Вероятностное обоснование дистилляции моделей машинного обучения *

А. В. Грабовой[1], **В. В. Стрижов**[2]

Аннотация: Данная работа посвящена методам для понижения сложности модели при помощи дистилляции. Предлагается вероятностное обоснование методов для понижения сложности моделей машинного обучения путем дистилляции и привилегированного обучения. В работе показаны общие выводы для произвольной параметрической функции с заданой сигнатурой, а также 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 выборка FashionMNIST и Twitter Sentiment Analysis.

Ключевые слова: дистилляция моделей, привилегированное обучения, выбор модели, байесовские методы.

DOI: 00.00000/00000000000000

1 Введение

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

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,

*Работа выполнена при поддержке РФФИ и правительства РФ.

[1]Московский физико-технический институт, grabovoy.av@phystech.edu

[2]Московский физико-технический институт, strijov@phystech.edu

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.

В работе [8] рассматривается метод дистилляции моделей машиного 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 сложностная нейросеть была дистиллирована в меньшую нейросеть. 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.

В работе [6] введено понятия *привилегированной информации* — информации которая доступная только в момент обучения. 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].

В данной работе предлагается рассмотреть общий подход дистилляции в рамках вероятностного подхода. Проводится обобщение на случай, когда привилегированная информация доступна не для всех объектов из обучающей выборки. Предлагается анализ и обобщение функции ошибки [8, 7] в рамках вероятностного подхода. Также предлагается рассмотреть частные случаи для задач классификации и регрессии.

В рамках вычислительного эксперимента анализируются модели, которые используют модель учителя при обучения и модели, которые не используют модель учителя при обучения. Для анализа используются реальные выборки для задачи классификации изображений FashionMNIST [10] и для задачи классификации текстов Twitter Sentiment Analysis [13]. В эксперименте использовалась выборка FashionMNIST, таккак выборка MNIST имеет хорошее качество аппроксимации даже для линейного классификатора. В рамках вычислительного эксперимента использовались модели разной сложности: линейные модели, полносвязная нейронная сеть, сверточная нейронная сеть [1], модель Bi-LSTM [2] и модель BERT [5].

2 Постановка задачи обучения с учителем

Пусть задано множество объектов $\mathbf{\Omega}$ и множество целевых переменных $\mathbb{Y}$:

$$\mathbf{\Omega}, \quad |\mathbf{\Omega}| = m,$$

где m — число объектов, множество $\mathbb{Y} = \{1, \cdots, K\}$ для задачи классификации, где K число классов, множество $\mathbb{Y} = \mathbb{R}$ для задачи регрессии. Для множества $\mathbf{\Omega}$ задано отображение в некоторое признаковое пространство $\mathbb{R}^n$:

$$\varphi : \mathbf{\Omega} \to \mathbb{R}^n,$$

где n размерность признакового пространства. Обозначим $\varphi(\mathbf{\Omega}) = \mathbf{X}$. Пусть для объектов $\mathbf{\Omega}^* \subset \mathbf{\Omega}$ задана привилегированная информация:

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

где $m^* \leq m$ — число объектов с привилегированной информацией, n^* — число признаков в пространстве привилегированной информации. Обозначим $\varphi^*(\mathbf{\Omega}^*) = \mathbf{X}^*$.

Множество индексов объектов, для которых известна привилегированная информация, обозначим $\mathcal{I}$:

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

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 $\mathbf{f}\bigl(\mathbf{x}^*\bigr)$:

$$\mathbf{f} : \mathbb{R}^{n^*} \to \mathbb{Y}^*, \tag{1}$$

где $\mathbb{Y}^* = \mathbb{Y}$ для задачи регрессии и $\mathbb{Y}^*$ является единичным симплексом $\mathcal{S}_K$ в пространстве размерности K для задачи классификации. Обозначим ответы модели $\mathbf{f}\bigl(\mathbf{x}_i^*\bigr) = \mathbf{s}_i$. Получим ответы $\mathbf{S}$ модели учителя $\mathbf{f}$.

Требуется построить модель $\mathbf{g}\bigl(\mathbf{x}\bigr)$ над множеством исходных признаков:

$$\mathbf{g} : \mathbb{R}^n \to \mathbb{Y}^*.$$

Пусть $\mathbf{g}$ выбирается из некоторого множества функций:

$$\mathcal{G} = \{\mathbf{g} | \mathbf{g} : \mathbb{R}^n \to \mathbb{Y}^*\}, \tag{2}$$

например для задачи классификации множество $\mathcal{G}$ может быть параметрическим семейством функций линейных моделей $\mathcal{G} = \bigl\{\mathbf{g}\bigl(\mathbf{W}, \mathbf{x}\bigr) | \mathbf{g}\bigl(\mathbf{W}, \mathbf{x}\bigr) = \text{softmax}\bigl(\mathbf{W}\mathbf{x}\bigr)\bigr\}$.

3 Постановка задачи: Хинтон & Вапник

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

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

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

В данной постановке рассматривается параметрическое семейство функций:

$$\mathcal{G}_{\text{cl}} = \bigl\{\mathbf{g} | \mathbf{g} = \text{softmax}\bigl(\mathbf{z}\bigl(\mathbf{x}\bigr)/T\bigr), \quad \mathbf{z} : \mathbb{R}^n \to \mathbb{R}^K\bigr\},$$

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

$$\mathcal{F}_{\text{cl}} = \bigl\{\mathbf{f} | \mathbf{f} = \text{softmax}\bigl(\mathbf{v}\bigl(\mathbf{x}\bigr)/T\bigr), \quad \mathbf{v} : \mathbb{R}^n \to \mathbb{R}^K\bigr\},$$

где $\mathbf{v}$ — это дифференцируемая параметрическая функция, T — параметр температуры.

Параметр температуры T имеет следующие свойства:

1) при $T \to 0$ получаем one hot вектора;

1) при $T \to \infty$ получаем равновероятные классы.

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

$$\begin{aligned}\mathcal{L}(\mathbf{g}) = &-\underbrace{\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log \mathbf{g}(\mathbf{x}_i)\bigr|_{T=1}}_{\text{исходная функция потерь}} \\ &-\underbrace{\sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{f}(\mathbf{x}_i)\bigr|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)\bigr|_{T=T_0}}_{\text{слагаемое дистилляция}},\end{aligned} \tag{3}$$

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

Получаем оптимизационную задачу:

$$\begin{aligned}\hat{\mathbf{g}} = \arg\min_{\mathbf{g}\in\mathcal{G}_{\text{cl}}} &-\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log \mathbf{g}(\mathbf{x}_i)\bigr|_{T=1} \\ &-\sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{f}(\mathbf{x}_i)\bigr|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)\bigr|_{T=T_0}.\end{aligned} \tag{4}$$

В работе [7] метод [8] имеет обобщение. Решение задачи оптимизации (4) зависит от модели учителя $\mathbf{f}$, только через вектор ответов учителя. Следовательно признаковые пространства учителя и ученика могут различатся. В этом случае получаем следующую постановку задачи:

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

где $\mathbf{x}_i$ это информация доступна на этапах обучения и контроля, а $\mathbf{x}_i^*$ это информация доступна только на этапе обучения.

В данном случае в качестве функции учителя выбирается функция $\mathbf{f}$ не из множества $\mathcal{F}_{cl}$, а из множества $\mathcal{F}_{cl}^*$:

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

где $\mathbf{v}^*$ — это дифференцируемая параметрическая функция, T — параметр температуры.

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 $\mathbf{f} \in \mathcal{F}^*_{\text{cl}}$;

2) выбираем оптимальную модель ученика $\mathbf{g} \in \mathcal{G}_{\text{cl}}$ используя дистилляцию [8].

Модель ученика — это функция, которая минимизирует (3). Модель учителя — это функция, которая минимизирует Cross Entropy Loss:

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

4 Постановка задачи: вероятностный подход

4.1 Метод максимального правдоподобия

Для поиска $\hat{\mathbf{g}}$ воспользуемся методом максимального правдоподобия:

$$p(\mathbf{Y}|\mathbf{X},\mathbf{g}) = \prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{x}_i,\mathbf{g}). \tag{5}$$

В качестве $\hat{\mathbf{g}}$ выбирается функция, которая максимизирует правдоподобие модели (5):

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathcal{G}} \prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{x}_i,\mathbf{g}), \tag{6}$$

где множество $\mathcal{G}$ задается (2).

4.2 Подход дистилляции модели учителя в модель ученика

Рассмотрим вероятностную постановку, в которой должны быть выполнены ограничения:

1) 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;

2) если $|\mathbf{\Omega}^*| = 0$ то решение должно соответствовать решению (6);

3) рассмотрим параметр $\lambda \in [0, 1]$ как уровень доверия к ответам модели $\mathbf{f}$, которая задана в (1).

Рассмотрим совместное правдоподобие истинных меток и меток учителя:

$$p(\mathbf{Y},\mathbf{S}|\mathbf{X},\mathbf{g},\mathcal{I},\lambda) = \prod_{i\not\in\mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i,\mathbf{g}) \prod_{i\in\mathcal{I}} p(\mathbf{y}_i,\mathbf{s}_i|\mathbf{x}_i,\mathbf{g},\lambda). \tag{7}$$

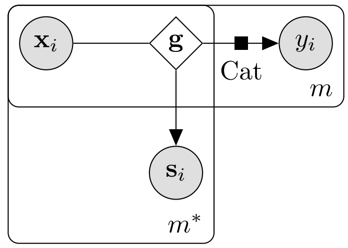

Рис. 1: Общий вид модели в формате plate notation.

Рассмотрим $p\bigl(\mathbf{y}_i, \mathbf{s}_i|\mathbf{X}, \mathbf{g}, \lambda\bigr)$ следующего вида:

$$p\bigl(\mathbf{y}_i, \mathbf{s}_i|\mathbf{X}, \mathbf{g}, \lambda\bigr) = p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} \bigl(\mathbf{s}_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}\bigr) \tag{8}$$

Подставляя выражения (8) в (7) получаем.

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

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:

$$p\bigl(\mathbf{Y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}, \lambda\bigr) = \prod_{i\not\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr). \tag{9}$$

Используя (9) получаем следующую оптимизационную задачу для поиска $\hat{\mathbf{g}}$

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathcal{G}} \prod_{i\not\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i\in\mathcal{I}} p\bigl(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr). \tag{10}$$

Для удобства, будем минимизировать логарифм, тогда из (10) получаем:

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathcal{G}} \sum_{i\not\in\mathcal{I}} \log p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) + (1-\lambda) \sum_{i\in\mathcal{I}} \log p\bigl(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}\bigr) + \lambda \sum_{i\in\mathcal{I}} \log p\bigl(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr), \tag{11}$$

где параметр λ введен для взвешивания ошибок на истинных ответах и ошибок относительно ответов учителя.

На рис. 1 показан общий вид вероятностной модели в графовой нотации. Для каждой реализации функции $\mathbf{g}$ данный вид изменяется. На рис 3 показана более подробная реализация в случае, когда модель $\mathbf{g}$ это линейная модель.

5 Частные случаи обучения с учителем

5.1 Случай классификации

Для задачи многоклассовой классификации рассматриваются следующие вероятностные предположения:

1) рассматривается функция учителя $\mathbf{f} \in \mathcal{F}^*_{\mathrm{cl}}$;

2) рассматривается функция ученика следующего вида $\mathbf{g} \in \mathcal{G}_{\mathrm{cl}}$;

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

4) для меток учителя рассматривается плотность распределения

$$p\bigl(\mathbf{s}|\mathbf{x}, \mathbf{g}\bigr) = C\prod_{k=1}^{K} g_k\bigl(\mathbf{x}\bigr)^{s^k}. \tag{12}$$

Теорема 1. *Пусть для всех* k *выполняется* $1 > 1 - \varepsilon > g_k\bigl(\mathbf{x}\bigr) > \varepsilon > 0$, *тогда при*

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

функция $p\bigl(\mathbf{s}|\mathbf{x}, \mathbf{g}\bigr)$ *определенная в* (12) *является плотностью распределения.*

Доказательство.

1) Покажем, что для произвольного $\mathbf{s} \in \mathcal{S}_K$ выполняется $p\bigl(\mathbf{s}|\mathbf{x}, \mathbf{g}\bigr) \geq 0$. Заметим, что для всех k выполняется, что $\log g_k\bigl(\mathbf{x}\bigr) < 0$, тогда

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

тогда с учетом того, что $g_k\bigl(\mathbf{x}\bigr) > 0$ и $C > 0$ получаем, что $p\bigl(\mathbf{s}|\mathbf{x}, \mathbf{g}\bigr) \geq 0$.

2) Покажем, что интеграл по всему пространству объектов $\mathcal{S}_K$ является конечным:

$$
\begin{aligned}
\int_{\mathcal{S}_K} p\bigl(\mathbf{s}|\mathbf{x},\mathbf{g}\bigr)ds &= \int_{\mathcal{S}_K}\prod_{k=1}^{K} g_k\bigl(\mathbf{x}\bigr)^{s^k}ds = \prod_{k=1}^{K}\int_{\mathcal{S}_K} g_k\bigl(\mathbf{x}\bigr)^{s^k}ds \\
&= \prod_{k=1}^{K}\int_0^1 \frac{r^{K-1}\sqrt{K}}{(K-1)!\sqrt{2^{K-1}}} g_k\bigl(\mathbf{x}\bigr)^r dr = \prod_{k=1}^{K}\underbrace{\frac{\sqrt{K}}{(K-1)!\sqrt{2^{K-1}}}}_{D}\int_0^1 r^{K-1} g_k\bigl(\mathbf{x}\bigr)^r dr \\
&= D^K\prod_{k=1}^{K}\int_0^1 r^{K-1}\exp\bigl(r\log g_k\bigl(\mathbf{x}\bigr)\bigr)dr \\
&= (-D)^K\prod_{k=1}^{K}\log g_k\bigl(\mathbf{x}\bigr)\left(\Gamma\bigl(K\bigr)-\Gamma\bigl(K,-\log g_k\bigl(\mathbf{x}\bigr)\bigr)\right) \\
&= (-D)^K\,(K-1)!^K\prod_{k=1}^{K}\log g_k\bigl(\mathbf{x}\bigr)\left(1-g_k\bigl(\mathbf{x}\bigr)\exp_{K-1}\bigl(-\log g_k\bigl(\mathbf{x}\bigr)\bigr)+g_k\bigl(\mathbf{x}\bigr)\right) \\
&= \frac{\left(-\sqrt{K}\right)^K}{2^{K(K-1)/2}}\prod_{k=1}^{K}\log g_k\bigl(\mathbf{x}\bigr)\left(1-g_k\bigl(\mathbf{x}\bigr)\exp_{K-1}\bigl(-\log g_k\bigl(\mathbf{x}\bigr)\bigr)+g_k\bigl(\mathbf{x}\bigr)\right)<\infty,
\end{aligned}
\tag{13}
$$

где $\Gamma\bigl(K\bigr)$ является гамма функцией, $\Gamma\bigl(K,-\log g_k\bigl(\mathbf{x}\bigr)\bigr)$ является неполной гамма функцией, $\exp_n\bigl(x\bigr)$ является суммой Тейлора из первых n слагаемых. В рамках приближенных расчетов будем считать, что $\exp_n\bigl(x\bigr)\approx\exp\bigl(x\bigr)$, тогда с учетом (13) получаем:

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

□

Используя предположения 1)–4) и подставляя в (11) получаем следующую оптимизационную задачу:

$$
\begin{aligned}
\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathcal{G}} &\sum_{i\not\in\mathcal{I}}\sum_{k=1}^{K} y_i^k\log g_k\bigl(\mathbf{x}_i\bigr)\big|_{T=1} \\
&+(1-\lambda)\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} y_i^k\log g_k\bigl(\mathbf{x}_i\bigr)\big|_{T=1}+\lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} s_{i,k}\log g_k\bigl(\mathbf{x}_i\bigr)\big|_{T=T_0} \\
&+\lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K}\left(\log g_k\bigl(\mathbf{x}_i\bigr)\big|_{T=T_0}+\log\log\frac{1}{g_k\bigl(\mathbf{x}_i\bigr)}\big|_{T=T_0}\right),
\end{aligned}
\tag{14}
$$

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

Проанализировав выражение (14) получаем, что первые три слагаемые совпадают со слагаемыми в выражении (3) при $\mathcal{I} = \{1, \cdots, m\}$, и $\lambda = \frac{1}{2}$, а третье слагаемое является некоторым регуляризатором, который получен из вида распределения.

Анализируя первые 3 слагаемых в выражении (14) 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 $\text{Cat}\bigl(\mathbf{y}\bigr)$ и распределения задаваемого моделью учителя $\text{Cat}\bigl(\mathbf{s}\bigr)$
2) второе распределение это распределение задаваемое моделью ученика $\text{Cat}\bigl(\mathbf{g}\bigl(\mathbf{x}\bigr)\bigr)$.

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 $\mathbf{f} \in \mathcal{F}^*_{\text{rg}}$:
$$\mathcal{F}^*_{\text{rg}} = \left\{\mathbf{f} | \mathbf{f} = \mathbf{v}^*\bigl(\mathbf{x}^*\bigr), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \to \mathbb{R}\right\},$$
где $\mathbf{v}^*$ — это дифференцируемая параметрическая функция;
2) рассматривается функция ученика $\mathbf{g} \in \mathcal{G}_{\text{rg}}$:
$$\mathcal{G}_{\text{rg}} = \left\{\mathbf{g} | \mathbf{g} = \mathbf{z}\bigl(\mathbf{x}\bigr), \quad \mathbf{z} : \mathbb{R}^{n} \to \mathbb{R}^K\right\},$$
где $\mathbf{z}$ — это дифференцируемая параметрическая функция;
3) истинные метки имеют нормальное распределение распределение
$$p\bigl(y | \mathbf{x}, \mathbf{g}\bigr) = \mathcal{N}\bigl(y | \mathbf{g}\bigl(\mathbf{x}\bigr), \sigma\bigr);$$
4) метки учителя распределены
$$p\bigl(s | \mathbf{x}, \mathbf{g}\bigr) = \mathcal{N}\bigl(s | \mathbf{g}\bigl(\mathbf{x}\bigr), \sigma_s\bigr);$$

Используя предположения 1)–4) и подставляя в (11) получаем следующую оптимизационную задачу:
$$\begin{aligned} \hat{g} = \arg\min_{g \in \mathcal{G}} & \sum_{i \not\in \mathcal{I}} \sigma^2 \left(y_i - \mathbf{g}\bigl(\mathbf{x}_i\bigr)\right)^2 \\ & + \left(1 - \lambda\right) \sum_{i \in \mathcal{I}} \sigma^2 \left(y_i - \mathbf{g}\bigl(\mathbf{x}_i\bigr)\right)^2 + \lambda \sum_{i \in \mathcal{I}} \sigma_s^2 \left(s_i - \mathbf{g}\bigl(\mathbf{x}_i\bigr)\right)^2. \end{aligned} \tag{15}$$

Выражение (15) 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* (15) *эквивалентно решению следующей задачи линейной регрессии:*

$$\mathbf{y}'' = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{16}$$

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

$$\begin{aligned} \sigma'_i &= \begin{cases} \sigma^2, \text{ если } i \notin \mathcal{I} \\ (1-\lambda)\,\sigma^2 + \lambda\sigma_s^2, \text{ иначе} \end{cases}, \\ \mathbf{y}'' &= \boldsymbol{\Sigma}\mathbf{y}', \\ y'_i &= \begin{cases} \sigma^2 y_i, \text{ если } i \notin \mathcal{I} \\ (1-\lambda)\,\sigma^2 y_i + \lambda\sigma_s^2 s_i, \text{ иначе} \end{cases}. \end{aligned} \tag{17}$$

Доказательство. В доказательстве используется обозначения $\mathbf{a}_\mathcal{J} = [a_i | i \in \mathcal{J}]^\mathsf{T}$, где $\mathbf{a}$ произвольный вектор, а $\mathcal{J}$ произвольное не пустое индексное множество. К примеру подвектор вектора ответов $\mathbf{y}$ для элементов которого доступна привилегированная информация обозначается $\mathbf{y}_\mathcal{I} = [y_i | i \in \mathcal{I}]^\mathsf{T}$. Аналогичная операция рассматривается для матрицы объектов $\mathbf{X}_\mathcal{I} = [\mathbf{x}_i | i \in \mathcal{I}]^\mathsf{T}$.

В случае линейной модели, когда $\mathbf{g}(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x}$ выражение (15) принимает вид:

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

Раскроем скобки и сгруппируем:

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

Продифференцируем выражение, приравняем к нулю и сгруппируем элементы:

$$\begin{aligned} \left(\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}} + (1-\lambda)\,\sigma^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I} + \lambda\sigma_s^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I}\right)\mathbf{w} =& 2\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{y}_{\bar{\mathcal{I}}} \\ &+ 2(1-\lambda)\,\sigma^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{y}_\mathcal{I} + 2\lambda\sigma_s^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{s}_\mathcal{I}. \end{aligned} \tag{18}$$

Легко получить следующие равенства:

$$\begin{aligned} \sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{X}_{\bar{\mathcal{I}}} + (1-\lambda)\,\sigma^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I} + \lambda\sigma_s^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{X}_\mathcal{I} = \mathbf{X}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{X}, \\ 2\sigma^2\mathbf{X}_{\bar{\mathcal{I}}}^\mathsf{T}\mathbf{y}_{\bar{\mathcal{I}}} + 2(1-\lambda)\,\sigma^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{y}_\mathcal{I} + 2\lambda\sigma_s^2\mathbf{X}_\mathcal{I}^\mathsf{T}\mathbf{s}_\mathcal{I} = 2\mathbf{X}\mathbf{y}', \end{aligned} \tag{19}$$

где $\boldsymbol{\Sigma}$ и $\mathbf{y}'$ из условия задачи (17).

Подставляя (19) в (18) получаем:

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

что соответсвует решению задачи (16). □

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.

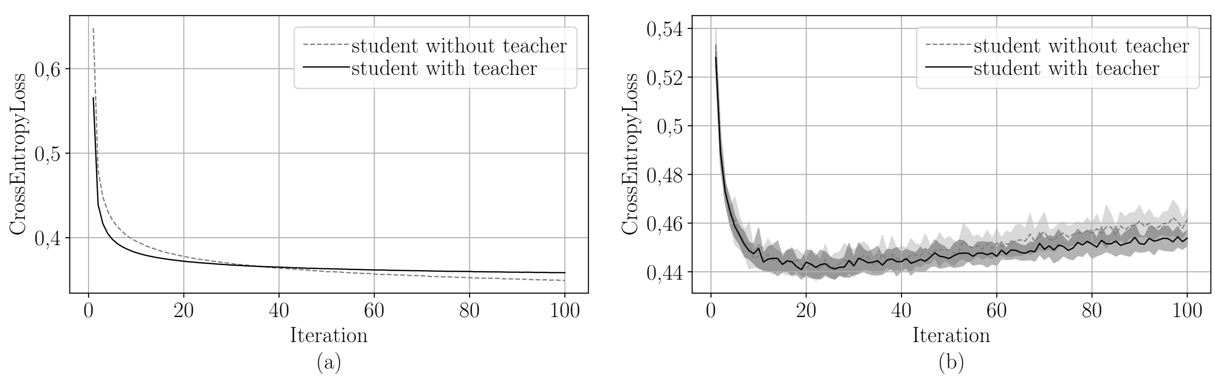

Рис. 2: Зависимость кросс–этропии между истинными метками и предсказанными учеников вероятностями классов: a) на обучающей выборке; b) на тестовой выборке.

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 (14) 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.

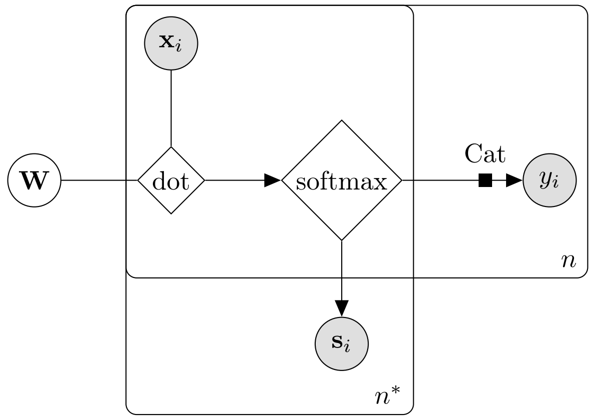

Рис. 3: Модель синтетического эксперимента в формате plate notation.

6.2 Синтетический эксперимент

Проанализируем модель на синтетической выборке. Выборка построенная следующим образом:

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

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 (14)) ошибку между one hot распределением и распределением, которое предсказывает линейная модель $\mathbf{g}$. Представим данную модель в виде plate-notation:

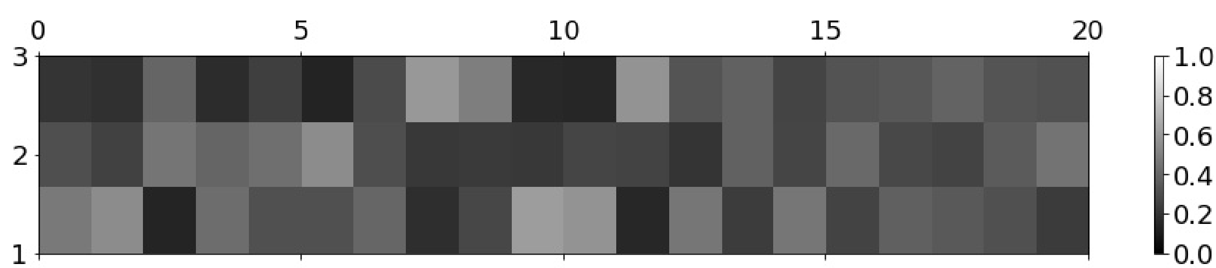

Рис. 4: Истинное распределение объектов по классам.

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.

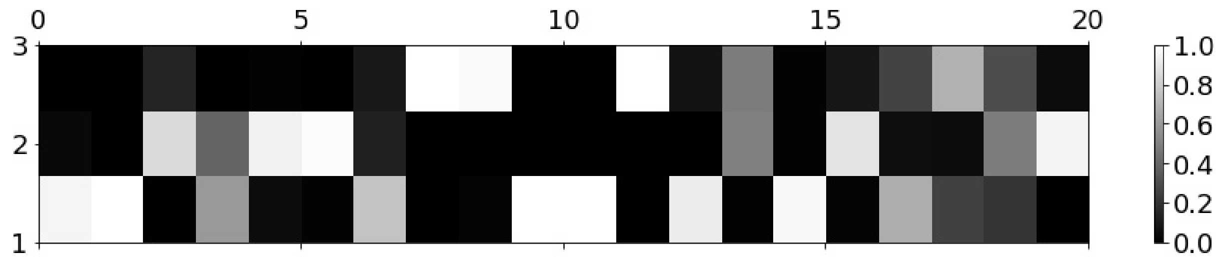

Рис. 5: Распределение предсказанное моделью без использования информации об истинном распределение на классах.

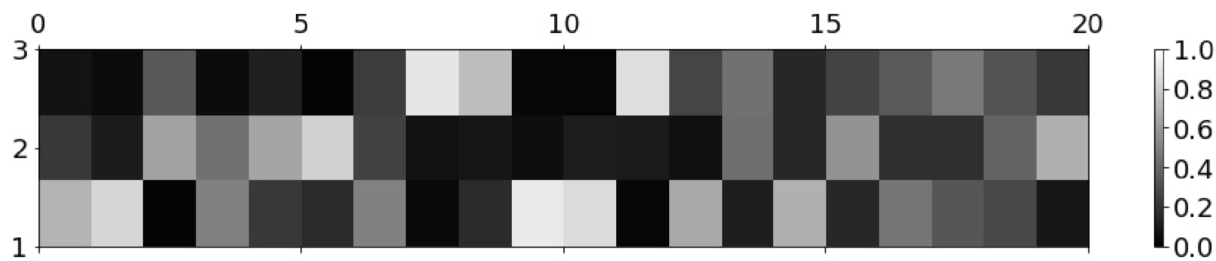

Рис. 6: Распределение предсказанное моделью с использования информации об истинном распределение на классах.

Рассмотрим модель, которая учитывает информацию о истинных распределениях на классах для каждого объекта. Для этого будем минимизировать первые три слагаемых в формуле (14), 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 6 показано распределение, которое дала модель в данном случае, видно, что распределения являются сглаженными и концентрации всей вероятности в одном классе 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 меткок.

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.

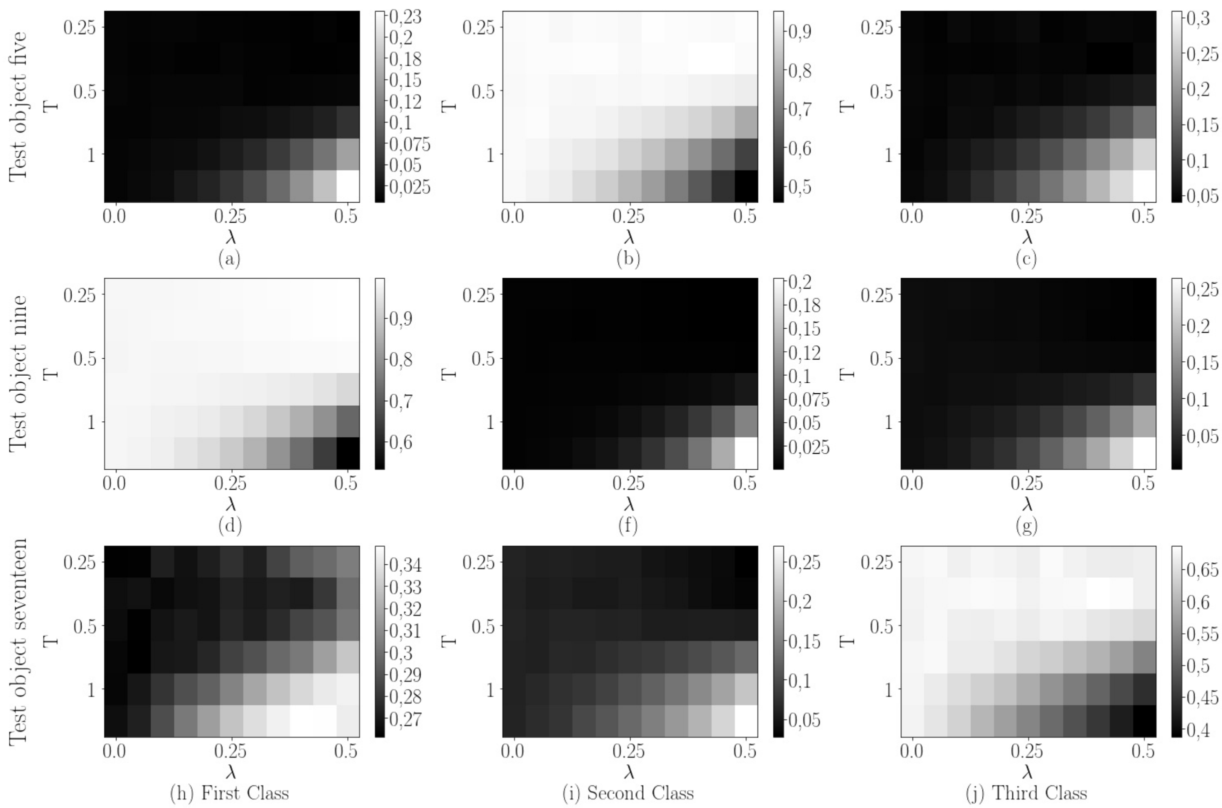

Рис. 7: Вероятности классов для разных объектов.

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”.

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. Результаты данной части эксперимента показаны в табл. 1.

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

Таблица 1: Сводная таблица результатов вычислительного эксперимента.

7 Выводы

В данной работе проанализирована задача обучения модели ученика используя модель учителя. Исследован метод дистилляции модели учителя в модель ученика. В работе предложено вероятностное обоснования дистилляции, которое было предложено в работах [8, 7]. Введены вероятностные предположения, которые описывают дистилляцию моделей. В рамках данных вероятностных предположений получен анализ некоторых моделей. Результат анализа сформулирован в виде теоремы 1 и теоремы 2.

В рамках теоремы 2 показано, 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.

Анализ задачи регрессии в вычислительном эксперименте не проводится, так как в теореме 2 была показана эквивалентность классическому решению задачи линейной регрессии. Для задачи классификации проведен вычислительный эксперимент. 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.

В дальнейшем предполагается обобщить метод описаный в пункте 4 используя Байесовский подход выбора моделей машинного обучения. Также в рамках байесовсокго подхода планируется улучшить методы для получения улучшения качества не только для задачи классификации, а 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.