Информативные априорные предположения в задаче привилегированного обучения

Нейчев Радослав Георгиев

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель:
д.ф.-м.н, В.В. Стрижов

21 июня 2018

Цели исследования

Цель исследования: создать метод построения моделей оптимальной сложности для задач распознавания и прогнозирования.
Проблема: Неустойчивая сходимость параметров моделей в зависимости от начальной инициализации.

Требуется предложить метод построения моделей, который

- соблюдает баланс между точностью и сложностью модели,
- использует дополнительную (априорную) информацию на этапе обучения и учитывает объекты, априорная информация о которых отсутствует (использует неполные описания объектов).

Обзор литературы

- Фильтрация объектов в задаче многоклассовой классификации.
  Object selection in credit scoring using covariance matrix of parameters estimations. A. Aduenko, A. Motrenko, V. Strijov, Annals of Operations Research, 2018.

- Использование привилегированного обучения применительно к SVM.
  Learning using privileged information: Similarity control and knowledge transfer. V.Vapnik, R.Izmailov. JMLR, 2015.

- Обобщение подходов Вапника и Хинтона к привилегированному обучению.
  Unifying distillation and privileged information. B.Schlölkopf, V.Vapnik, D.Lopez-Paz, L.Bottou. ICLR, 2016.

Постановка задачи декодирования, прогнозирования и классификации

Заданы матрица объект-признак $\mathbf{X}$ и целевая матрица $\mathbf{Y}$.

$$\left[\begin{array}{c|ccc} \hat{\mathbf{y}}' & \mathbf{x}'_0 & \mathbf{x}''_0 & \cdots \\ \hline \mathbf{y}'_1 & \mathbf{x}'_1 & \mathbf{x}''_1 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}'_m & \mathbf{x}'_m & \mathbf{x}''_m & \cdots \end{array}\right] = \left[\begin{array}{c|c} \underset{1\times r}{\hat{\mathbf{y}}} & \underset{1\times n}{\mathbf{x}_0} \\ \hline \underset{m\times r}{\mathbf{Y}} & \underset{m\times n}{\mathbf{X}} \end{array}\right].$$

Оптимальная модель $\hat{\mathbf{f}} : \mathbf{x} \to \mathbf{y},\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{y} \in \mathbb{R}^r$ минимизирует заданную функцию ошибки S при ограничении на сложность,

$$\hat{\mathbf{f}} = \operatorname{argmin}_{\mathbf{f}\in\mathfrak{F}} S(\mathbf{f}, \mathbf{X}, \mathbf{Y}),\ \text{при}\ |\hat{\mathbf{f}}|_c \le M_c.$$

Предлагается использовать *привилегированную априорную информацию* при построении $\hat{\mathbf{f}}$.

Шлюзовая функция $\pi_k(\mathbf{x}) : \mathbf{x} \to [0;1]$ определяет правдоподобие k-й модели на $\mathbf{x}$.

$$\pi_k(\mathbf{x}, \mathbf{V}) = \boldsymbol{\sigma}\Big(\mathbf{g}(\mathbf{x}, \boldsymbol{\omega}), \mathbf{V}\Big) = \frac{\exp \mathbf{v}_k^\top \mathbf{g}(\mathbf{x}, \boldsymbol{\omega})}{\sum_{k'=1}^{K} \exp \mathbf{v}_k'^\top \mathbf{g}(\mathbf{x}, \boldsymbol{\omega})},$$

где $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K, \boldsymbol{\omega}]$, $\boldsymbol{\sigma}$ — softmax,
$\mathbf{g}(\mathbf{x}, \boldsymbol{\omega})$ — преобразование над $\mathbf{x}$.

Смесь экспертов: $\mathbf{f}_{\text{me}}(\mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}, \mathbf{V}) \mathbf{k}(\mathbf{x})$.

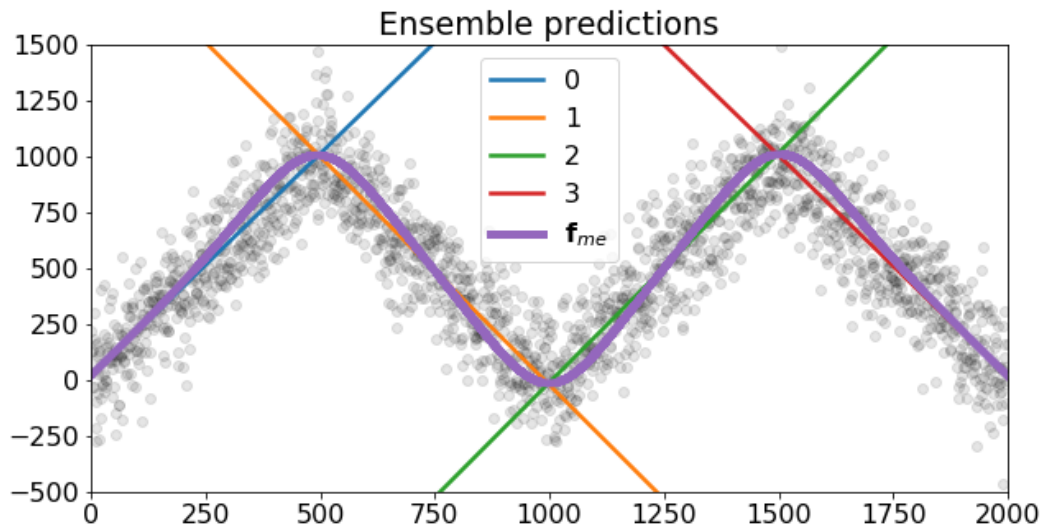
Апостериорное распределение на $\mathbf{y}$:

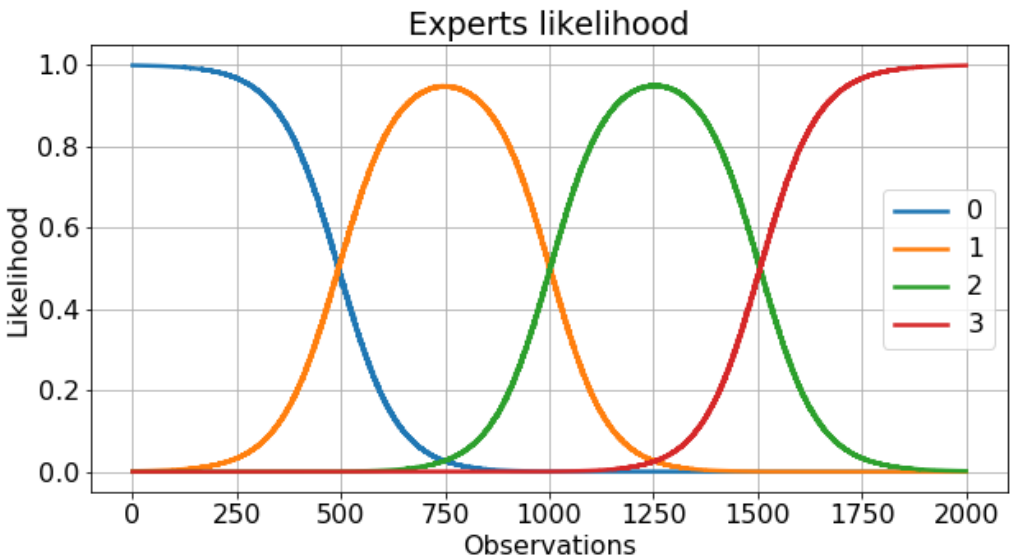
$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} p(k|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{y}|k, \mathbf{x}, \boldsymbol{\theta}) =$$

$$= \sum_{k=1}^{K} \pi_k(\mathbf{x}, \mathbf{V}) \exp\left(-\frac{1}{2\beta_k}\big(\mathbf{y} - f_k(\mathbf{x}, bw)\big)^2\right), \text{где}$$

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}_k(\mathbf{x}, \mathbf{w}_k), \beta_k), \quad \boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{V}, \boldsymbol{\beta}].$$

Смесь экспертов на синтетических данных





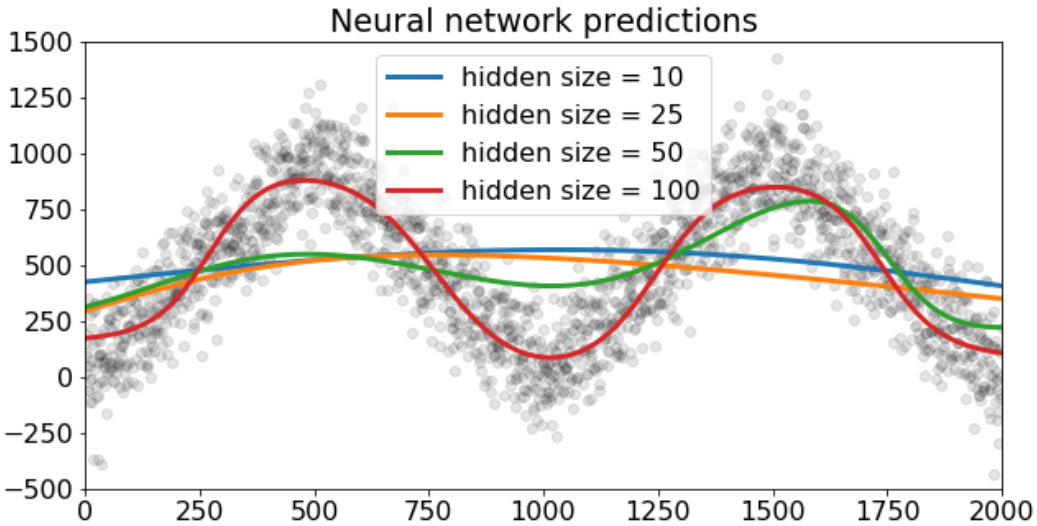
$$\mathbf{f}_{\text{me}} = \sum_{k=1}^{K} \pi_k \mathbf{f}_k,$$

$\boldsymbol{\pi}(\mathbf{x}, \mathbf{V})$ — нейросеть с одним скрытым слоем из 50 нейронов,

$$\mathbf{f}_k = \mathbf{w}_k \mathbf{x} + b_k$$

$$|\mathbf{f}_{\text{me}}|_C \sim 10^2$$

Сравнение нейросетей на синтетических данных



$$|\mathbf{f}_{\text{NN}}|_C \sim 10^4,$$

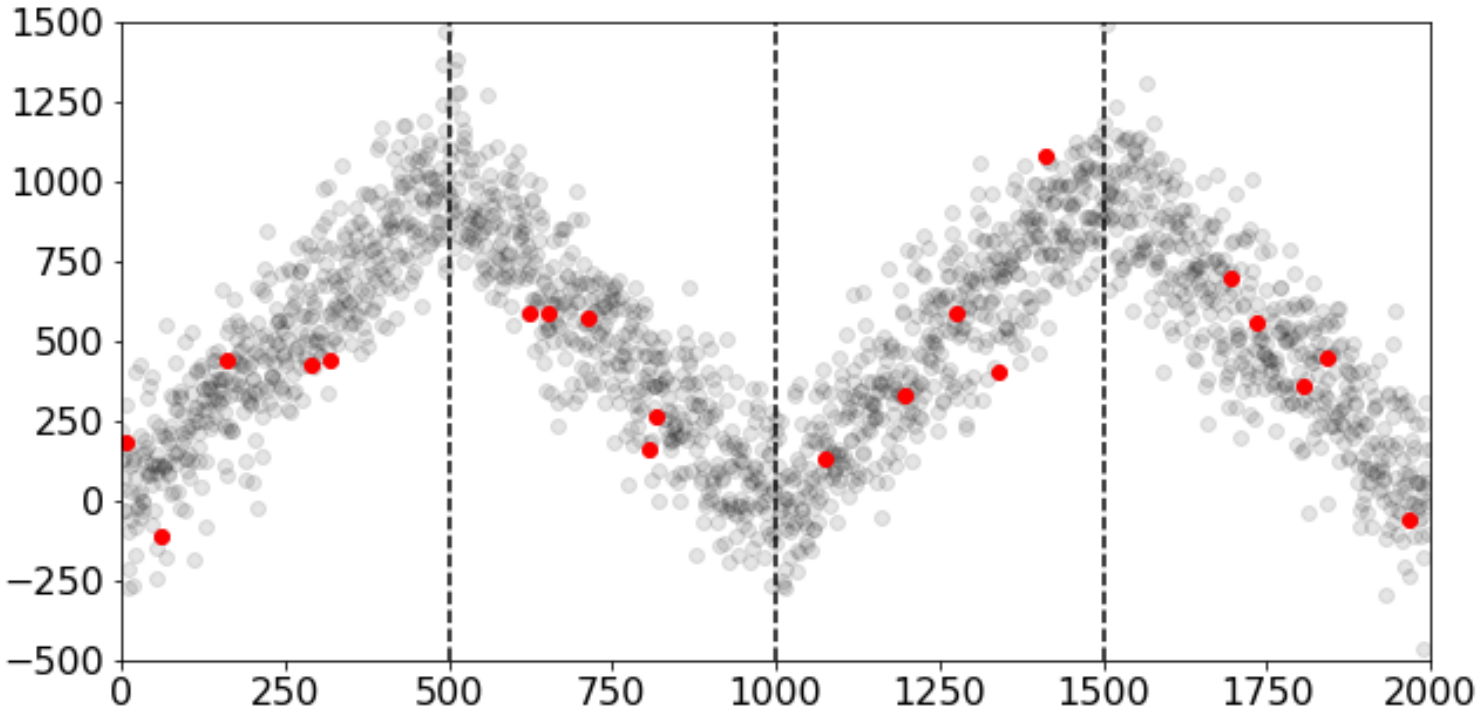
$$S(\mathbf{f}_{\text{me}}) \approx S(\mathbf{f}_{\text{NN}}),$$

$$|\mathbf{f}_{\text{me}}|_C \ll |\mathbf{f}_{\text{NN}}|_C.$$

Лишь экземпляр $\mathbf{f}_{\text{NN}}$ с размером скрытого слоя 100 смог адекватно описать данные.

Проблема: Параметры мультимодели $\mathbf{f}_{\text{me}}$ очень плохо сходится ($\sim$ в 10% запусков со случайной инициализацией параметров).

Решение: использовать привилегированную информацию о принадлежности некоторых объектов к различным экспертам. Случайные 5 точек для каждого сегмента зафиксированы за различными экспертами



С привлечением привилегированной информации сходимость достигается $\sim$ в 76% запусков.

Мета-обучение (distillation)

Пусть для некоторых объектов $\mathbf{x}$ доступна *привилегированная* информация $\mathbf{x}^*$. Введем функции ученика $\mathbf{f}_\mathsf{s} \in \mathfrak{F}_\mathsf{s}$ (student) и учителя $\mathbf{f}_\mathsf{t} \in \mathfrak{F}_\mathsf{t}$ (teacher):

$$\mathbf{f}_\mathsf{s} : \mathbf{x} \longrightarrow \mathbf{y}, \quad \mathbf{f}_\mathsf{t} : \mathbf{x}, \mathbf{x}^* \longrightarrow \mathbf{y}.$$

$$\mathbf{f}_\mathsf{s} = \operatorname{argmin}_{\mathbf{f} \in \mathfrak{F}_s} \frac{1}{n} \sum_{i=1}^{n} \Big[(1-\lambda) S\big(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i)\big) + \lambda S\big(\mathbf{s}_i, \mathbf{f}(\mathbf{x}_i)\big) \Big],$$

в задаче классификации (T — температура сглаживания):

$$\mathbf{s}_i = \boldsymbol{\sigma}(\mathbf{f}_\mathsf{t}(\mathbf{x}_i)/T), \; S\big(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i)\big) = -\sum_{k=1}^{c} \mathbf{y}_k \log \boldsymbol{\sigma}\big(\mathbf{f}(\mathbf{x}_i)\big), \; \boldsymbol{\sigma} \text{ — softmax},$$

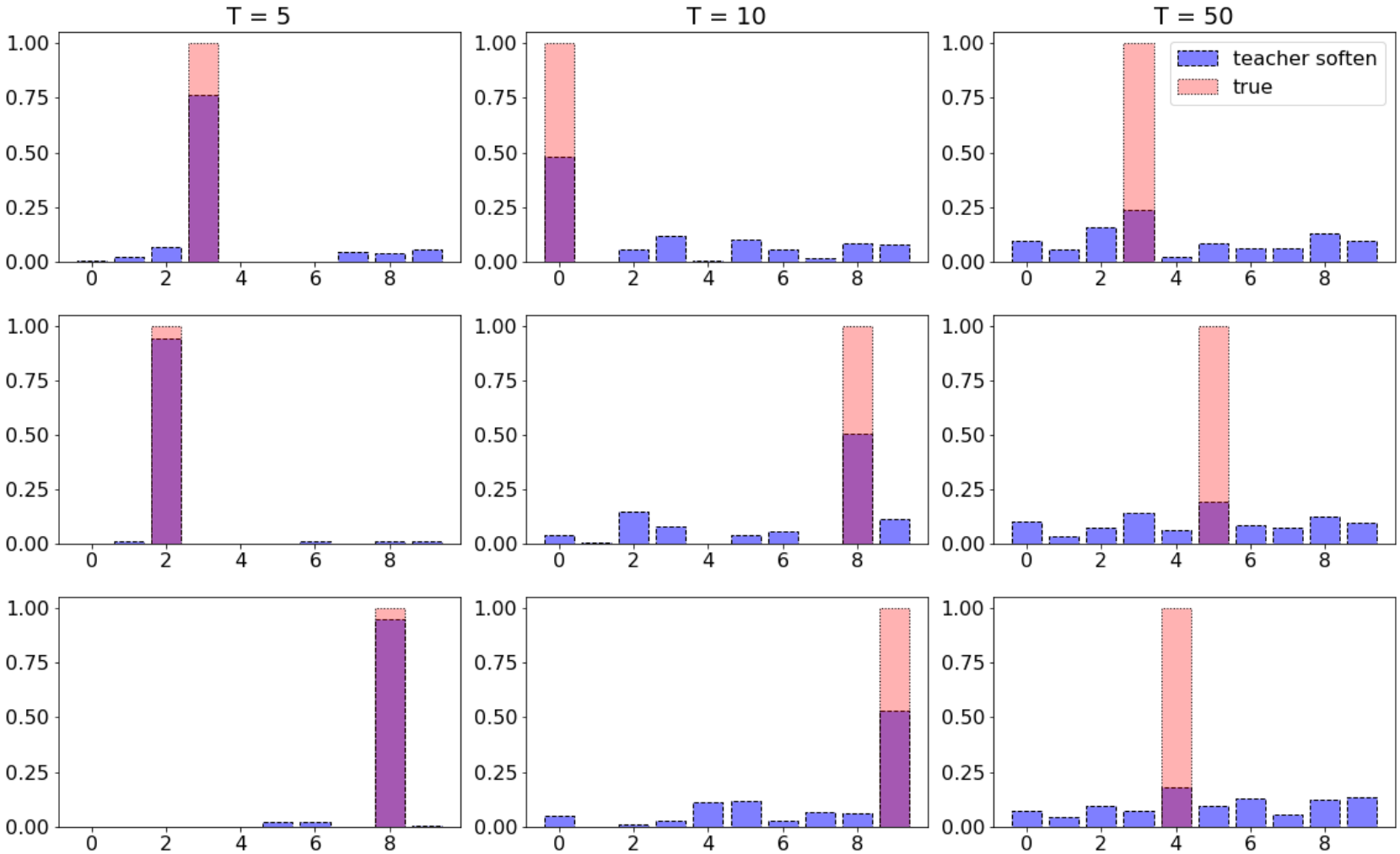
в задаче декодирования ($\mathbf{T}$ — ширина окна):

$$\mathbf{s}_i = \big[\mathbf{f}_\mathsf{t}(\mathbf{x}_i) - \mathbf{T}; \mathbf{f}_\mathsf{t}(\mathbf{x}_i) + \mathbf{T}\big], \; S\big(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i)\big) = \|\mathsf{mean}(\mathbf{y}_i) - \mathbf{f}(\mathbf{x}_i)\|_2 \mathbb{I}\big(\mathbf{f}(\mathbf{x}_i) \notin \mathbf{y}_i\big).$$

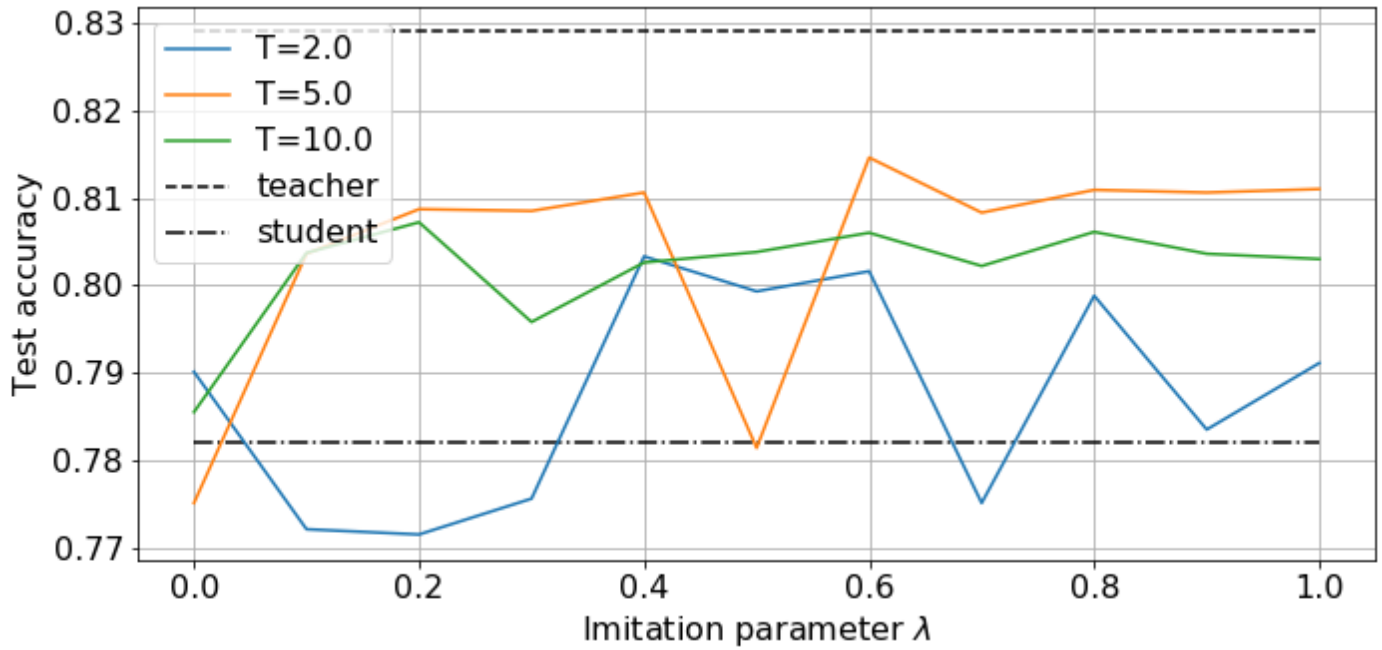
$|\mathcal{F}_\mathsf{t}|_C \gg |\mathcal{F}_\mathsf{s}|_C$, $\mathbf{x}^* = \emptyset$ — дистилляция (Хинтон).

$|\mathcal{F}_\mathsf{t}|_C \ll |\mathcal{F}_\mathsf{s}|_C$, $\mathbf{x}^* \neq \emptyset$ — привилегированное обучение (Вапник).

Иллюстрация сглаженных предсказаний учителя $\mathbf{s}_i$ в зависимости от значения параметра T на примере классификации датасета MNIST.



Качество классификации ученика, обученного методом дистилляции в зависимости от параметров T и λ.



Обучающая выборка — 500 изображений из датасета MNIST, $\mathbf{x}^*$ — исходные изображения, $\mathbf{x}$ — изображения с разрешением в 4 раза меньше, $\mathbf{f}_t$ и $\mathbf{f}_s$ — нейросети с двумя скрытыми слоями из 50 нейронов и ReLU-активациями. Число параметров ученика значительно меньше, чем учителя:

$$|\mathbf{f}_s|_C = 1.5 \cdot 10^3 \ll 1.5 \cdot 10^4 = |\mathbf{f}_t|_C.$$

Заключение

- Предложен метод построения модели меньшей сложности с использованием привилегированной информации.
- Предложенный метод позволяет использовать частично доступные данные.
- Создан фреймворк, позволяющий применять дистилляцию для обучения сторонних моделей.

Публикации по теме:

- *Выбор оптимального набора признаков из мультикоррелирующего множества в задаче прогнозирования.* Нейчев Р.Г., Катруца А.М., Стрижов В.В., Заводская лаборатория, 2016.
- *Heterogeneous model selection for multiscale time series forecasting.* R.Neychev, A.Motrenko, E.Gaussier and V.Strijov, Рассматривается редколлегией Journal of Applied Mathematics and Computation.
- *LHCb trigger streams optimization.* R.Neychev, N.Kazeev, A.Panin et al, CHEP-2016 proceedings.
- *Machine-Learning-based global particle-identification algorithms at the LHCb experiment.* R.Neychev, D.Derkach et al, Journal of Physics: Conference series, 2017

Сопутствующие результаты:

Разработана и запущена в эксплуатацию автоматическая система прогнозирования энергопотребления ДЦ компании Яндекс.

Backup

Определения

Априорные знания — информация о предметной области/ограничениях на решение, не представленная в обучающей выборке в явном виде.
Привилегированная информация — дополнительная информация об объектах обучающей выборки, доступная только на этапе обучения.
Сложность модели $|f|_C.|.|_C$ используется оценка числа простейших арифметических операций на единичном входе