

Привилегированная информация и дистилляция моделей

Грабовой Андрей Валериевич

Московский физико-технический институт

МФТИ, г. Долгопрудный

Цель: предложить вероятностную постановку задачи дистилляции модели 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 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.

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

Когда возникает задача:

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

Сам слайд мне не нравится нужно переписать....

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.

Задано:

- ❶ множество объектов Ω , где $|\Omega| = m$;
- ❷ множество объектов Ω^* , где $|\Omega^*| = m^*$;
- ❸ множество целевых переменных \mathbb{Y} , причем $\mathbf{y}_i = \mathbf{y}(\omega_i)$;
- ❹ отображение $\varphi : \Omega \rightarrow \mathbb{R}^n$, обозначим $\mathbf{x}_i = \varphi(\omega_i)$;
- ❺ отображение $\varphi^* : \Omega^* \rightarrow \mathbb{R}^{n^*}$, обозначим $\mathbf{x}_i^* = \varphi^*(\omega_i)$.

Введем множество объектов, для которых известна привилегированная информация:

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

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

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

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

Заметим:

- ❶ множество $\mathbb{Y}^* = \mathbb{Y}$ для задачи регрессии;
- ❷ 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:

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

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

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

Рассматривается:

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

Обозначим y_i — класс объекта, а \mathbf{y}_i вектор вероятности для i -го объекта.

Параметрическое семейство учителя и ученика:

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

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

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

Параметр температуры T имеет следующие свойства:

- 1 при $T \rightarrow 0$ получаем вектор, в котором один из классов имеет единичную вероятность;
- 2 при $T \rightarrow \infty$ 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:

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

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

Получаем оптимизационную задачу:

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

Рассматривается:

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

Параметрическое семейство учителя и ученика:

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

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

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 равен 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:

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

Вероятностные предположения:

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

Совместное правдоподобие истинных меток и меток учителя:

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

Совместное правдоподобие истинных меток и меток учителя:

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

По формуле условной вероятности:

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

Получаем:

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

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

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

Совместное правдоподобие:

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

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

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

Для удобства минимизируется логарифм:

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

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

Для задачи многоклассовой классификации рассматриваются следующие вероятностные предположения:

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

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

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

Theorem (Грабовой 2020)

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

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

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:

$$\begin{aligned} \hat{\mathbf{g}} = \arg \max_{\mathbf{g} \in \mathcal{G}} & \sum_{i \notin \mathcal{I}} \sum_{k=1}^K y_i^k \log g_k(\mathbf{x}_i) \Big|_{T=1} \\ & + (1 - \lambda) \sum_{i \in \mathcal{I}} \sum_{k=1}^K y_i^k \log g_k(\mathbf{x}_i) \Big|_{T=1} + \lambda \sum_{i \in \mathcal{I}} \sum_{k=1}^K s_{i,k} \log g_k(\mathbf{x}_i) \Big|_{T=T_0} \\ & + \lambda \sum_{i \in \mathcal{I}} \sum_{k=1}^K \left(\log g_k(\mathbf{x}_i) \Big|_{T=T_0} + \log \log \frac{1}{g_k(\mathbf{x}_i)} \Big|_{T=T_0} \right). \end{aligned}$$

Задача регрессии имеет вероятностные предположения:

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

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

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

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

- ③ истинные метки имеют нормальное распределение

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

- ④ метки учителя распределены

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

Оптимизационная задача:

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

Оптимизационная задача:

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

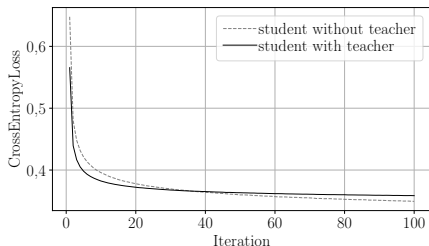
Theorem (Грабовой 2020)

Пусть множество \mathcal{G} описывает класс линейных функций вида $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. 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 $\boldsymbol{\Sigma}^{-1} = \text{diag}(\boldsymbol{\sigma}')$ и \mathbf{y}'' имеют следующий вид:

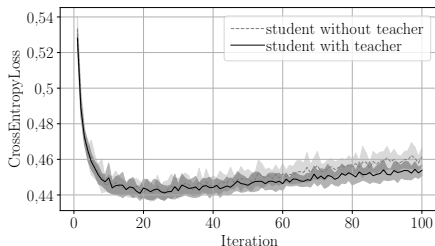
$$\begin{aligned} \sigma'_i &= \begin{cases} \sigma^2, & \text{если } i \notin \mathcal{I} \\ (1 - \lambda) \sigma^2 + \lambda \sigma_s^2, & \text{иначе} \end{cases}, \\ \mathbf{y}'' &= \boldsymbol{\Sigma} \mathbf{y}', \\ y'_i &= \begin{cases} \sigma^2 y_i, & \text{если } i \notin \mathcal{I} \\ (1 -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

Вычислительный эксперимент состоит из следующих частей:

- ❶ эксперимент с выборкой FashionMNIST;
- ❷ эксперимент на синтетической выборке;
- ❸ эксперимент на выборке Twitter Sentiment Analysis.



(a)



(b)

Зависимость кросс-энтропии между истинными метками и предсказанными учеников вероятностями классов: а) на обучающей выборке; б) на тестовой выборке.

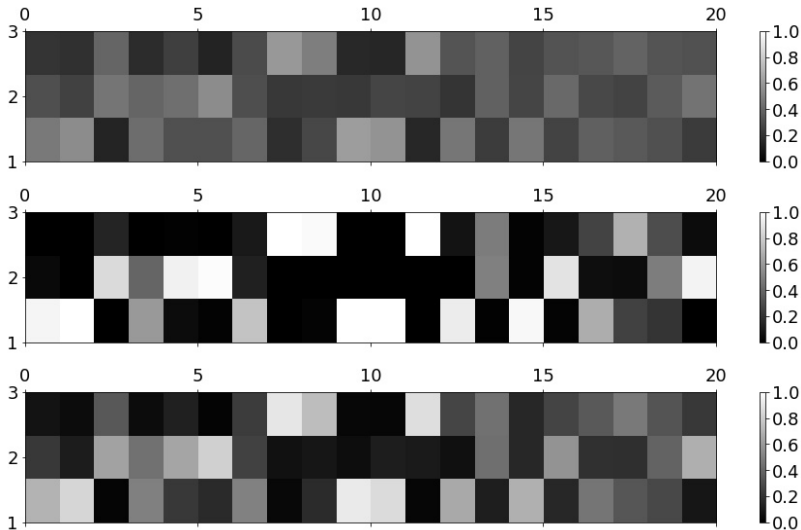
Выборка построенная следующим образом:

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

где функция softmax берется построчно. Строки матрицы \mathbf{S} будем рассматривать как предсказание учителя, то есть учитель знает истинные вероятности каждого класса.

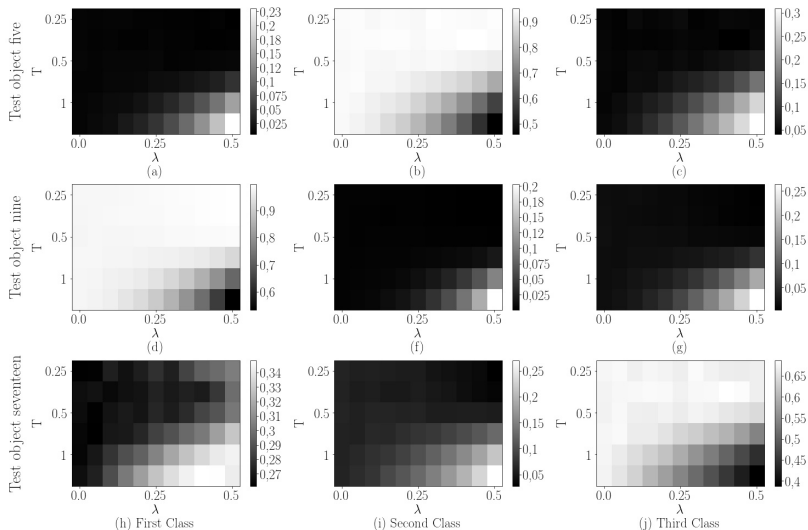
В эксперименте число признаков $n = 10$, число классов $K = 3$, для обучения было сгенерировано $m_{\text{train}} = 1000$ и $m_{\text{test}} = 100$ объектов.

Синтетический эксперимент: распределение классов



Сверху вниз: истинное распределение; без учителя; с учителем

Синтетический эксперимент: анализ параметра λ и T



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:

- ❶ поставлена вероятностная задача дистилляции моделей глубокого обучения;
- ❷ проведен теоретический анализ предложенной вероятностной задачи;
- ❸ проведен вычислительный эксперимент для анализа предложенной модели.

Планируется:

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