© 2020 г. А.В. ГРАБОВОЙ
В.В. СТРИЖОВ, д-р физ-мат. наук
(Московский физико-технический институт, Москва)

Анализ моделей привилегированного обучения и дистиляции[1]

Данная работа посвящена методам понижения сложности аппроксимирующих моделей. Предлагается вероятностное обоснование методов дистилляции и привилегированного обучения. В работе приведены общие выводы для произвольной параметрической функции с наперед заданной структурой. Показано теоретическое обоснование для частных случаев: линейной и логистической регрессии. Теоретические результаты анализируются в вычислительном эксперименте на синтетических выборках и реальных данных. В качестве реальных данных рассматривается выборки FashionMNIST и Twitter Sentiment Analysis.

1. Введение

Повышение точности аппроксимации в задачах машинного обучения влечет за собой повышение сложности моделей и как следствие снижает их интерпретируемость. Примером такого усложнения являются следующие модели: трансформеры [4], BERT [5], ResNet [3] а также ансамбли этих моделей.

При построении модели машинного обучения используется два свойства: сложность модели и точность аппроксимации модели. Сложность влияет на время, которое модель требуется для принятия решения, а также на интерпретируемость модели, следовательно модель которая имеют меньшую сложность является более предпочтительной [12]. С другой стороны точность аппроксимации модели нужно максимизировать. В данной работе рассматривается метод *дистилляции* модели. Данные метод позволяет строить новые модели на основе ранее обученых моделей.

Определение 1. *Дистилляция модели — уменьшение сложности модели путем выбора модели в множестве более простых моделей с использованием ответов более сложной модели.*

В работе [8] Дж. Е. Хинттоном рассматривается метод дистилляции моделей машинного обучения для задачи классификации. В работе проведен ряд экспериментов, в которых проводилась дистилляции моделей для разных задач машинного обучения. Эксперимент на выборке MNIST [9], в котором избыточно сложно нейросеть была дистиллирована в нейросеть меньшей сложности. Эксперимент по Speech Recognition, в котором ансамбль моделей был *дистиллирован* в одну модель. Также в работе [8] был проведен эксперимент по обучению экспертных моделей на основе одной большой модели.

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

[1]Работа выполнена при поддержке ...(грант № ...).

В работе [6] В. Н. Вапником введено понятия *привилегированной информации*. В работе [7] метод дистилляции [8] используется вместе с привилегированным обучениям [6]. В предложенном методе на первом этапе обучается модель *учителя* в пространстве привилегированной информации, после чего обучается модель *ученика* в исходном признаковом пространстве используя *дистилляцию* [8]. Для обучения строится функция ошибки специального вида, анализируемая в данной работе. Эта функция состоит из нескольких слагаемых, включая ошибки учителя, ученика и регуляризирующие элементы. Первые варианты подобной функции ошибки были предложены А. Г. Ивахненко [14].

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

Определение 4. *Ученик — модель, которая выбирается согласно какого-либо критерия.*

В данной работе предлагается рассмотреть вероятностный подход к решению задачи дистилляции модели и задачи привилегированного обучения. Подход обобщается на случай, когда привилегированная информация доступна не для всех объектов из обучающей выборки. В рамках вероятностного подхода предлагается анализ и обобщение функции ошибки [8, 7]. Рассматриваются частные задачи классификации и регрессии [14].

В рамках вычислительного эксперимента анализируются методы использующие и не использующие модель учителя при обучение модели ученика. Для анализа используются реальные выборки для задачи классификации изображений FashionMNIST [10] и для задачи классификации текстов Twitter Sentiment Analysis [13]. Выборка FashionMNIST использовалась вместо общепринятой выборки MNIST, так как последняя имеет приемлемое качество аппроксимации даже для линейного классификатора. Вычислительный эксперимент использует модели разной сложности: линейная модель, полносвязная нейронная сеть, сверточная нейронная сеть [1], модель Bi-LSTM [2] и модель BERT [5].

2. Постановка задачи обучения с учителем

Задано множество объектов $\mathbf{\Omega}$ и множество целевых переменных $\mathbb{Y}$. Множество $\mathbb{Y} = \{1, \cdots, K\}$ 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 $\mathbf{y}_i = \mathbf{y}\bigl(\omega_i\bigr)$. Множество целевых переменных для всех объектов обозначим $\mathbf{Y}$. Для множества $\mathbf{\Omega}$ задано отображение в некоторое признаковое пространство $\mathbb{R}^n$:

$$\varphi : \mathbf{\Omega} \to \mathbb{R}^n, \quad |\mathbf{\Omega}| = m, \tag{1}$$

где n размерность признакового пространства, а m количество 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:

$$\varphi^* : \mathbf{\Omega}^* \to \mathbb{R}^{n^*}, \quad |\mathbf{\Omega}^*| = m^*, \tag{2}$$

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 $\mathcal{I}$:

$$\mathcal{I} = \{1 \leqslant i \leqslant m | \text{ для } i\text{-го объекта задана привилегированная информация}\}, \tag{3}$$

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

Пусть на множестве привилегированных признаков задана функция учителя $\mathbf{f}(\mathbf{x}^*)$:

$$\mathbf{f} : \mathbb{R}^{n^*} \to \mathbb{Y}^*, \tag{4}$$

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

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

$$\mathfrak{G} = \{\mathbf{g} | \mathbf{g} : \mathbb{R}^n \to \mathbb{Y}^*\}, \tag{5}$$

например для задачи классификации множество $\mathfrak{G}$ может быть параметрическим семейством функций линейных моделей:

$$\mathfrak{G}_{\text{lin,cl}} = \left\{\mathbf{g}(\mathbf{W}, \mathbf{x}) | \mathbf{g}(\mathbf{W}, \mathbf{x}) = \mathbf{softmax}(\mathbf{W}\mathbf{x}), \quad \mathbf{W} \in \mathbb{R}^{n \times K}\right\}. \tag{6}$$

3. Постановка задачи: Хинтона и Вапника

Рассмотрим описание метода предложеного в работах [8, 7]. В рамках данных работ предполагается, что для всех данных доступна привилегированная информация $\mathcal{I} = \{1, 2, \cdots, m\}$. В работе [8] решается задача классификации вида:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \qquad \mathbf{x}_i \in \mathbb{R}^n, \quad y_i \in \mathbb{Y} = \{1, \cdots, K\}, \tag{7}$$

где y_i — это класс объекта, также будем обозначать $\mathbf{y}_i$ вектором вероятности для класса y_i.

В постановке Хинтона рассматривается параметрическое семейство функций:

$$\mathfrak{G}_{\text{cl}} = \left\{\mathbf{g} | \mathbf{g} = \text{softmax}(\mathbf{z}(\mathbf{x})/T), \quad \mathbf{z} : \mathbb{R}^n \to \mathbb{R}^K\right\}, \tag{8}$$

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

$$\mathfrak{F}_{\text{cl}} = \left\{\mathbf{f} | \mathbf{f} = \text{softmax}(\mathbf{v}(\mathbf{x})/T), \quad \mathbf{v} : \mathbb{R}^n \to \mathbb{R}^K\right\}, \tag{9}$$

где $\mathbf{v}$ — это дифференцируемая параметрическая функция заданной структуры, T — параметр температуры. Параметр температуры T имеет следующие свойства:

1) при $T \to 0$ получаем вектор, в котором один из классов имеет единичную 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 следующий вид:

$$\begin{aligned}\mathcal{L}_{st}(\mathbf{g}) = &-\underbrace{\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log \mathbf{g}(\mathbf{x}_i)\big|_{T=1}}_{\text{исходная функция потерь}} \\ &-\underbrace{\sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{f}(\mathbf{x}_i)\big|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)\big|_{T=T_0}}_{\text{слагаемое дистилляция}},\end{aligned} \tag{10}$$

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

Получаем оптимизационную задачу:

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g}\in\mathfrak{G}_{\text{cl}}} \mathcal{L}_{st}(\mathbf{g}). \tag{11}$$

Работа [7] обобщает метод предложенный в работе [8]. Решение задачи оптимизации (11) зависит только от вектора ответов модели учителя $\mathbf{f}$. Следовательно признаковые пространства учителя и ученика могут различаться. В этом случае получаем следующую постановку задачи:

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{x}_i^*, y_i)\}_{i=1}^{m}, \qquad \mathbf{x}_i \in \mathbb{R}^n, \quad \mathbf{x}_i^* \in \mathbb{R}^{n^*}, \quad y_i \in \{1, \cdots, K\}, \tag{12}$$

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

$$\mathfrak{F}_{\text{cl}}^* = \left\{\mathbf{f} | \mathbf{f} = \text{softmax}\big(\mathbf{v}^*(\mathbf{x}^*)/T\big), \quad \mathbf{v}^*: \mathbb{R}^{n^*} \to \mathbb{R}^K\right\}, \tag{13}$$

где $\mathbf{v}^*$ — это дифференцируемая параметрическая функция заданной структуры, T — параметр температуры. Множество моделей $\mathfrak{F}_{cl}^*$ отличается от множества моделей $\mathfrak{F}_{cl}$ из выражения (9). В множестве $\mathfrak{F}_{cl}$ модели используют пространство исходных признаков, а в множестве $\mathfrak{F}_{cl}^*$ модели используют пространство привилегированных признаков. Функция потерь (10) в случае модели учителя $\mathbf{f} \in \mathfrak{F}_{cl}^*$ переписывается в следующем виде:

$$\mathcal{L}_{st}(\mathbf{g}) = -\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log \mathbf{g}(\mathbf{x}_i)\big|_{T=1} - \sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{f}(\mathbf{x}_i^*)\big|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)\big|_{T=T_0}, \tag{14}$$

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

Требуется построить модель, которая использует привилегированную информацию $\mathbf{x}_i^*$ при обучении. Для этого рассмотрим двухэтапную модель обучения предложенную в работе [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 [8].

Модель ученика — это функция, которая минимизирует (14). Модель учителя — это функция, которая минимизирует кросс–энтропийную функции ошибки:

$$\mathcal{L}_{th}\bigl(\mathbf{f}\bigr) = -\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log \mathbf{f}\bigl(\mathbf{x}_i^*\bigr). \tag{15}$$

4. Постановка задачи: вероятностный подход

4.1. Метод максимального правдоподобия

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

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathfrak{G}} \prod_{i=1}^{N} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr), \tag{16}$$

где множество $\mathfrak{G}$ задается (5).

4.2. Подход дистилляции модели учителя в модель ученика

Рассмотрим вероятностную постановку, в которой должны быть выполнены ограничения:

1) задано распределение целевой переменной $p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr)$;

2) задано совместное распределение целевой переменной и ответов модели учителя $p\bigl(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr)$;

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 (16).

Рассмотрим совместное правдоподобие истинных меток и меток учителя:

$$p\bigl(\mathbf{Y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}\bigr) = \prod_{i\not\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr). \tag{17}$$

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

$$p\bigl(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr) = p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr)p\bigl(\mathbf{s}_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}\bigr) \tag{18}$$

Подставляя выражения (18) в (17) получаем.

$$p\bigl(\mathbf{Y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}\bigr) = \prod_{i\not\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{s}_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}\bigr). \tag{19}$$

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

$$p\bigl(\mathbf{Y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}\bigr) = \prod_{i\not\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr). \tag{20}$$

Используя (20) получаем следующую оптимизационную задачу для поиска $\hat{\mathbf{g}}$

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathcal{G}} \prod_{i\not\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr). \tag{21}$$

Для удобства, будем минимизировать логарифм, тогда из (21) получаем:

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathcal{G}} \sum_{i\not\in\mathcal{I}} \log p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) + (1-\lambda) \sum_{i\in\mathcal{I}} \log p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) + \lambda \sum_{i\in\mathcal{I}} \log p\bigl(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr), \tag{22}$$

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

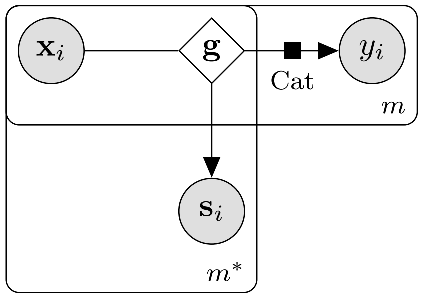

Рис. 1. Вероятностная модель в формате плоских нотаций.

На рис. 1 показан вид вероятностной модели в графовой нотации, для произвольной функции $\mathbf{g}$. Для каждой реализации $\mathbf{g}$ соответсвующий блок требует уточнение. На рис 3 показана более подробная реализация в случае, когда модель $\mathbf{g}$ это линейная модель.

5. Обучение с учителем для задачи классификации и регрессии

5.1. Случай классификации

Для задачи многоклассовой классификации рассматриваются следующие вероятностные предположения:

1) рассматривается функция учителя $\mathbf{f} \in \mathfrak{F}^*_{\text{cl}}$ (13);

2) рассматривается функция ученика следующего вида $\mathbf{g} \in \mathfrak{G}_{\text{cl}}$ (8);

3) для истинных меток рассматривается категориальное распределение $p\bigl(y|\mathbf{x}, \mathbf{g}\bigr) = \text{Cat}\bigl(\mathbf{g}(\mathbf{x})\bigr)$, где $\mathbf{g}(\mathbf{x})$ задает вероятность каждого класса;

4) для меток учителя введем плотность распределения

$$p\bigl(\mathbf{s}|\mathbf{x}, \mathbf{g}\bigr) = C \prod_{k=1}^{K} g_k\bigl(\mathbf{x}\bigr)^{s^k}, \tag{23}$$

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

Теорема 1. Пусть вероятнось каждого класса отделима от нуля и единицы, то есть для всех k выполняется $1 > 1 - \varepsilon > g_k\bigl(\mathbf{x}\bigr) > \varepsilon > 0$, тогда при

$$C = (-1)^K \frac{K^{K/2}}{2^{K(K-1)/2}} \prod_{k=1}^{K} g_k\bigl(\mathbf{x}\bigr) \log g_k\bigl(\mathbf{x}\bigr) \tag{24}$$

функция $p(\mathbf{s}|\mathbf{x},\mathbf{g})$ определенная в (23) является плотностью распределения.

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

$$C = \underbrace{\frac{K^{K/2}}{2^{K(K-1)/2}}}_{>0} \prod_{k=1}^{K} \underbrace{g_k(\mathbf{x})}_{>\varepsilon} \underbrace{\left(-\log g_k(\mathbf{x})\right)}_{>0} > 0, \tag{25}$$

тогда с учетом того, что $g_k(\mathbf{x}) > 0$ 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:

(26)

$$\begin{aligned}
\int_{\mathcal{S}_K} p(\mathbf{s}|\mathbf{x},\mathbf{g}) ds &= \int_{\mathcal{S}_K} \prod_{k=1}^{K} g_k(\mathbf{x})^{s^k} ds = \prod_{k=1}^{K} \int_{\mathcal{S}_K} g_k(\mathbf{x})^{s^k} ds \\
&= \prod_{k=1}^{K} \int_0^1 \frac{r^{K-1}\sqrt{K}}{(K-1)!\sqrt{2^{K-1}}} g_k(\mathbf{x})^r dr = \prod_{k=1}^{K} \underbrace{\frac{\sqrt{K}}{(K-1)!\sqrt{2^{K-1}}}}_{D} \int_0^1 r^{K-1} g_k(\mathbf{x})^r dr \\
&= D^K \prod_{k=1}^{K} \int_0^1 r^{K-1} \exp\left(r \log g_k(\mathbf{x})\right) dr \\
&= (-D)^K \prod_{k=1}^{K} \log g_k(\mathbf{x}) \left(\Gamma(K) - \Gamma\left(K, -\log g_k(\mathbf{x})\right)\right) \\
&= (-D)^K (K-1)!^K \prod_{k=1}^{K} \log g_k(\mathbf{x}) \left(1 - g_k(\mathbf{x}) \exp_{K-1}\left(-\log g_k(\mathbf{x})\right) + g_k(\mathbf{x})\right) \\
&= \frac{\left(-\sqrt{K}\right)^K}{2^{K(K-1)/2}} \prod_{k=1}^{K} \log g_k(\mathbf{x}) \left(1 - g_k(\mathbf{x}) \exp_{K-1}\left(-\log g_k(\mathbf{x})\right) + g_k(\mathbf{x})\right) < \infty,
\end{aligned}$$

где $\Gamma(K)$ является гамма функцией, $\Gamma\left(K, -\log g_k(\mathbf{x})\right)$ 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 (26) получаем:

$$C(\mathbf{g},\mathbf{x}) = \int_{\mathcal{S}_K} p(\mathbf{s}|\mathbf{x},\mathbf{g}) ds \approx (-1)^K \frac{K^{K/2}}{2^{K(K-1)/2}} \prod_{k=1}^{K} g_k(\mathbf{x}) \log g_k(\mathbf{x}) \tag{27}$$

Полученное выражение (27) заканчивает доказательство теоремы.

Из теоремы 1 следует, что плотность введенная для меток учителя является плотностью распределения, следовательно можно воспользоваться выражением (22). Используя предположения 1)–4) и подставляя в (22) получаем следующую оптимизационную задачу:

$$
\begin{aligned}
\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathcal{G}} &\sum_{i\not\in\mathcal{I}}\sum_{k=1}^{K} y_i^k \log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=1} \\
&+ (1-\lambda)\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} y_i^k \log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=1} + \lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} s_{i,k}\log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=T_0} \\
&+ \lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K}\left(\log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=T_0} + \log\log\frac{1}{g_k\bigl(\mathbf{x}_i\bigr)}\bigr|_{T=T_0}\right),
\end{aligned}
\tag{28}
$$

где выражение $\cdot\bigr|_{T=t}$ обозначает, что в предыдущую функцию softmax требуется подставить значение температуры T равное некоторому значению t.

Проанализировав выражение (28) получаем, что первые три слагаемые совпадают со слагаемыми в выражении (10) при $\mathcal{I} = \{1, \cdots, m\}$, и $\lambda = \frac{1}{2}$, а третье слагаемое является некоторым регуляризатором, который получен из вида распределения.

Анализируя первые три слагаемых в выражении (28) получаем, что при $T_0 = 1$ получаем сумму кросс энтропий между двумя распределениями для каждого объекта:

1) первое распределение это выпуклая комбинация с весом $1-\lambda$ и λ: распределения задаваемое метками объектов $\mathrm{Cat}\bigl(\mathbf{y}\bigr)$ и распределения задаваемого моделью учителя $\mathrm{Cat}\bigl(\mathbf{s}\bigr)$

2) второе распределение это распределение задаваемое моделью ученика $\mathrm{Cat}\bigl(\mathbf{g}(\mathbf{x})\bigr)$.

Получаем, что модель ученика восстанавливает плотность не исходных меток, а новую плотность, которая является выпуклой комбинаций плотности исходных меток и меток учителя.

5.2. Случай регрессии

Для задачи регрессии рассматриваются следующие вероятностные предположения:

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

$$
\mathfrak{F}^*_{\mathrm{rg}} = \left\{\mathbf{f} | \mathbf{f} = \mathbf{v}^*\bigl(\mathbf{x}^*\bigr), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \to \mathbb{R}\right\}, \tag{29}
$$

где $\mathbf{v}^*$ — это дифференцируемая параметрическая функция;

2) рассматривается функция ученика $\mathbf{g} \in \mathfrak{G}_{\mathrm{rg}}$:

$$
\mathfrak{G}_{\mathrm{rg}} = \left\{\mathbf{g} | \mathbf{g} = \mathbf{z}\bigl(\mathbf{x}\bigr), \quad \mathbf{z} : \mathbb{R}^{n} \to \mathbb{R}^K\right\}, \tag{30}
$$

где $\mathbf{z}$ — это дифференцируемая параметрическая функция;

3) истинные метки имеют нормальное распределение

$$
p\bigl(y|\mathbf{x}, \mathbf{g}\bigr) = \mathcal{N}\bigl(y|\mathbf{g}(\mathbf{x}), \sigma\bigr); \tag{31}
$$

4) метки учителя распределены

$$
p\bigl(s|\mathbf{x}, \mathbf{g}\bigr) = \mathcal{N}\bigl(s|\mathbf{g}(\mathbf{x}), \sigma_s\bigr); \tag{32}
$$

Используя предположения 1)–4) и подставляя в (22) получаем следующую оптимизационную задачу:

$$\begin{aligned}\hat{g} = \arg\min_{g\in\mathcal{G}} &\sum_{i\notin\mathcal{I}} \sigma^2 \left(y_i - \mathbf{g}(\mathbf{x}_i)\right)^2 \\ &+ (1-\lambda)\sum_{i\in\mathcal{I}} \sigma^2 \left(y_i - \mathbf{g}(\mathbf{x}_i)\right)^2 + \lambda\sum_{i\in\mathcal{I}} \sigma_s^2 \left(s_i - \mathbf{g}(\mathbf{x}_i)\right)^2.\end{aligned} \tag{33}$$

Выражение (33) записано с точностью до аддитивной константы относительно $\mathbf{g}$.

Теорема 2. Пусть множество $\mathcal{G}$ описывает класс линейных функций вида $\mathbf{g}(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x}$. Тогда решение оптимизационной задачи (33) *эквивалентно решению следующей задачи линейной регрессии:*

$$\mathbf{y}'' = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{34}$$

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

$$\begin{aligned}\sigma'_i &= \begin{cases}\sigma^2, \text{ если } i\notin\mathcal{I}\\ (1-\lambda)\,\sigma^2 + \lambda\sigma_s^2, \text{ иначе}\end{cases},\\ \mathbf{y}'' &= \boldsymbol{\Sigma}\mathbf{y}',\\ y'_i &= \begin{cases}\sigma^2 y_i, \text{ если } i\notin\mathcal{I}\\ (1-\lambda)\,\sigma^2 y_i + \lambda\sigma_s^2 s_i, \text{ иначе}\end{cases}.\end{aligned} \tag{35}$$

Доказательство. Обозначеним $\mathbf{a}_\mathcal{J} = [a_i | i\in\mathcal{J}]^\mathsf{T}$, где $\mathbf{a}$ произвольный вектор, а $\mathcal{J}$ произвольное не пустое 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 $\mathbf{g}(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x}$ выражение (33) принимает вид:

$$\begin{aligned}\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}}\ &\sigma^2\left(\mathbf{y}_{\bar{\mathcal{I}}} - \mathbf{X}_{\bar{\mathcal{I}}}\mathbf{w}\right)^\mathsf{T}\left(\mathbf{y}_{\bar{\mathcal{I}}} - \mathbf{X}_{\bar{\mathcal{I}}}\mathbf{w}\right)\\ &+ \sigma^2(1-\lambda)\left(\mathbf{y}_\mathcal{I} - \mathbf{X}_\mathcal{I}\mathbf{w}\right)^\mathsf{T}\left(\mathbf{y}_\mathcal{I} - \mathbf{X}_\mathcal{I}\mathbf{w}\right) + \sigma_s^2\lambda\left(\mathbf{s}_\mathcal{I} - \mathbf{X}_\mathcal{I}\mathbf{w}\right)^\mathsf{T}\left(\mathbf{s}_\mathcal{I} - \mathbf{X}_\mathcal{I}\mathbf{w}\right).\end{aligned} \tag{36}$$

Раскроем скобки и сгруппируем:

$$\begin{aligned}\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}}\ &\sigma^2\left(\mathbf{w}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}}\mathbf{w} - 2\mathbf{y}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}}\mathbf{w}\right)\\ &+ (1-\lambda)\,\sigma^2\left(\mathbf{w}^\mathsf{T}\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\mathbf{w} - 2\mathbf{y}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\mathbf{w}\right) + \lambda\sigma_s^2\left(\mathbf{w}^\mathsf{T}\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\mathbf{w} - 2\mathbf{s}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\mathbf{w}\right)\end{aligned} \tag{37}$$

Продифференцируем выражение, приравняем к нулю и сгруппируем элементы:

$$\begin{aligned}\left(\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}} + (1-\lambda)\,\sigma^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I} + \lambda\sigma_s^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\right)\mathbf{w} =& 2\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{y}_{\bar{\mathcal{I}}}\\ &+ 2\,(1-\lambda)\,\sigma^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{y}_\mathcal{I} + 2\lambda\sigma_s^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{s}_\mathcal{I}.\end{aligned} \tag{38}$$

Воспользуемся следующими равенствами:

$$\begin{aligned}\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}} + (1-\lambda)\,\sigma^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I} + \lambda\sigma_s^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I} &= \mathbf{X}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{X},\\ 2\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{y}_{\bar{\mathcal{I}}} + 2\,(1-\lambda)\,\sigma^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{y}_\mathcal{I} + 2\lambda\sigma_s^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{s}_\mathcal{I} &= 2\mathbf{X}\mathbf{y}',\end{aligned} \tag{39}$$

где $\mathbf{\Sigma}$ и $\mathbf{y}'$ из условия задачи (35).

Подставляя (39) в (38) получаем:

$$\mathbf{w} = 2\left(\mathbf{X}^{\mathsf{T}}\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{\Sigma}^{-1}\mathbf{y}'', \tag{40}$$

что соответсвует решению задачи (34).

Теорема 2 показывает, что обучения с учителем для задачи регрессии можно свести к классической задачи оптимизации для задачи линейной регрессии.

6. Вычислительный эксперимент

Проводится вычислительный эксперимент для анализа качества моделей, которые получены путем дистилляции модели учителя в модель ученика. Как показано в теореме 2 задачу регрессии с учителем можно свести к задачи регрессии без учителя, поэтому в эксперименте более подробно рассматривается случай классификации. Во всех частях вычислительного эксперимента для поиска оптимальных параметров нейросетей использовался градиентный метод оптимизации Адам [11].

6.1. Выборка FashionMNIST

В данной части проводится эксперимент для задачи классификации для выборки FashionMNIST [10]. В качестве модели учителя $\mathbf{f}$ рассматривается модель нейросети 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.

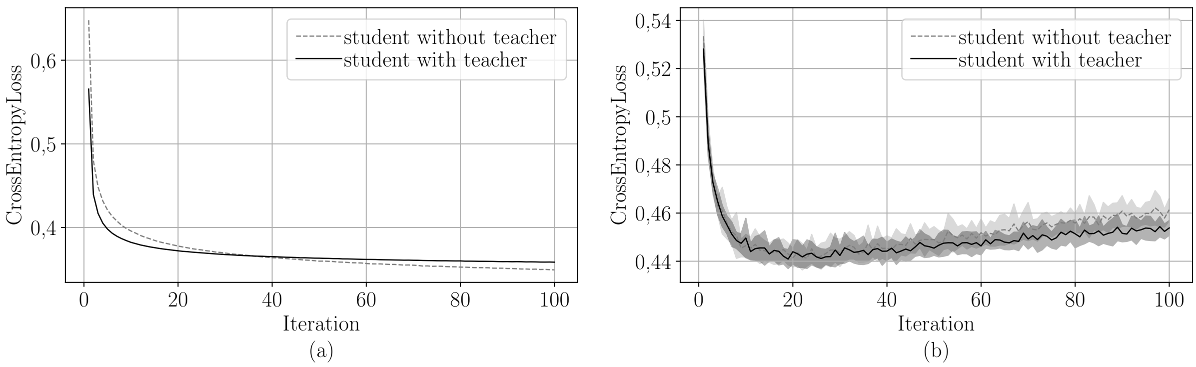

Рис. 2. Зависимость кросс–этропии между истинными метками и предсказанными учеников вероятностями классов: a) на обучающей выборке; b) на тестовой выборке.

На рис. 2 показан график зависимости кросс–энтропии между истинными метками объектов и вероятностями, которые предсказывает модель ученика. На графике сравнивается моделя, которая обучалась без учителя (в задаче оптимизации (28) присутствует только первое слагаемое) с моделью, которая была получена путем дистилляции модели нейросети в линейную модель. На графике видно, что обе модели начинают переобучатся после 30-й итерации, но модель, которая получена путем дистилляции переобувается не так быстро, что следует из того, что ошибка на тестовой выборке растет медленней, а на обучающей выборке падает также медленней.

6.2. Синтетический эксперимент

Проанализируем модель на синтетической выборке. Выборка построенная следующим образом:

$$\begin{aligned} \mathbf{W} &= \left[\mathcal{N}\big(w_{jk}|0,1\big)\right]_{n\times K}, \qquad & \mathbf{X} &= \left[\mathcal{N}\big(x_{ij}|0,1\big)\right]_{m\times n}, \\ \mathbf{S} &= \mathrm{softmax}\left(\mathbf{XW}\right), \qquad & \mathbf{y} &= \left[\mathrm{Cat}\big(y_i|\mathbf{s}_i\big)\right], \end{aligned} \tag{41}$$

где функция softmax берется построчно. Строки матрицы $\mathbf{S}$ будем рассматривать как предсказание учителя, то есть учитель знает истинные вероятности каждого класса. На рис. 3 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 каждого объекта обучающей выборки. Видно, что все классы являются равновероятными.

Построим в качестве ученика простую линейную модель, которая минимизирует крос–энтропийную (первое слагаемое в формуле (28)). Представление данной модели в виде графовой модели показано на рис. 3.

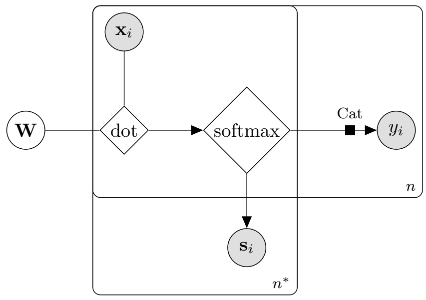

Рис. 3. Вероятностная модель используемая в синтетическом эксперименте.

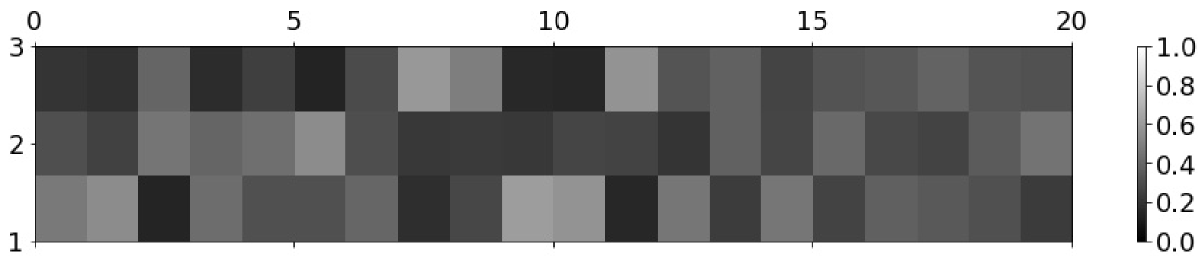

Рис. 4. Истинное распределение объектов по классам.

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

Рассмотрим модель, которая учитывает информацию о истинных распределениях на классах для каждого объекта. Для этого будем минимизировать первые три слагаемых в формуле (28), при $T_0 = 1$ и $\lambda = 0{,}75$. В качестве меток учителя $s_{i,k}$ использовались истинные вероятности для каждого класса для данного объекта. На

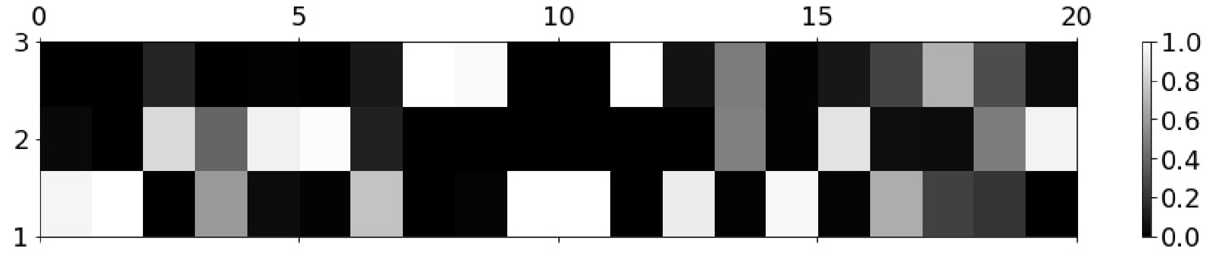

Рис. 5. Распределение предсказанное моделью без использования информации об истинном распределение на классах.

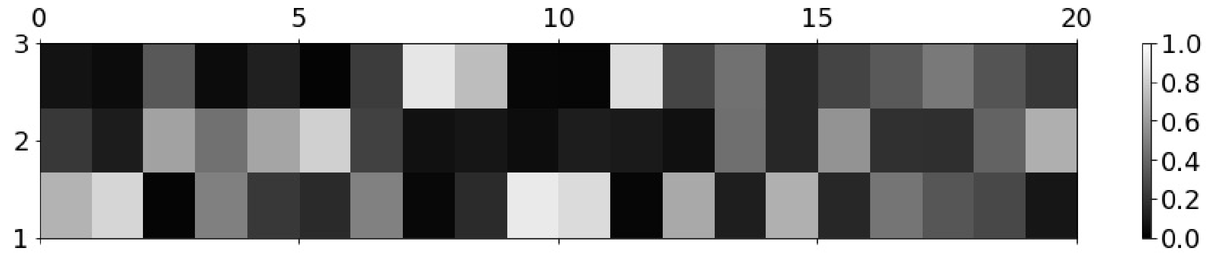

Рис. 6. Распределение предсказанное моделью с использования информации об истинном распределение на классах.

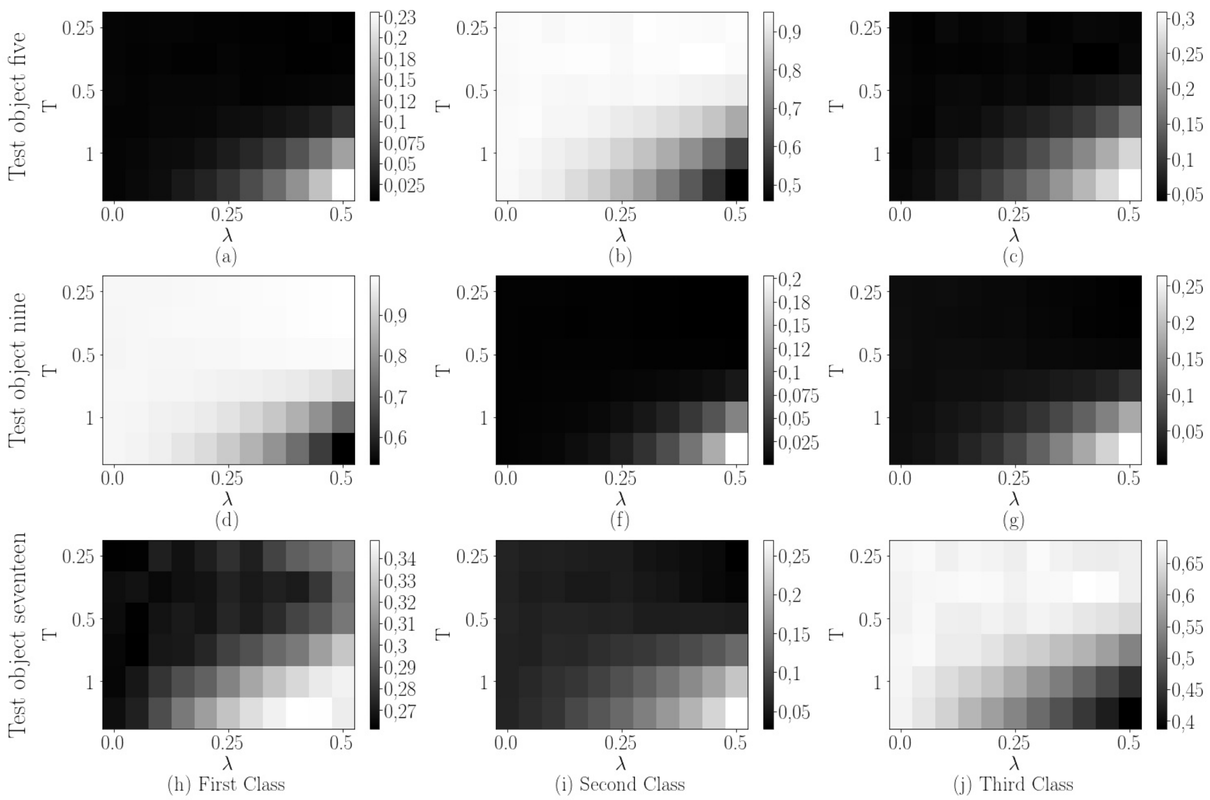

Рис. 7. Вероятности классов для разных объектов.

рис 6 показано распределение, которое дала модель в данном случае, видно, что распределения являются сглаженными и концентрации всей вероятности в одном классе

Таблица 1. Сводная таблица результатов вычислительного эксперимента.

Dataset	Model	CrossEntropyLoss	Accuracy	StudentSize
FashionMnist	without teacher	$0{,}461 \pm 0{,}005$	$0{,}841 \pm 0{,}002$	7850
	with teacher	$0{,}453 \pm 0{,}003$	$0{,}842 \pm 0{,}002$	7850
Synthetic	without teacher	$0{,}225 \pm 0{,}002$	$0{,}831 \pm 0{,}002$	33
	with teacher	$0{,}452 \pm 0{,}001$	$0{,}828 \pm 0{,}001$	33
Twitter	without teacher	$0{,}501 \pm 0{,}006$	$0{,}747 \pm 0{,}005$	1538
	with teacher	$0{,}489 \pm 0{,}003$	$0{,}764 \pm 0{,}004$	1538

не наблюдается.

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

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

На рис. 7 показана зависимость вероятности верного класса от температуры T и параметра доверия λ для одного из объекта из тестовой выборке. Видно, что при увеличении темпертуры распределение на классас становится более равномерным.

6.3. Выборка Twitter Sentiment Analysis

В данной части проводится эксперимент на выборке Twitter Sentiment Analysis. Данная выборка содержит короткие сообщения, для которых нужно предсказать эмоциональный окрас: содержит твит позитивный окрас или негативный. Выборка разделена на 1,18 миллиона твитов для обучения и 0,35 миллиона твитов для тестирования. В твитах была выполнена следующая 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 табл. 1. В качестве модели учителя использовалась модель Bi-LSTM с 170 тысячами параметров для обучения. В качестве эмбедингов обучалась матрица из 30 миллионов параметров в единой процедуре с моделью BI-LSTM. Обученная модель предсказывает с точностью 0,835. В качестве модели ученика рассматривается модель с 1538 параметрами, но в качестве эмбедингов рассматривается переобученная модель BERT.

7. Заключение

В данной работе проанализирована задача обучения модели ученика с помощью модели учителя. Исследован метод дистилляции и привилигированного обучения. Предложено вероятностное обоснования дистилляции. Введены вероятностные предположения описывающие дистилляцию моделей. В рамках данных вероятностных предположений проанализированы модели для задачи классификации и регрессии. Результат анализа сформулирован в виде теоремы 1 и теоремы 2.

Теорема 2 показала, что обучения линейной регрессии с учителем эквивалентно замене обучающей выборке 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 переписана в виде классической задачи классификации. Для использования данной информации требуется использовать распределение, которое представлено в теореме 1.

В вычислительном эксперименте сравнивается модель ученика, которая обучена без использования учителя и с использованием модели учителя. В таблице 1 показаны результаты вычислительного эксперимента для разных выборок. Из таблицы видно, что точность аппроксимации выборки учеником улучшается при использовании модели учителя. Задачи регрессии не приведена в вычислительном эксперименте, так как в теореме 2 была показана эквивалентность классическому решению задачи линейной регрессии. Для задачи классификации проведен вычислительный эксперимент.

В дальнейшем предполагается обобщить метод максимального правдоподобия для дистилляции моделей используя Байесовский подход выбора моделей машинного обучения. Также в рамках байесовсокго подхода планируется улучшить методы для получения улучшения качества не только для задачи классификации, но и для задачи регрессии.

СПИСОК ЛИТЕРАТУРЫ

1. *Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel* Backpropagation Applied to Handwritten Zip Code Recognition // Neural Computation. 1989. Vol.1 No 4. pp. 541–551.

2. *Sepp Hochreiter; Jürgen Schmidhuber* Long short-term memory // Neural Computation. 1997. Vol. 9, No 8. pp. 1735–1780.

3. *Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun* Deep Residual Learning for Image Recognition // CoRR. 2015

4. *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin* Attention Is All You Need // CoRR. 2017

5. *Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. 2018

6. *Vladimir Vapnik, Rauf Izmailov* Learning Using Privileged Information: Similarity Control and Knowledge Transfer // Journal of Machine Learning Research. 2015. No 16. pp. 2023–2049.

7. *David Lopez-Paz, Leon Bottou, Bernhard Scholkopf, Vladimir Vapnik* UNIFYING DISTILLATION AND PRIVILEGED INFORMATION // Published as a conference paper at ICLR. 2016.

8. *Geoffrey Hinton, Oriol Vinyals, jeff Dean* Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. 2015.

9. *LeCun Y., Cortes C., Burges C.* The MNIST dataset of handwritten digits, 1998. http://yann.lecun.com/exdb/mnist/index.html

10. *Han Xiao and Kashif Rasul and Roland Vollgraf* Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms // arXiv, 2017.

11. *Diederik P. Kingma and Jimmy Ba* Adam: A Method for Stochastic Optimization // arXiv, 2014.

12. *О. Ю. Бахтеев, В. В. Стрижов* Выбор моделей глубокого обучения субоптимальной сложности // Автоматика и телемеханика, 2018.

13. *Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter* SemEval-2013 task 2: Sentiment analysis in twitter // In Proceedings of the International Workshop on Semantic Evaluation, SemEval '13. 2013.

14. *Madala, H. R. and Ivakhnenko, A. G.* Inductive Learning Algorithms for Complex Systems Modeling // CRC Press Inc., Boca Raton, 1994.

Грабовой А.В., *Московский физико-технический институт, студент, Москва*, `grabovoy.av@phystech.edu`

Стрижов В.В., *Московский физико-технический институт, профессор, Москва*, `strijov@phystech.edu`