

# Информативные априорные предположения в задаче привилегированного обучения

Радослав Нейчев, Вадим Стрижов

Московский физико-технический институт

ИОИ-2018, Гаэта, Италия

# Цели исследования

**Цель исследования:** Создать метод построения мультимodelей оптимальной сложности для задач распознавания и декодирования.

**Проблемы:**

- ▶ Неустойчивая сходимость параметров моделей в зависимости от начальной инициализации.
- ▶ Сложность использования информативной априорной информации.

Требуется предложить метод построения моделей, который

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

- ▶ 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 классификации и декодирования

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

$$\left[ \begin{array}{c|c} \hat{\mathbf{y}} & \mathbf{x}_0 \\ \hline 1 \times r & 1 \times n \\ \hline \mathbf{Y} & \mathbf{X} \\ m \times r & m \times n \end{array} \right]$$

Оптимальная модель  $\hat{\mathbf{f}} : \mathbf{x} \rightarrow \mathbf{y}$ ,  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{y} \in \mathbb{R}^r$  аппроксимирует истинную зависимость, минимизируя заданную функцию ошибки  $S$  при ограничении на сложность,

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

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

# Априорная информация

Априорная информация может быть представлена как:

- ▶ экспертная разметка для части данных,
- ▶ ограничения на решение, основанные на знания о предметной области и особенностях решаемой задачи,
- ▶ начальная оценка параметров распределений параметров модели.

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

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

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

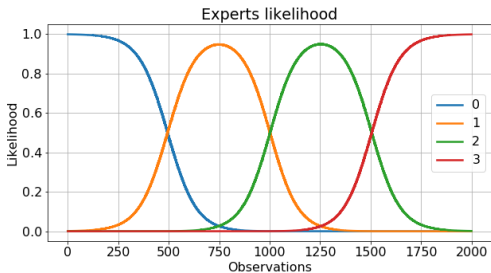
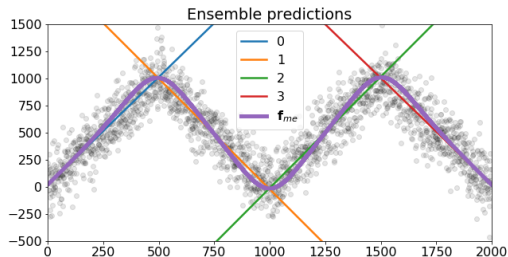
$$\text{Смесь экспертов: } \mathbf{f}_{\text{me}}(\mathbf{x}) = \sum_{k=1}^K \pi_k(\mathbf{x}, \mathbf{V}) \mathbf{k}(\mathbf{x}).$$

Апостериорное распределение на  $\mathbf{y}$ :

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^K p(k|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{y}|k, \mathbf{x}, \boldsymbol{\theta}) = \\ = \sum_{k=1}^K \pi_k(\mathbf{x}, \mathbf{V}) \exp\left(-\frac{1}{2\beta_k} (\mathbf{y} - f_k(\mathbf{x}, \mathbf{w}_k))^2\right), \text{ где}$$

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

# Смесь экспертов на синтетических данных



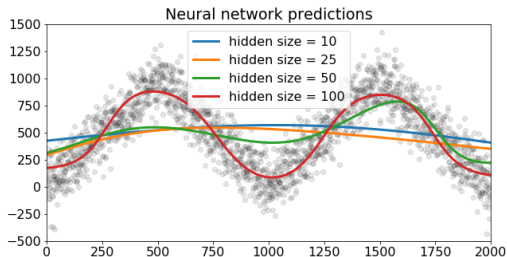
$$\mathbf{f}_{me} = \sum_{k=1}^K \pi_k \mathbf{f}_k,$$

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

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

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

# Сравнение нейронных сетей на синтетических данных



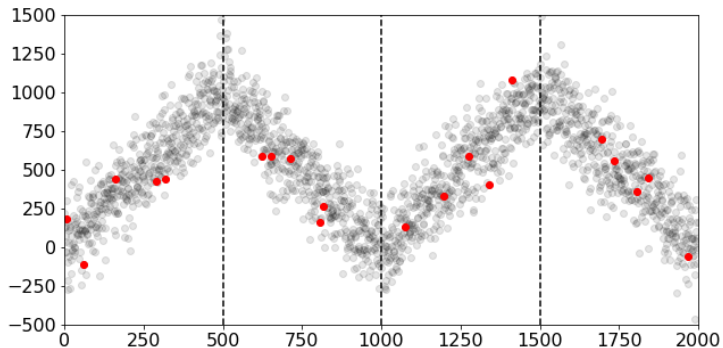
$$|\mathbf{f}_{\text{NN}}|_C \sim 10^4,$$
$$S(\mathbf{f}_{\text{me}}) \approx S(\mathbf{f}_{\text{NN}}),$$
$$|\mathbf{f}_{\text{me}}|_C \ll |\mathbf{f}_{\text{NN}}|_C.$$

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

**Проблема:** Параметры мультимодели  $\mathbf{f}_{\text{me}}$  очень плохо сходятся ( $\sim$  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  $\mathbf{f}_s \in \mathfrak{F}_s$  (student) и учителя  $\mathbf{f}_t \in \mathfrak{F}_t$  (teacher):

$$\mathbf{f}_s : \mathbf{x} \longrightarrow \mathbf{y}, \quad \mathbf{f}_t : \mathbf{x}, \mathbf{x}^* \longrightarrow \mathbf{y}.$$

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

в задаче классификации ( $T$  — температура сглаживания):

$$\mathbf{s}_i = \sigma(\mathbf{f}_t(\mathbf{x}_i)/T), \quad S(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i)) = - \sum_{k=1}^c \mathbf{y}_k \log \sigma(\mathbf{f}(\mathbf{x}_i)), \quad \sigma — \text{softmax},$$

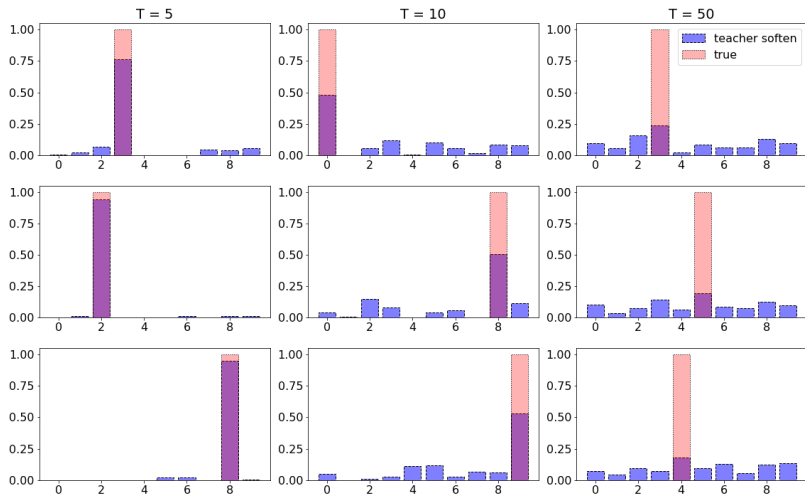
в задаче декодирования ( $\mathbf{T}$  — ширина окна):

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

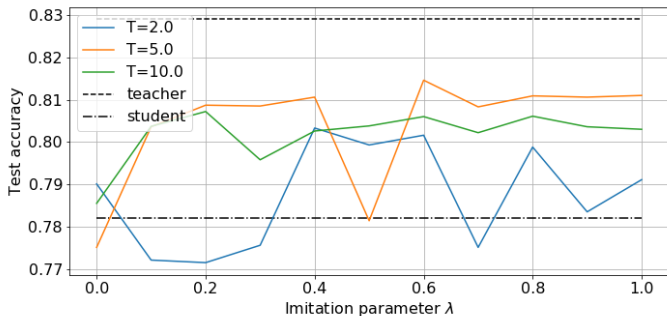
$|\mathcal{F}_t|_C \gg |\mathcal{F}_s|_C, \mathbf{x}^* = \emptyset$  — дистилляция (Хинтон).

$|\mathcal{F}_t|_C \ll |\mathcal{F}_s|_C, \mathbf{x}^* \neq \emptyset$  — привилегированное обучение (Вапник). 10/13

Иллюстрация сглаженных предсказаний учителя  $s_i$  в зависимости от значения параметра  $T$  на примере классификации датасета MNIST.



Качество классификации ученика, обученного методом дистилляции в зависимости от параметров  $T$  и  $\lambda$ .



Обучающая выборка — 500 изображений из датасета MNIST,  $\mathbf{x}^*$  — исходные изображения,  $\mathbf{x}$  — изображения с разрешением в 4 раза меньше,  $\mathbf{f}_t$  и  $\mathbf{f}_s$  — нейронные сети 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

- ▶ Использование априорной информации позволяет строить более устойчивые и простые модели.
- ▶ Отсутствие привилегированных описаний для части данных не является проблемой.

# Backup

## Публикации и ссылки по теме

- ▶ Information Bottleneck Learning Using Privileged Information for Visual Recognition. Saeid Motiian, Marco Piccirilli et.al.
- ▶ A new learning paradigm: Learning using privileged information. Vladimir Vapnik, Akshay Vashist.
- ▶ Understanding LUPI (Learning using Privileged Information) Ahmadreza Momeni, Kedar Tatwawadi.
  
- ▶ Stack-exange post about incorporating new information in prior Dirichle distribution
- ▶ Great Quora thread about prior knowledge.

**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**  $|f|_{C \cdot} |C$  используется оценка числа простейших арифметических операций на единичном входе