big|_{T=T_0} + \log\log\frac{1}{g_k\big(\mathbf{x}_i\big)}\big|_{T=T_0}\right).$$

Проанализировав выражение (17), получаем, что первые три слагаемых совпадают со слагаемыми в выражении (4) при $\mathcal{I} = \{1,\ldots,m\}$ и $\lambda = \frac{1}{2}$, а четвертое слагаемое является некоторым регуляризатором, который получен из вида распределения. Анализируя первые три слагаемых в выражении (17) при $T_0 = 1$, получаем сумму кросс-энтропий между двумя распределениями для каждого объекта:

1) первое распределение — это выпуклая комбинация с весами $1-\lambda$ и λ распределения, задаваемого метками объектов $\mathrm{Cat}\big(\mathbf{y}\big)$, и распределения, задаваемого моделью учителя $\mathrm{Cat}\big(\mathbf{s}\big)$;
2) второе распределение — это распределение, задаваемое моделью ученика $\mathrm{Cat}\big(\mathbf{g}\big(\mathbf{x}\big)\big)$.

Следовательно, модель ученика восстанавливает плотность не исходных меток, а новую плотность, которая является выпуклой комбинацией плотности исходных меток и меток учителя.

5.2. Случай регрессии

Для задачи регрессии рассматриваются вероятностные *предположения*:

1) *рассматривается функция учителя* $\mathbf{f} \in \mathfrak{F}^*_{rg}$,

$$\mathfrak{F}^*_{rg} = \Big\{\mathbf{f}|\mathbf{f} = \mathbf{v}^*\big(\mathbf{x}^*\big), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \to \mathbb{R}\Big\},$$

   где $\mathbf{v}^*$ *— дифференцируемая параметрическая функция*;
2) *рассматривается функция ученика* $\mathbf{g} \in \mathfrak{G}_{rg}$,

$$\mathfrak{G}_{rg} = \Big\{\mathbf{g}|\mathbf{g} = \mathbf{z}\big(\mathbf{x}\big), \quad \mathbf{z} : \mathbb{R}^{n} \to \mathbb{R}^{K}\Big\},$$

   где $\mathbf{z}$ *— дифференцируемая параметрическая функция*;
3) *истинные метки имеют нормальное распределение*

$$p\big(y|\mathbf{x},\mathbf{g}\big) = \mathcal{N}\big(y|\mathbf{g}\big(\mathbf{x}\big),\sigma\big);$$

4) *метки учителя имеют распределение*

$$p\big(s|\mathbf{x},\mathbf{g}\big) = \mathcal{N}\big(s|\mathbf{g}\big(\mathbf{x}\big),\sigma_s\big).$$

Используя предположения 1–4 и подставляя в (13), получаем оптимизационную задачу:

$$\hat{g} = \arg\min_{g\in\mathcal{G}} \sum_{i\notin\mathcal{I}} \sigma^2 \left(y_i - \mathbf{g}\big(\mathbf{x}_i\big)\right)^2 + \tag{18}$$
$$+ (1-\lambda) \sum_{i\in\mathcal{I}} \sigma^2 \left(y_i - \mathbf{g}\big(\mathbf{x}_i\big)\right)^2 + \lambda \sum_{i\in\mathcal{I}} \sigma_s^2 \left(s_i - \mathbf{g}\big(\mathbf{x}_i\big)\right)^2.$$

Выражение (18) записано с точностью до аддитивной константы относительно $\mathbf{g}$.

Теорема 2. Пусть множество $\mathcal{G}$ описывает класс линейных функций вида $\mathbf{g}\big(\mathbf{x}\big) = \mathbf{w}^\mathsf{T}\mathbf{x}$. Тогда решение оптимизационной задачи (18) *эквивалентно решению задачи линейной регрессии:*

$$\mathbf{y}'' = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}\big(\mathbf{0}, \boldsymbol{\Sigma}\big), \tag{19}$$

где $\boldsymbol{\Sigma}^{-1} = \operatorname{diag}\big(\boldsymbol{\sigma}'\big)$ и $\mathbf{y}''$ имеют вид:

$$\sigma_i' = \begin{cases} \sigma^2, & \text{если } i \notin \mathcal{I}, \\ (1-\lambda)\,\sigma^2 + \lambda\sigma_s^2, & \text{иначе}, \end{cases}$$

$$\mathbf{y}'' = \boldsymbol{\Sigma}\mathbf{y}', \tag{20}$$

$$y_i' = \begin{cases} \sigma^2 y_i, & \text{если } i \notin \mathcal{I}, \\ (1-\lambda)\,\sigma^2 y_i + \lambda\sigma_s^2 s_i, & \text{иначе}. \end{cases}$$

Доказательство. Обозначим $\mathbf{a}_\mathcal{J} = [a_i | i \in \mathcal{J}]^\mathsf{T}$, где $\mathbf{a}$ — произвольный вектор, а $\mathcal{J}$ — произвольное непустое индексное множество. Подвектор вектора ответов $\mathbf{y}$, для элементов которого доступна привилегированная информация, обозначим $\mathbf{y}_\mathcal{I} = [y_i | i \in \mathcal{I}]^\mathsf{T}$. Аналогично обозначим матрицу $\mathbf{X}_\mathcal{I} = [\mathbf{x}_i | i \in \mathcal{I}]^\mathsf{T}$.

В случае линейной модели $\mathbf{g}\big(\mathbf{x}\big) = \mathbf{w}^\mathsf{T}\mathbf{x}$ выражение (18) принимает вид:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \sigma^2 \left(\mathbf{y}_{\bar{\mathcal{I}}} - \mathbf{X}_{\bar{\mathcal{I}}}\mathbf{w}\right)^\mathsf{T} \left(\mathbf{y}_{\bar{\mathcal{I}}} - \mathbf{X}_{\bar{\mathcal{I}}}\mathbf{w}\right) +$$
$$+ \sigma^2 (1-\lambda) \left(\mathbf{y}_\mathcal{I} - \mathbf{X}_\mathcal{I}\mathbf{w}\right)^\mathsf{T} \left(\mathbf{y}_\mathcal{I} - \mathbf{X}_\mathcal{I}\mathbf{w}\right) + \sigma_s^2\lambda \left(\mathbf{s}_\mathcal{I} - \mathbf{X}_\mathcal{I}\mathbf{w}\right)^\mathsf{T} \left(\mathbf{s}_\mathcal{I} - \mathbf{X}_\mathcal{I}\mathbf{w}\right).$$

Раскроем скобки и сгруппируем:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \sigma^2 \left(\mathbf{w}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}}\mathbf{w} - 2\mathbf{y}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}}\mathbf{w}\right) +$$
$$+ (1-\lambda)\,\sigma^2 \left(\mathbf{w}^\mathsf{T}\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\mathbf{w} - 2\mathbf{y}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\mathbf{w}\right) + \lambda\sigma_s^2 \left(\mathbf{w}^\mathsf{T}\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\mathbf{w} - 2\mathbf{s}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\mathbf{w}\right).$$

Продифференцируем выражение, приравняем к нулю и сгруппируем элементы:

$$\left(\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}} + (1-\lambda)\,\sigma^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I} + \lambda\sigma_s^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\right)\mathbf{w} = \tag{21}$$
$$= 2\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{y}_{\bar{\mathcal{I}}} + 2\,(1-\lambda)\,\sigma^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{y}_\mathcal{I} + 2\lambda\sigma_s^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{s}_\mathcal{I}.$$

Воспользуемся равенствами:

$$\begin{gathered}\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^{\mathsf{T}}\mathbf{X}_{\bar{\mathcal{I}}} + (1-\lambda)\,\sigma^2\mathbf{X}_{\mathcal{I}}^{\mathsf{T}}\mathbf{X}_{\mathcal{I}} + \lambda\sigma_s^2\mathbf{X}_{\mathcal{I}}^{\mathsf{T}}\mathbf{X}_{\mathcal{I}} = \mathbf{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X},\\ 2\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^{\mathsf{T}}\mathbf{y}_{\bar{\mathcal{I}}} + 2\,(1-\lambda)\,\sigma^2\mathbf{X}_{\mathcal{I}}^{\mathsf{T}}\mathbf{y}_{\mathcal{I}} + 2\lambda\sigma_s^2\mathbf{X}_{\mathcal{I}}^{\mathsf{T}}\mathbf{s}_{\mathcal{I}} = 2\mathbf{X}\mathbf{y}',\end{gathered} \tag{22}$$

где $\boldsymbol{\Sigma}$ и $\mathbf{y}'$ из условия задачи (20).

Подставляя (22) в (21), получаем:

$$\mathbf{w} = 2\left(\mathbf{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{y}'',$$

что соответствует решению задачи (19). Теорема 2 доказана.

Теорема 2 показывает, что обучение с учителем для задачи регрессии можно свести к задаче оптимизации в линейной регрессии.

6. Вычислительный эксперимент

Проводится вычислительный эксперимент для анализа моделей, которые получены путем дистилляции модели учителя в модель ученика. Как показано в теореме 2, задачу регрессии с учителем можно свести к задаче регрессии без учителя, поэтому в эксперименте рассматривается только случай классификации. Во всех частях вычислительного эксперимента для поиска оптимальных параметров нейросетей использовался градиентный метод оптимизации Adam [21].

6.1. Выборка FashionMNIST

Эксперимент проводился для задачи классификации для выборки FashionMNIST [17]. В качестве модели учителя $\mathbf{f}$ рассматривается нейросеть с двумя сверточными слоями и с тремя полносвязными слоями, в качестве функции активации рассматривается ReLu. Модель учителя содержит 30 тысяч обучаемых параметров. В качестве модели ученика рассматривается модель логистической регрессии для многоклассовой классификации. Модель ученика содержит 7850 обучаемых параметров.

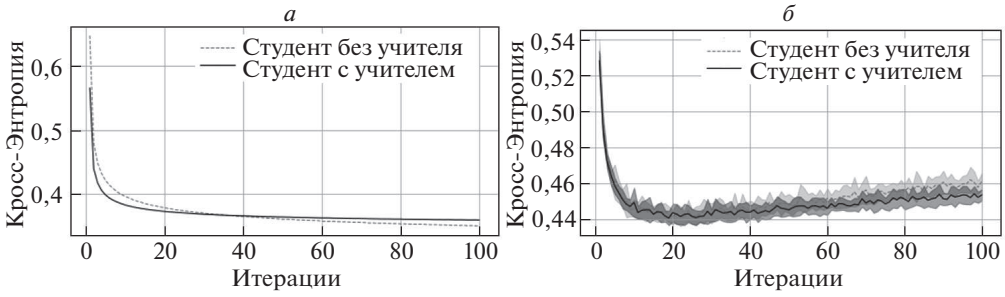

Рис. 3. Зависимость кросс-энтропии между истинными метками и предсказанными учеником вероятностями классов: *а* — на обучающей выборке; *б* — на тестовой выборке.

На рис. 3 показан график зависимости кросс–энтропии между истинными метками объектов и вероятностями, которые предсказывает модель ученика. На графике сравнивается модель, которая обучалась без учителя (в задаче оптимизации (17) присутствует только первое слагаемое) с моделью, которая была получена путем дистилляции модели нейросети в линейную модель. Из графика видно, что обе модели начинают переобучаться после 30-й итерации. Но модель, которая получена путем дистилляции, переобучается не так быстро: ошибка на тестовой выборке растет медленнее, а на обучающей выборке падает также медленнее.

В таблице показано, что для выборки FashionMnist итоговые модели ученика с учителем и без учителя сравнимы по точности и кросс-энтропийной ошибке, если учитывать дисперсию этих величин.

6.2. Синтетический эксперимент

Проанализируем модель на синтетической выборке. Выборка построена следующим образом:

$$\mathbf{W} = \left[\mathcal{N}\left(w_{jk}|0,1\right)\right]_{n\times K}, \quad \mathbf{X} = \left[\mathcal{N}\left(x_{ij}|0,1\right)\right]_{m\times n},$$
$$\mathbf{S} = \operatorname{softmax}\left(\mathbf{X}\mathbf{W}\right), \quad \mathbf{y} = \left[\operatorname{Cat}\left(y_i|\mathbf{s}_i\right)\right],$$

где функция softmax берется построчно. Строки матрицы $\mathbf{S}$ будем рассматривать как предсказание учителя, т.е. учитель знает истинные вероятности каждого класса. На рис. 2 показана вероятностная модель в графовой нотации. В эксперименте число признаков $n = 10$, число классов $K = 3$, для обучения было сгенерировано $m_{\text{train}} = 1000$ и $m_{\text{test}} = 100$ объектов.

На рис. 4 показано распределение по классам для 20 объектов из обучающей выборки. Каждому столбцу на графике соответствует объект, а каждой строке соответствует вероятность класса. Видно, что для каждого рассмотренного объекта вероятности разных классов близки. Получается, что если в качестве истинных меток взять класс с максимальной вероятностью, то выборка будет сильно зашумленной и модель будет описывать эти данные некорректно.

Построим в качестве ученика линейную модель, которая минимизирует кросс-энтропийную (первое слагаемое в формуле (17)). Представление данной модели в виде графовой модели показано на рис. 2.

На рис. 5 показано распределение вероятностей классов, которое предсказала модель. Видно, что полученное распределение не соответствует истинному, так как модель сосредотачивает всю вероятность в одном классе.

Рассмотрим модель, которая учитывает информацию об истинных распределениях на классах для каждого объекта. Для этого будем минимизировать первые три слагаемых в формуле (17) при $T_0 = 1$ и $\lambda = 0{,}75$. В качестве меток учителя $s_{i,k}$ использовались истинные вероятности для каждого класса данного объекта. На рис. 6 показано распределение, которое дала модель. В данном случае видно, что распределения являются сглаженными. Концентрации всей вероятности в одном классе не наблюдается.

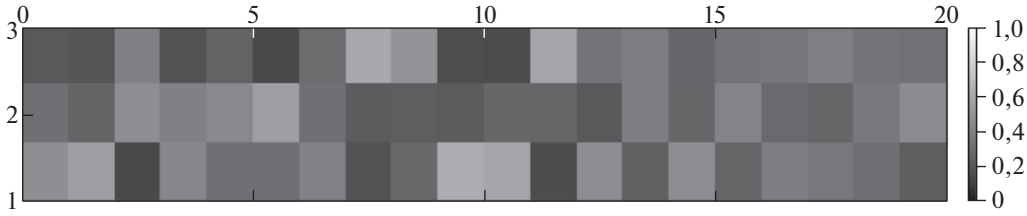

Рис. 4. Истинное распределение объектов по классам.

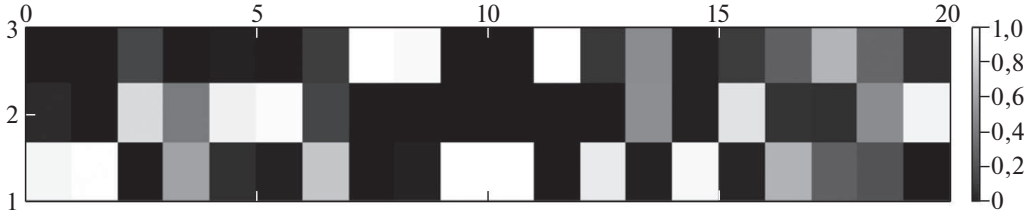

Рис. 5. Распределение, предсказанное моделью без использования информации об истинном распределении на классах.

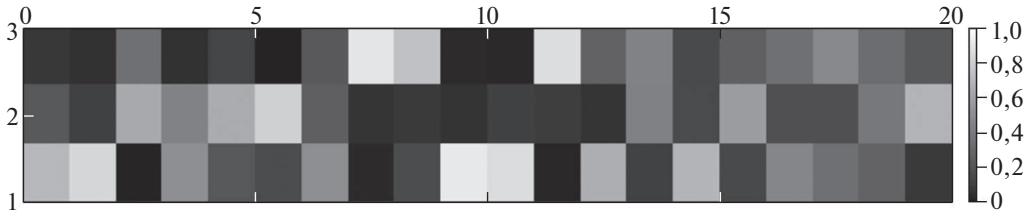

Рис. 6. Распределение, предсказанное моделью с использованием информации об истинном распределении на классах.

Заметим, что в данном примере предполагается, что модель учителя учитывает не только метки классов, но и распределение на метках классов, в то время как в выборке $\{\mathbf{X}, \mathbf{y}\}$ имеются только точечные оценки в виде меток.

В данном примере используются истинные распределения в качестве предсказаний учителя, но их можно заменить предсказаниями модели учителя, которая предсказывает не только сами метки, но и их распределение для каждого объекта.

На рис. 7 показана зависимость вероятности верного класса от температуры T и параметра доверия λ для одного из объектов из тестовой выборки. На рис. 7 видно, что изменение температуры T влечет изменение концентрации вероятностной меры. При уменьшении параметра температуры и приближении его к нулю наблюдаем, что вероятность одного из классов приближается к единице, а остальных классов — к нулю. С другой стороны, при увеличении параметра температуры вероятности классов сглаживаются и распределение классов для каждого объекта становится близким к равномерному.

В таблице в колонке “Кросс-энтропийная ошибка с реальными вероятностями” показано сравнение кросс-энтропии в случае, если в качестве истин-

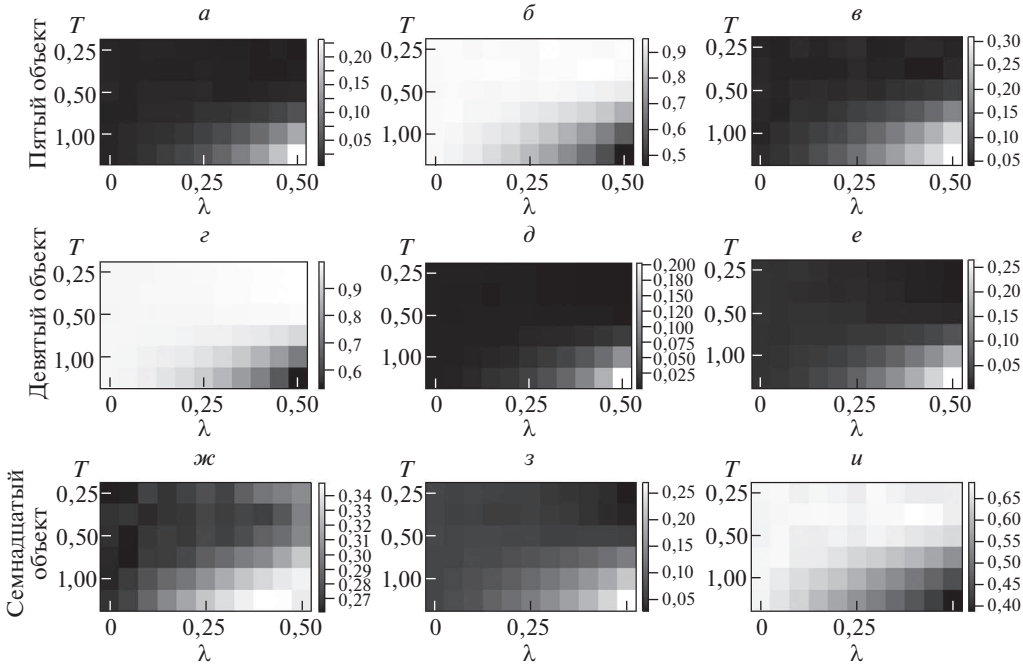

Рис. 7. Иллюстрация распределения вероятности предсказания классов при различных значениях λ и T.

ных вероятностей меток рассмотреть не onehot-кодированные вероятности классов, а истинные вероятности:

$$\mathcal{L}_{\text{real}}(\mathbf{g}) = -\sum_{i=1}^{m}\sum_{k=1}^{K} s_i^k \log g_k(\mathbf{x}_i),$$

где $\mathbf{g}$ — модель ученика. Видно, что модель с учителем лучше аппроксимирует истинные вероятности классов. Также в таблице представлено среднее значение разницы максимальной вероятности с минимальной вероятностью для каждого объекта:

$$\mathcal{L}_{\text{maxmin}}(\mathbf{g}) = \frac{1}{m}\sum_{i=1}^{m}\left(\max_k g_k(\mathbf{x}_i) - \min_k g_k(\mathbf{x}_i)\right).$$

Видно, что модель учителя имеет меньшую разницу между вероятностями классов, т.е. вероятности классов не концентрируются в одном классе.

6.3. Выборка Twitter Sentiment Analysis

Проводится эксперимент на выборке Twitter Sentiment Analysis. Данная выборка содержит короткие сообщения, для которых требуется предсказать эмоциональный окрас: содержит твит позитивный окрас или негативный. Выборка разделена на 1,18 млн твитов для обучения и 0,35 млн твитов для

Таблица. Сводная таблица результатов вычислительного эксперимента

Выборка	Модель	Кросс-Энтропийная ошибка	Кросс-Энтропийная ошибка с реальными вероятностями	Вероятностная разница	Точность	Число параметров
Fashion-Mnist	с учителем	$0{,}453 \pm 0{,}003$	–	$0{,}84 \pm 0{,}13$	$0{,}842 \pm 0{,}002$	7850
	без учителя	$0{,}461 \pm 0{,}005$	–	$0{,}86 \pm 0{,}18$	$0{,}841 \pm 0{,}002$	7850
Systetic	с учителем	$0{,}618 \pm 0{,}001$	$1{,}17 \pm 0{,}05$	$0{,}45 \pm 0{,}20$	$0{,}828 \pm 0{,}002$	33
	без учителя	$0{,}422 \pm 0{,}002$	$2{,}64 \pm 0{,}02$	$0{,}75 \pm 0{,}22$	$0{,}831 \pm 0{,}001$	33
Twiter	с учителем	$0{,}489 \pm 0{,}003$	–	$0{,}79 \pm 0{,}17$	$0{,}764 \pm 0{,}005$	1538
	без учителя	$0{,}501 \pm 0{,}006$	–	$0{,}83 \pm 0{,}22$	$0{,}747 \pm 0{,}004$	1538

тестирования. В твитах была выполнена предобработка: все твиты были переведены в нижний регистр, все никнеймы вида “@andrey” были заменены на токен “name”, все цифры были заменены на токен “number”.

Результаты данной части эксперимента показаны в таблице. В качестве модели учителя использовалась модель Bi-LSTM с линейным слоем на выходе. В качестве векторного представления токенов обучалась матрица параметров. В ней каждая строка соответствует токену из обучающей выборки. Суммарное число обучаемых параметров модели учителя составляет более 30 млн. Обученная модель учителя имеет точность предсказания 0,835. В качестве модели ученика рассматривается линейная модель с 1538 параметрами, где в качестве векторного представления предложения рассматривается выход предобученной модели BERT с размерностью векторного пространства 768. Признаковое описание модели учителя и модели ученика различаются. Модель учителя в качестве признакового описания рассматривает исходные слова в предложении. Модель ученика в качестве признакового описания использует готовое векторное представление предложения, которое получено при помощи модели BERT.

В таблице показано качество модели ученика с использованием предсказания модели учителя и без него. В рамках данных результатов качество модели ученика с дистилляцией выше, чем модели ученика без дистилляции, но разница находится в пределах погрешности, что не позволяет говорить о значительных улучшениях качества.

Программное обеспечение для проведения экспериментов и проверки результатов находится в [22].

7. Заключение

В данной статье проанализирована задача обучения модели ученика с помощью модели учителя. Исследован метод дистилляции и привилегированно-

го обучения. Предложено вероятностное обоснование дистилляции. Введены вероятностные предположения, описывающие дистилляцию моделей. В рамках данных вероятностных предположений проанализированы модели для задачи классификации и регрессии. Результат анализа сформулирован в виде теорем 1 и 2.

Теорема 2 показала, что обучение линейной регрессии с учителем эквивалентно замене обучающей выборки и вероятностных предположений о распределении истинных ответов. Для задачи классификации ответы учителя дают дополнительную информацию в виде распределения классов для каждого объекта из обучающей выборки. Данная информация не может быть представлена в виде задачи классификации. Требуется ввести распределение, которое представлено в теореме 1.

В вычислительном эксперименте сравниваются модели ученика, которые обучены с использованием модели учителя и без него. В таблице показаны результаты вычислительного эксперимента для разных выборок. Показано, что точность аппроксимации выборки учеником улучшается при использовании модели учителя. Задача регрессии не приведена в вычислительном эксперименте, так как в теореме 2 была показана ее эквивалентность задаче линейной регрессии. Для задачи классификации проведен вычислительный эксперимент. Из вычислительного эксперимента видно, что дистилляция влияет на распределение классов в рамках одного объекта. Вероятности классов для каждого объекта являются более разреженными, а не концентрируются в одном классе. Данное свойство хорошо видно в синтетической выборке, так как она генерировалась с максимальной дисперсией в вероятностях классов.

Основным результатом данной статьи является вероятностная интерпретация задачи дистилляции. Рассмотрен частный случай, когда признаковые описания модели учителя и ученика совпадают. В рамках вычислительного эксперимента проведен анализ ответов модели ученика с использованием модели учителя и без нее. Из результатов эксперимента видно, что модель ученика наследует распределение вероятностей по классам от модели учителя. Когда модель учителя адекватно описывает данные, описание данных моделью ученика также улучшается, что показано в вычислительном эксперименте на синтетических данных.

В дальнейшем предполагается обобщить метод максимального правдоподобия для дистилляции моделей с помощью байесовского подхода выбора моделей машинного обучения. Также в рамках байесовского подхода планируется развить методы повышения качества не только для задачи классификации, но и для задачи регрессии.

СПИСОК ЛИТЕРАТУРЫ

1. *Vaswani A., Gomez A., Jones L., Kaiser L., Parmar N., Polosukhin I., Shazeer N., Uszkoreit J.* Attention Is All You Need // Advances in Neural Information Processing Syst. 2017. V. 5. P. 6000–6010.
2. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. 2019 Conf. North Amer-

ican Chapter of the Association for Computational Linguistics: Human Language Technologies. Minnesota. 2019. V. 1. P. 4171–4186.

3. *He K., Ren S., Sun J., Zhang X.* Deep Residual Learning for Image Recognition // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas. 2016. P. 770–778.
4. *Бахтеев О.Ю., Стрижов В.В.* Выбор моделей глубокого обучения субоптимальной сложности // АиТ. 2018. № 8. С. 129–147.

   Bakhteev O.Yu., Strijov V.V. Deep Learning Model Selection of Suboptimal Complexity // Automat. Remote Control. 2018. V. 79. P. 1474–1488.
5. *Hinton G., Dean J., Vinyals O.* Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. 2015.
6. *Bucilu C., Caruana R., Mizil A.* Model compression // Proc. ACM SIGKDD Conf. on Knowledge Discovery and Data mining. Philadelphia. 2006. P. 535–541.
7. *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying Distillation and Privileged Information // Int. Conf. on Learning Representations. Puerto Rico. 2016.
8. *Tang Z., Wang D., Zhang Z.* Recurrent neural network training with dark knowledge transfer // Proc. IEEE Conf. on Acoustics, Speech and Signal Processing. Shanghai. 2016. V. 2. P. 5900–5904.
9. *Darrell T., Hoffman J., Saenko K., Tzeng E.* Simultaneous deep transfer across domains and tasks // Proc. IEEE Conf. on Computer Vision. Santiago. 2015. V. 2. P. 4068–4076.
10. *Ahn S., Dai Z., Damianou A., Hu S., Lawrence N.* Variational information distillation for knowledge transfer // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach. 2019. P. 9163–9171.
11. *Burges C., Cortes C., LeCun Y.* The MNIST dataset of handwritten digits. 1998. http://yann.lecun.com/exdb/mnist/index.html.
12. *Che Z., Chen Y., Guoping H., Liu W., Wang T., Ziqing Y.* TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing // Proc. 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Online. 2020.
13. *Huang Z., Naiyan W.* Like What You Like: Knowledge Distill via Neuron Selectivity Transfer // arXiv:1707.01219. 2017.
14. *Fu T., Lei Z., Liao S., Mei T., Wang S., Wang X.* Exclusivity-Consistency Regularized Knowledge Distillation for Face Recognition // Lect. Notes in Computer Sci. 2020. V. 1 P. 23–69.
15. *Vapnik V., Izmailov R.* Learning Using Privileged Information: Similarity Control and Knowledge Transfer // J. of Machine Learning Research. 2015. V. 16. P. 2023–2049.
16. *Ivakhnenko A., Madala H.* Inductive Learning Algorithms for Complex Systems Modeling. Boca Raton: CRC Press Inc. 1994.
17. *Rasul K., Vollgraf R., Xiao H.* Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms // arXiv preprint arXiv:1708.07747. 2017.
18. *Kozareva Z., Nakov P., Ritter A., Rosenthal S., Stoyanov V., Wilson T.* SemEval-2013 Task 2: Sentiment Analysis in Twitter // Proc. Seventh Int. Workshop on Semantic Evaluation (SemEval 2013). Atlanta. 2013. P. 312–320.
19. *Boser B., Denker J., Henderson D., Howard R., Hubbard W., Jackel L., LeCun Y.* Backpropagation Applied to Handwritten Zip Code Recognition // Neural Computation. 1989. V. 1. No. 4. P. 541–551.

20. *Hochreiter S., Schmidhuber J.* Long Short-Term Memory // Neural Computation. 1997. V. 9. No. 8. P. 1735–1780.
21. *Ba J., Kingma D.* Adam: A Method for Stochastic Optimization // Int. Conf. on Learning Representations. San Diego. 2014.
22. Код вычислительного эксперимента. URL: `https://github.com/andriygav/PrivilegeLearning`

Статья представлена к публикации членом редколлегии О.П. Кузнецовым.

Поступила в редакцию 29.08.2020

После доработки 14.08.2021

Принята к публикации 29.08.2021