Привилигированная информация и дистиляция моделей

Грабовой Андрей Валериевич

Московский физико-технический институт

МФТИ, г. Долгопрудный

Цель: предложить вероятностную постановку задачи дистилляции моделей глубокого обучения на основе существующих методов дистилляции и привилегированного обучения.

Задачи

1. Поставить вероятностную задачу дистилляции для задачи классификации и регрессии.
2. Провести теоретический анализ предложенной вероятностной постановки задачи для линейных моделей.

Исследуемая проблема

1. Снижение размерности пространства параметров моделей глубокого обучения.

Метод решения

Предлагается поставить вероятностную постановку задачи дистилляции моделей глубокого обучения. В качестве базовой дистилляции предлагается использовать методы предложенные Дж. Хинтоном и В. Вапником.

Список литературы

1. *Грабовой А. В., Стрижов В. В.* Анализ моделей привилегированного обучения и дистилляции // Автоматика и Телемеханика (на рассмотрении)
2. *Christopher Bishop*, Pattern Recognition and Machine Learning, 2016.
3. *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying Distillation and Privileged Information // In International Conference on Learning Representations. Puerto Rico, 2016.
4. *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. 2015.
5. *Madala H., Ivakhnenko A.* Inductive Learning Algorithms for Complex Systems Modeling. Boca Raton: CRC Press Inc., 1994.

Когда возникает задача:

1. изменение признакового описания объектов;
2. использования информации из “будущего”;
3. уменьшение сложности модели;
4. использования нескольких типо признаков.

Сам слайд мне не нравится нужно переписать... Что сюда интересное можно придумать?

Необходимые понятия

Definition

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

Definition

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

Definition

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

Definition

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

Постановка задачи обучения с учителем

Задано:

1. множество объектов $\mathbf{\Omega}$, где $|\mathbf{\Omega}| = m$;
2. множество объектов $\mathbf{\Omega}^*$, где $|\mathbf{\Omega}^*| = m^*$;
3. множество целевых переменных $\mathbb{Y}$, причем $\mathbf{y}_i = \mathbf{y}\bigl(\omega_i\bigr)$;
4. отображение $\varphi : \mathbf{\Omega} \to \mathbb{R}^n$, обозначим $\mathbf{x}_i = \varphi(\omega_i)$;
5. отображение $\varphi^* : \mathbf{\Omega}^* \to \mathbb{R}^{n^*}$, обозначим $\mathbf{x}_i^* = \varphi^*(\omega_i)$.

Введем множество объектов, для которых известна привилигированя информация:

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

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

Постановка задачи обучения с учителем

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

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

Заметим:

1. множество $\mathbb{Y}^* = \mathbb{Y}$ для задачи регрессии;
2. множество $\mathbb{Y}^*$ является единичным симплексом $\mathcal{S}_K$ в пространстве размерности K для задачи классификации.

Для удобства введем обозначения: $\mathbf{f}(\mathbf{X}^*) = \mathbf{S}$.
Требуется выбрать модель ученика $\mathbf{g}(\mathbf{x})$ из множества:

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

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

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

Постановка задачи: Хинтон

Рассматривается:

1. привилегированная информация $\mathcal{I} = \{1, 2, \cdots, m\}$;
2. классификация $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \mathbf{x}_i \in \mathbb{R}^n, y_i \in \mathbb{Y} = \{1, \cdots, K\}$.

Обозначим y_i — класс объекта, а $\mathbf{y}_i$ вектор вероятности для i-го объекта.
Параметрическое семейство учителя и ученика:

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

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

где $\mathbf{z}, \mathbf{v}$ — это дифференцируемые параметрические функции 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:

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

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

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

Рассматривается:

1. привилегированная информация $\mathcal{I} = \{1, 2, \cdots, m\}$;
2. классификация $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{x}_i^*, y_i)\}_{i=1}^{m}, \mathbf{x}_i \in \mathbb{R}^n, \mathbf{x}_i^* \in \mathbb{R}^{n^*}, y_i \in \{1, \cdots, K\}$.

Параметрическое семейство учителя и ученика:

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

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

где $\mathbf{z}, \mathbf{v}^*$ — это дифференцируемые параметрические функции заданной структуры, T — параметр температуры.

Функция потерь:

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

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 поиске оптимальной модели $\mathbf{g} \in \mathfrak{G}_{\text{cl}}$.
Рассматривается двухэтапная модель обучения:

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.

Модель ученика — это функция, которая минимизирует $\mathcal{L}_{st}$.
Модель учителя — это функция, которая минимизирует кросс–энтропийную функции ошибки:

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

Принцип максимума правдоподобия:

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

Вероятностные предположения:

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 для одних и тех же объектов;
4. если $|\mathbf{\Omega}^*| = 0$ то решение должно соответствовать решению максимума правдоподобия.

Совместное правдоподобие истинных меток и меток учителя:

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

Совместное правдоподобие истинных меток и меток учителя:

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

По формуле условной вероятности:

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

Получаем:

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

Заметим, что $\mathbf{y}_i$ и $\mathbf{s}_i$ зависимы только через переменную $\mathbf{x}_i$:

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

Совместное правдоподобие:

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

Получаем оптимизационную задачу для поиска $\hat{\mathbf{g}}$:

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

Для удобства минимизируется логарифм:

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

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

Частный случай: классификация

Для задачи многоклассовой классификации рассматриваются следующие вероятностные предположения:

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

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

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

Частный случай: классификация

Theorem (Грабовой 2020)

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

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

функция $p(\mathbf{s}|\mathbf{x}, \mathbf{g}) = C \prod_{k=1}^{K} g_k(\mathbf{x})^{s^k}$ является плотностью распределения.

Получаем оптимизационную задачу:

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

Задача регрессии имеет вероятностные предположения:

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

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

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

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

3. истинные метки имеют нормальное распределение

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

4. метки учителя распределены

$$p\left(s|\mathbf{x}, \mathbf{g}\right) = \mathcal{N}\left(s|\mathbf{g}\left(\mathbf{x}\right), \sigma_s\right);$$

Оптимизационная задача:

$$\hat{g} = \arg\min_{g \in \mathcal{G}} \sum_{i \not\in \mathcal{I}} \sigma^2 \left(y_i - \mathbf{g}\left(\mathbf{x}_i\right)\right)^2 + (1-\lambda) \sum_{i \in \mathcal{I}} \sigma^2 \left(y_i - \mathbf{g}\left(\mathbf{x}_i\right)\right)^2 + \lambda \sum_{i \in \mathcal{I}} \sigma_s^2 \left(s_i - \mathbf{g}\left(\mathbf{x}_i\right)\right)^2.$$

Оптимизационная задача:

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

Theorem (Грабовой 2020)

Пусть множество $\mathcal{G}$ описывает класс линейных функций вида $\mathbf{g}(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x}$. Тогда решение оптимизационной задачи эквивалентно решению следующей задачи линейной регрессии $\mathbf{y}'' = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, где $\boldsymbol{\Sigma}^{-1} = diag(\boldsymbol{\sigma}')$ и $\mathbf{y}''$ имеют следующий вид:

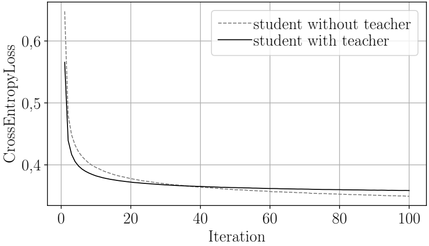
$$\sigma_i' = \begin{cases} \sigma^2, \text{ если } i \notin \mathcal{I} \\ (1-\lambda)\,\sigma^2 + \lambda\sigma_s^2, \text{ иначе} \end{cases},$$

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

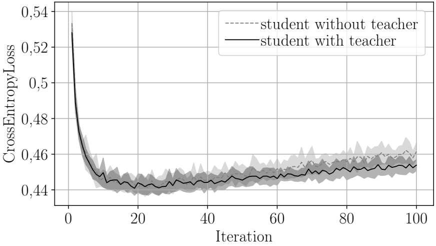
$$y_i' = \begin{cases} \sigma^2 y_i, \text{ если } i \notin \mathcal{I} \\ (1-\lambda)\,\sigma^2 y_i + \lambda\sigma_s^2 s_i, \text{ иначе} \end{cases}.$$

Вычислительный эксперимент состоит из следующих частей:

1. эксперимент с выборкой FashionMNIST;
2. эксперимент на синтетической выборке;
3. эксперимент на выборке Twitter Sentiment Analysis.



(a)



(b)

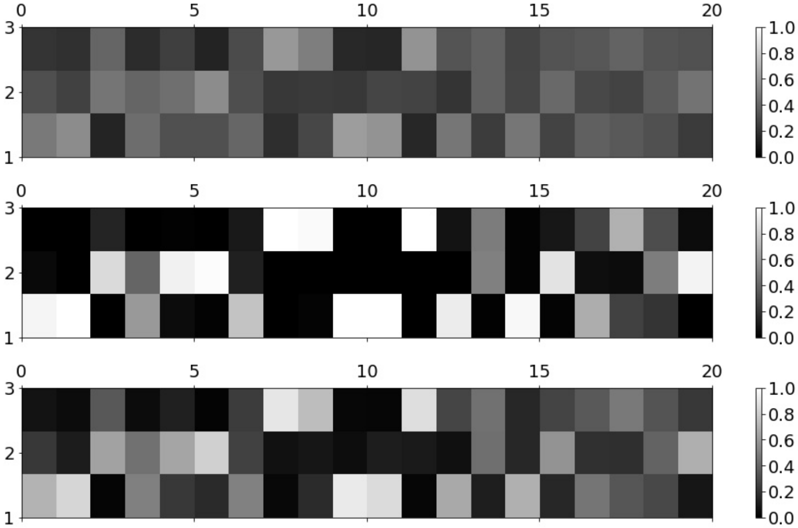
Зависимость кросс–этропии между истинными метками и предсказанными учеников вероятностями классов: a) на обучающей выборке; b) на тестовой выборке.

Синтетический эксперимент

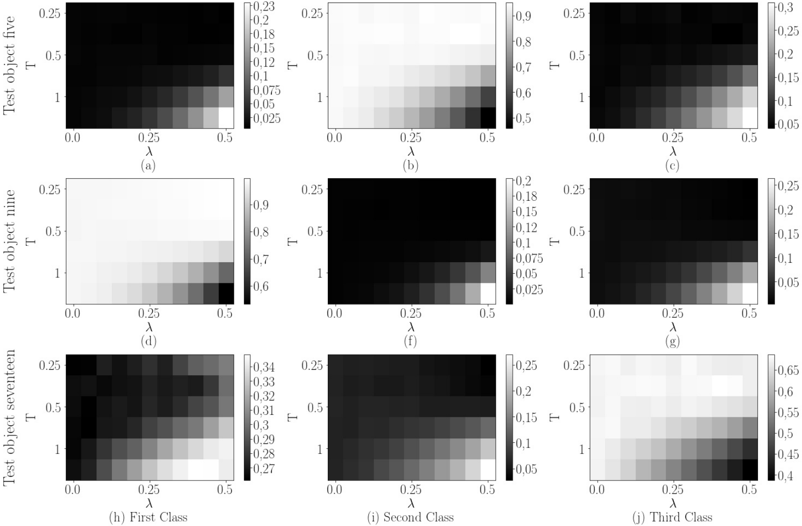
Выборка построенная следующим образом:

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

где функция softmax берется построчно. Строки матрицы $\mathbf{S}$ будем рассматривать как предсказание учителя, то есть учитель знает истинные вероятности каждого класса.
В эксперименте число признаков $n = 10$, число классов $K = 3$, для обучения было сгенерировано $m_{\text{train}} = 1000$ и $m_{\text{test}} = 100$ объектов.



Сверху вниз: истинное распределение; без учителя; с учителем



Зависимость распределения по классам при разных параметрах λ и T

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".

Описание моделей:

- модель учителя: модель на основе Bi-LSTM с ≈ 30 миллионов настраиваемых параметров;
- модель ученика: модель на основе предобученной модели BERT с 1538 настраиваемых параметров.

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

Сделано:

1. поставлена вероятностная задача дистилляции моделей глубокого обучения;
2. проведен теоретический анализ предложенной вероятностной задачи;
3. проведен вычислительный эксперимент для анализа предложенной модели.

Планируется:

1. обобщить предложенный метод на случай задачи регрессии более корректно;
2. использовать байесовский подход выбора моделей машинного обучения для решения данной задачи.