Министерство науки и высшего образования
Российской Федерации
"<Московский физико-технический институт
(государственный университет)">
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Нейчев Радослав Георгиев

Информативные априорные предположения в задачах привилегированного обучения

03.04.01 — Прикладные математика и физика

Выпускная квалификационная работа
(магистерская диссертация)

Научный руководитель:
д. ф.-м. н. Стрижов Вадим Викторович

Москва,
2018

АННОТАЦИЯ

В данной работе рассматриваются различные подходы к построению моделей оптимальной сложности. Вычислительная сложность модели играет основную роль во многих жизненных сценариях. Носимая электроника и защищенные устройства для решения задач биометрии, устройства автоматической обработки телеметрических данных, системы потоковой аналитики результатов коллизий Большого Адронного Колайдера — все они требуют не только качественного решения соответствующих задач, но и быстрого и энергоэффективного решения. Для снижения сложности моделей рассматриваются различные подходы. В мультимоделях используется идея разделения пространства объектов на подобласти, в каждой из которых данные описываются определенной моделью или их композицией. В мета-обучении предполагается использование парадигмы учителя-и-ученика, где на этапе обучения модели-ученика используются как истинные ответы на соответствующих объектах, так и предсказания модели-учителя. Данные подходы позволяют добиться значительно лучших результатов в случае привлечения дополнительной априорной *привилегированной* информации на этапе обучения. Использование привилегированной информации улучшает сходимость оптимизационного процесса на этапе обучения, повышает итоговое качество предсказаний модели и позволяет снижать сложность модели без значимых потерь в качестве предсказаний. В ходе вычислительного эксперимента рассмотренные подходы проверяются на синтетических и реальных данных.

ВВЕДЕНИЕ

Актуальность темы. В настоящее время различные системы мониторинга и обработки данных внедряются практически во все области, от сельского хозяйства и медицины до политики и технических производств. Присутствует необходимость обработки поступающих данных локально, независимо от наличия соединения с некоторым сервером (и независимо от его существования). Оперативная обработка поступающих данных [1–3] и принятие оптимальных решений позволит снизить издержки, повысить отказоустойчивость и использовать различные устройства с большей эффективностью.

Мультимодели показывают отличные результаты во многих задачах [4, 5]. Классические подходы к ансамблированию: беггинг и градиентный бустинг [6]. На их основе построено множество других методов: комбинация беггинга и метода случайных подпространств привели к появлению Случайного леса (Random Forest) [7], который представляет собой один из лучших универсальных алгоритмов и часто используется в качестве базового подхода (наряду с линейными моделями).

Более современный подход к мультимоделированию предполагает, что вклад членов ансамбля в ответ должен зависеть от конкретного объекта. Смесь экспертов, метод, предложенный более 20 лет назад [8] в настоящее время продолжает развиваться и приносить новые значимые результаты [4]. Смесь экспертов базируется на понятии шлюзовой функции, которая определяет значимость предсказаний каждого из экспертов — членов ансамбля. В роли шлюзовой функции может выступать стандартный softmax, процесс Дирихле [9], нейронная сеть [4] и др. Свое применение данный подход нашел и в задаче прогнозирования временных рядов [10, 11].

Несмотря на значимые успехи, сходимость относительно сложных (мульти-) моделей (и смеси экспертов в частности) на этапе обучения сильно зависит от начальной инициализации параметров [12]. Частично с этой проблемой позволяет справиться использование априорной информации о решаемой задаче и/или признаковом пространстве на этапе обучения. Априорная привилегированная информация содержит в себе дополнительную разметку имеющихся данных, зачастую гораздо

более информативную, но при этом доступна только на этапе обучения и не для всех объектов.

Привилегированная информация отлично дополняет парадигму "машины учат машины" в мета-обучении [13]. Использование дополнительных моделей для нахождения лучшего решения задачи позволяет использовать неполные признаковые описания на этапе обучения и косвенно учитывать эту информацию даже для объектов, не обладающих дополнительным привилегированным описанием. Дополнительно, подобный подход позволяет снижать значимость аномальных объектов, тем самым производя отбор объектов, что благосклонно сказывается на качестве решения задачи [14].

Цель работы. Создать метод построения моделей оптимальной сложности для задач распознавания и прогнозирования.

Научная новизна. Использование дополнительной априорной информации для снижения итоговой сложности модели.

Использование априорной информации для достижения более устойчивой сходимости при оптимизации параметров мультимоделей.

Практическая ценность. Предложенный метод использования дополнительной априорной информации на этапе построения и обучения моделей позволяет снижать итоговую сложность моделей и делает процесс обучения более устойчивым. Некоторые результаты данной работы были использованы при разработке системы автоматического прогнозирования энергопотребления датацентров компании Яндекс.

Положения, выносимые на защиту:

- Предложен и математически обоснован метод отбора моделей входящих мультимодель. Проведена проверка применимости данного метода в рамках вычислительного эксперимента.
- Предложен метод повышения качества предсказаний простых моделей за счет использования привилегированной информации. Данный метод также

позволяет использовать частично доступные данные.

- Создан демонстрационный стенд, позволяющий применять полученные результаты в других экспериментах.

ПОСТАНОВКА ЗАДАЧИ.

Пусть существуют пространство объектов $\mathbb{X}$. Каждый объект x обладает некоторым признаковым описанием $\mathbf{x} \in \mathbb{R}^m$, некоторые из объектов обладают дополнительным, *привилегированным* признаковым описанием $\mathbf{x}^* \in \mathbb{R}^p$. Матрица плана $\hat{\mathbf{X}}$ включает в себя матрицу объект-признак $\mathbf{X} = [\mathbf{x}_i]_{i=1}^n$ и матрицу ответов $\mathbf{Y} = [\mathbf{y}_i]_{i=1}^n$, которая содержит метки классов, векторы распределения или векторы целевых значения в зависимости от задачи.

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

Требуется найти оптимальную модель $\hat{\mathbf{f}} : \mathbb{R}^m \to \mathbb{R}^r$, которая минимизирует заданную функцию ошибки $S(\mathbf{f}, \mathbf{X}, \mathbf{Y})$ при ограничении на сложность:

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} S(\mathbf{f}, \mathbf{X}, \mathbf{Y}), \text{ при } |\hat{\mathbf{f}}|_c \le M_c,$$

где $|\cdot|_C$ — мера сложности, M_C — некоторая константа.

3.1 Задача многоклассовой классификации.

В задаче многоклассовой классификации на r классов матрица ответов $\mathbf{Y}$ состоит из векторов $\mathbf{y}_i \in \mathbf{\Delta}^r$, где $\mathbf{\Delta}^r$ есть пространство векторов вероятности размерности r:

$$\mathbf{y}_i = [y_i^1, \dots, y_i^r],$$
$$0 \le y_i^k \le 1 \quad \forall k,$$
$$\sum_{k=1}^{r} y_i^k = 1.$$

В качестве функции ошибки S используется кросс-энтропия:

$$S\big(\mathbf{y}_i, \hat{\mathbf{f}}(\mathbf{x}_i)\big) = -\sum_{k=1}^{r} y_i^k \log \boldsymbol{\sigma}\big(\hat{f}(\mathbf{x}_i)^k\big),$$

где $\boldsymbol{\sigma} : \mathbb{R}^r \to \mathbf{\Delta}^r$ — операция softmax:

$$\sigma(\hat{\mathbf{y}})^k = \frac{\exp y^k}{\sum_{k'=1}^{r} \exp y^{k\prime}}.$$

3.2 Задача декодирования.

В задаче декодирования матрица ответов $\mathbf{Y}$ состоит из действительных векторов $\mathbf{y}_i \in \mathbb{R}^r$. В качестве функции ошибки S используются MSE, MAE, MAPE и др.

$$\text{MAE}\big(\mathbf{y}_i, \hat{\mathbf{f}}(\mathbf{x}_i)\big) = \Big\|\mathbf{y}_i - \hat{\mathbf{f}}(\mathbf{x}_i)\Big\|_1,$$

$$\text{MAPE}\big(\mathbf{y}_i, \hat{\mathbf{f}}(\mathbf{x}_i)\big) = \Big\|\frac{\big(\mathbf{y}_i - \hat{\mathbf{f}}(\mathbf{x}_i)\big)}{\mathbf{y}_i}\Big\|_1,$$

$$\text{MSE}\big(\mathbf{y}_i, \hat{\mathbf{f}}(\mathbf{x}_i)\big) = \Big\|\mathbf{y}_i - \hat{\mathbf{f}}(\mathbf{x}_i)\Big\|_2.$$

3.3 Задача прогнозирования временных рядов как задача декодирвания.

Определение: *Временной ряд* $\mathbf{s} = [s_T, \ldots, s_i, \ldots, s_1]$ — последовательность наблюдений $s_i = s(t_i)$ (обратим внимание, что в данной нотации время течет из настоящего в прошлое, что нестандартно).

Пусть задан набор из нескольких временных рядов $\mathfrak{D} = \{\mathbf{s}^q\}$, $\mathbf{s} \in \mathbb{R}^T$, $q = 1, \ldots, Q$. Каждому временному ряду $\mathbf{s}^q$ соответствует частота семплирования $1/\tau^{(q)} : t_i^{(q)} = i \cdot \tau^{(q)}$. Для временных доступна предыстория длиной Δt_p. Необходимо получить прогноз целевого временного ряда $\hat{\mathbf{s}}$ на некоторый промежуток времени Δt_r, то есть $[\hat{s}(t_i)] : T_\text{max} + \Delta t_\text{r} \geq t_i > T_\text{max}$.

3.3.1 Построение матрицы плана.

Пара векторов $\mathbf{y}_i^{(q)}$, $\mathbf{x}_i^{(q)}$ соответствует i -тому сегменту временного ряда $\mathbf{s}_i^q$:

$$[\mathbf{y}_i^q | \mathbf{x}_i^q] = \tag{1}$$

$$\underbrace{s^q(t_i), \ldots, s^q(t_i - \Delta t_\text{r})}_{\mathbf{y}_i^{(q)}}, \underbrace{\ldots, s^q(t_i - \Delta t_\text{r} - \Delta t_\text{p})}_{\mathbf{x}_i^{(q)}}],$$

Для построения матрицы плана $\hat{\mathbf{X}}$ выберем множество моментов разбиения $\{t_i\}, i = 0 \ldots (n)$ таким образом, чтобы сегменты $\mathbf{s}_i^q = [\mathbf{y}_i | \mathbf{x}_i]$, покрывающие временной ряд $\mathbf{s}^q$, были упорядочены:

$$t_{i+1} > t_i \quad \forall i, \tag{2}$$

и разобьем каждый из временных рядов $\{\mathrm{s}^q\}$. Таким образом, матрица плана будет представлять собой набор сегментов временных рядов $\{\mathrm{s}^q\}$, причем левая часть будет содержать лишь значения целевого (целевых) ряда.

$$\hat{\mathbf{X}} = \left[\begin{array}{c|c} \underset{1\times r}{\mathbf{y}_i^{(\text{target})_l}}, \ldots, \mathbf{y}_i^{(\text{target})_1} & \underset{1\times m}{\mathbf{x}_i^{(q)}}, \ldots, \mathbf{x}_i^{(1)} \end{array} \right]_{i=n}^{0}. \tag{3}$$

Оптимальная модель $\hat{\mathbf{f}} : \mathbb{R}^m \to \mathbb{R}^r$ доставляет минимум заданной функции ошибки S.

3.3.2 Процедура скользящего контроля.

Для проверки адекватности модели $\hat{\mathbf{f}}$ на базе исторических данных предлагается *процедура скользящего контроля* (4). В ее рамках прогнозирование происходит на V сегментах времени, упорядоченных хронологически. Каждый сегмент фиксированной длины $\Delta t_{\mathfrak{V}}$ соответствует матрице плана $\hat{\mathbf{X}}_{\mathfrak{v}}$ и начинается в момент времени $t_{\mathfrak{v}}$.

Описание процедуры:

1) Зафиксируем некоторое семейство функций $\mathfrak{F}$, среди которых будем искать оптимальную модель $\hat{\mathbf{f}}$. Положим $\mathfrak{v} = 0$.

2) Построим пару векторов $\mathbf{y}_{\text{val},\mathfrak{v}}$, $\mathbf{x}_{\text{val},\mathfrak{v}}$, соответствующую промежутку длиной Δt_{r} как первую сточку матрицы плана $\hat{\mathbf{X}}_{\mathfrak{v}}$.

3) Дополним матрицу плана локальной предысторией, соответствующей промежутку длиной $\Delta t_{\mathfrak{V}} - \Delta t_{\text{r}}$:

$$\hat{\mathbf{X}}_{\mathfrak{v}} = \left[\begin{array}{c|c} \ldots & \ldots \\ \hline \underset{1\times r}{\mathbf{y}_{\text{val},\mathfrak{v}}} & \underset{1\times m}{\mathbf{x}_{\text{val},\mathfrak{v}}} \\ \underset{n\times r}{\mathbf{Y}_{\text{train},\mathfrak{v}}} & \underset{n\times m}{\mathbf{X}_{\text{train},\mathfrak{v}}} \\ \hline \ldots & \ldots \end{array} \right] \Bigg\uparrow_{\mathfrak{v}}. \tag{4}$$

4) Найдем $\hat{\mathbf{f}}$ как $\hat{\mathbf{f}} = \arg\min_{\mathbf{f}\in\mathfrak{F}} S(\mathbf{f}, \mathbf{X}_{\text{train},\mathfrak{v}}, \mathbf{Y}_{\text{train},\mathfrak{v}})$ (при необходимости можно провести дополнительную процедуру кросс-валидации с использованием $\hat{\mathbf{X}}_{\mathfrak{v}}$.)

5) Оценим ошибку модели $\hat{\mathbf{f}}$ на паре $\mathbf{y}_{\text{val},\mathfrak{v}}$, $\mathbf{x}_{\text{val},\mathfrak{v}}$.

6) Увеличим $\mathfrak{v}$ на 1 и вернемся к шагу 2.

ПОСТРОЕНИЕ МУЛЬТИМОДЕЛЕЙ.

Определение: *Шлюзовая функция* $\boldsymbol{\pi} : \mathbb{R}^m \to \boldsymbol{\Delta}^K$ — отображение, определяющее правдоподобие π_k модели $\mathbf{f}_k$ для всех $k = 1, \dots K$ в мультимодели из K моделей.

$$\boldsymbol{\pi} = [\pi_1, \dots, \pi_k],$$
$$0 \leq \pi_k \leq 1 \; \forall k, \quad \sum_{k=1}^{K} \pi_k = 1.$$

Определение: *Мультимодель* — модель $\bar{\mathbf{f}} : \mathbb{R}^m \to \mathbb{R}^r$, агрегирующая предсказания других моделей $\mathbf{f}_1, \dots, \mathbf{f}_K : \mathbb{R}^m \to \mathbb{R}^r$, действующих в одних и тех же пространтсвах.

$$\bar{\mathbf{f}} = \sum_{k=1}^{K} \pi_k \mathbf{f}_k.$$

Здесь и далее будем считать, что вектор ответов $\mathbf{y}$ представляет собой предсказание неизвестной модели $\mathbf{f}$ со случайным шумом, распределенным нормально:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \mathbf{w}) + \boldsymbol{\varepsilon},$$
$$\boldsymbol{\varepsilon} \sim \mathcal{N}\Big(0, \beta\Big),$$
$$\mathbf{y} \sim \mathcal{N}\Big(\mathbf{f}(\mathbf{x}, \mathbf{w}), \beta\Big),$$

где $\mathbf{w}$ — вектор параметров модели $\mathbf{f}$.

4.1 Смесь моделей.

Определение: *Смесь моделей* — мультимодель, ответы которой представляют собой взвешенную сумму ответов всех задействованных моделей независимо от объекта.

$$\bar{\mathbf{f}} = \sum_{k=1}^{K} \pi_k \mathbf{f}_k,$$
$$\pi_k = \text{const} \;\; \forall k = 1 \dots K.$$

Пусть вектор ответов $\mathbf{y}$ порождается одной из моделей, входящих в состав мультимодели. В вероятностной постановке распределение вектора ответов $\mathbf{y}$ будет смесью нормальных гауссовских распределений:

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{y}|\mathbf{f}_k(\mathbf{x}, \mathbf{w}_k), \beta_k) = \text{/для среднеквадратичной ошибки/} = \quad (5)$$

$$= \sum_{k=1}^{K} \frac{1}{(2\pi\beta_k)^{m/2}} \exp\left((-\frac{1}{2\beta_k}\big(\mathbf{y} - \mathbf{f}_k(\mathbf{x}, \mathbf{w}_k)\big)^\top \big(\mathbf{y} - \mathbf{f}_k(\mathbf{x}, \mathbf{w}_k)\big)\right),$$

$\boldsymbol{\theta}$ — вектор гиперпараметров

$$\boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_k, \boldsymbol{\pi}, \beta_1, \ldots, \beta_K]^{\mathrm{T}},$$

где $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_k]$ — вектор правдоподобия моделей (можно рассматривать как веса моделей).

Определение гиперпараметров: Необходимо найти такие гиперпараметры $\hat{\boldsymbol{\theta}}$, чтобы функция правдоподобия достигала своего максимального значения:

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \ln p(\mathbf{y}|\boldsymbol{\theta}), \tag{6}$$

$$\ln p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}|\mathbf{f}_k(\mathbf{x}_i, \mathbf{w}_k), \beta_k) \right).$$

Для вычисления оценки максимального правдоподобия (6) вектора гиперпараметров $\boldsymbol{\theta}$ смеси моделей (5), ведем скрытые индикаторные переменные

$$Z = [\mathbf{z}_1, \ldots, \mathbf{z}_m], \quad z_{ik} \in \{0, 1\},$$

$$z_{ik} = 1 \Leftrightarrow \mathbf{y}_i \sim \mathcal{N}(\mathbf{f}_k(\mathbf{x}_i, \mathbf{w}_k), \beta_k).$$

Тогда функция правдоподобия $p(\mathbf{y}|\mathbf{x}, Z, \boldsymbol{\theta})$ примет вид

$$p(\mathbf{y}|X, Z, \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\ln \pi_k + \ln \mathcal{N}(\mathbf{y}_i|\mathbf{f}_k(\mathbf{x}_i, \mathbf{w}_k), \beta_k) \right) =$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\ln \pi_k - \frac{1}{2\beta} (\mathbf{y}_i - \mathbf{f}_k(\mathbf{x}_i, \mathbf{w}_k))^2 + \frac{n \ln \beta_k}{2} + \text{const} \right).$$

Ввиду того, что в $p(\mathbf{y}|\mathbf{x}, Z, \boldsymbol{\theta})$ появилась зависимость от случайных величин z_{ik}, для нахождения оптимальных гиперпараметров можно максимизировать математическое ожидание правдоподобия по Z

$$\mathsf{E}_Z[p(\mathbf{y}, Z|\mathbf{x}, \boldsymbol{\theta})] =$$

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \Big(\ln \pi_k - \frac{1}{2\beta} (\mathbf{y}_i - \mathbf{f}_k(\mathbf{x}_i, \mathbf{w}_k))^2 + \frac{n \ln \beta_k}{2} \Big),$$

где $\gamma_{ik} = \mathsf{E}[z_{ik}|\mathbf{y}, \mathbf{x}]$.

Для нахождения оптимальной оценки вектора гиперпараметров $\hat{\boldsymbol{\theta}}$, максимизирующей $\mathsf{E}_Z[p(\mathbf{y}, Z|\mathbf{x}, \boldsymbol{\theta})]$ предлагается использовать итеративный двухшаговый алгоритм Expectation-Maximization (п. 4.3).

4.2 Смесь экспертов.

Определение: *Смесь экспертов* — мультимодель, определяющая правдоподобие каждой π_k каждой модели $\mathbf{f}_k$ на объекте $\mathbf{x}$ на основе его признакового описания.

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

$$\pi_k(\mathbf{x}, \mathbf{V}) : \mathbb{R}^m \to [0; 1] \quad \forall k = 1 \dots K.$$

Теперь будем рассматривать $\boldsymbol{\pi}$ как случайный вектор. Тогда каждая модель $\mathbf{f}_k$, входящая в состав мультимодели порождает пару объект-ответ $(\mathbf{x}, \mathbf{y})$ с вероятностью $p(k|\mathbf{x}, \mathbf{w}_k)$. Тогда распределение вектора ответов $\mathbf{y}$ может быть представлено в виде:

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

$$= p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}, \mathbf{v}_k) \mathcal{N}(\mathbf{y}|\mathbf{f}(\mathbf{x}, \mathbf{w}_k), \beta_k), \tag{7}$$

где шлюзовая функция может быть представлена в виде softmax:

$$\pi_k(\mathbf{x}, \mathbf{v}_k) = \boldsymbol{\sigma}(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}) = \frac{\exp(\mathbf{v}_k^{\mathrm{T}}\mathbf{x})}{\sum_{k'=1}^{K} \exp(\mathbf{v}_{k'}^{\mathrm{T}}\mathbf{x})}.$$

Шлюзовая функция может рассматриваться как классификатор, где в качестве классов выступают модели $\mathbf{f}_k$, входящие в мультимодель. В общем случае, она может представлять собой сложную функцию.

Обозначим вектор гиперпараметров $\boldsymbol{\theta}$

$$\boldsymbol{\theta} = [\mathbf{w}_1, \dots, \mathbf{w}_K, \mathbf{V}, \boldsymbol{\beta}],$$

содержащий теперь вектор параметров $\mathbf{V}$ шлюзовой функции. Тогда распределение на вектор ответов $\mathbf{y}$ можно расписать в следующем виде:

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

$$= \text{/для среднеквадратичной ошибки/} =$$

$$= \sum_{k=1}^{K} \frac{\exp(\mathbf{v}_k^{\mathrm{T}}\mathbf{x})}{\sum_{k'=1}^{K}\exp(\mathbf{v}_{k'}^{\mathrm{T}}\mathbf{x})} \exp\left(-\frac{1}{2\beta_k}\big(\mathbf{y}-\mathbf{f}_k(\mathbf{x},\mathbf{w}_k)\big)^2\right).$$

Введем матрицу $\mathbf{\Gamma} = [\gamma_{ik}]$, где за γ_{ik} обозначено правдоподобие модели $\mathbf{f}_k$ на объекте $\mathbf{x}_i$. Строки матрицы $\mathbf{\Gamma}$ содержат значения шлюзовой функции $\boldsymbol{\pi}$ на объектах выборки.

4.3 EM-алгоритм.

Для оптимизации вектора гиперпараметров $\boldsymbol{\theta}$ предлагается использовать двухшаговый алгоритм Expectation-Maximization, состоящий из E- и M- шагов соответственно.

E-шаг: Используя текущие оценки $\mathbf{w}_1^s, \ldots, \mathbf{w}_K^s, \mathbf{V}^s, \boldsymbol{\beta}^s$ пересчитать матрицу

$$\mathbf{\Gamma}^{(s+1)} = [\pi_1(\mathbf{X}), \ldots, \pi_K(\mathbf{X})]$$

следующим образом:

$$\gamma_{ik}^{(s+1)} = \mathsf{E}(z_{ik}) = p(k|\mathbf{x}_i, \boldsymbol{\theta}^{(s)}) = \tag{8}$$

$$= \frac{\pi_k(\mathbf{x}_i)\mathcal{N}(y_i|\mathbf{f}_k(\mathbf{x}_i,\mathbf{w}_k^{(s)}),\beta_k^{(s)})}{\sum_{k'=1}^{K}\pi_{k'}(\mathbf{x}_i)\mathcal{N}(\mathbf{y}_i|\mathbf{f}_k(\mathbf{x}_i,\mathbf{w}_{k'}^{(s)}),\beta_k^{(s)}),}$$

где s — номер итерации. **M-шаг:** Используя новую оценку значений γ_{ik} матрицы $\mathbf{\Gamma}^{(s+1)}$ оптимизировать параметры моделей $\mathbf{f}_k$, входящих в смесь:

$$\mathbf{v}_k = \operatorname{argmax}_{\mathbf{v}} \sum_{i=1}^{n} \gamma_{ik}^{s+1} \ln \pi_k(\mathbf{x}_i, \mathbf{v}),$$

$$\mathbf{w}_k = \operatorname{argmax}_{\mathbf{w}_k} \left[-\sum_{i=1}^{m} \gamma_{ik}^{s+1} \left(\mathbf{y}_i - \mathbf{f}_k(\mathbf{x}_i, \mathbf{w}_k)\right)^2\right],$$

$$\beta_k = \operatorname{argmax}_{\beta} \left[n\ln\beta - \sum_{i=1}^{m} \frac{1}{\beta} \left(\mathbf{y}_i - \mathbf{f}_k(\mathbf{x}_i, \mathbf{w}_k)\right)^2\right].$$

Начальная инициализация параметров моделей $\mathbf{f}_k$ и шлюзовой функции играют важную роль в сходимости данного метода. Для качественной инициализации предлагается использовать априорные знания о природе данных или множественные запуски на различных подмножествах данных.

4.4 Отбор моделей с помощью шлюзовой функции.

Определение: Будем называть модель *незначимой*, если ее правдоподобие близко к нулю на всех объектах обучающей выборки.

Теорема: Незначимая модель $\mathbf{f}_{k_{\text{weak}}}$ может быть исключена из мультимодели $\bar{\mathbf{f}}$ без потери в качестве описания данных мультимоделью.

Доказательство: Правдоподобие вектора ответов $\mathbf{y}$ представимо в виде:

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

$$= \sum_{k'\neq k_{\text{weak}}} p(k'|\mathbf{x},\boldsymbol{\theta})p(\mathbf{y}|k',\mathbf{x},\boldsymbol{\theta}) + p(k_{\text{weak}}|\mathbf{x},\boldsymbol{\theta})p(\mathbf{y}|k_{\text{weak}},\mathbf{x},\boldsymbol{\theta}) \approx$$

$$\approx \sum_{k'\neq k_{\text{weak}}} p(k'|\mathbf{x},\boldsymbol{\theta})p(\mathbf{y}|k',\mathbf{x},\boldsymbol{\theta}) + 0\cdot p(\mathbf{y}|k_{\text{weak}},\mathbf{x},\boldsymbol{\theta}) = \sum_{k\neq k_{\text{weak}}} p(\mathbf{y},k|\mathbf{x},\boldsymbol{\theta}).$$

Следовательно, правдоподобие вектора ответов $\mathbf{y}$ не изменится при исключении модели $\mathbf{f}_{k_{\text{weak}}}$ из мультимодели, а значит качество описания данных не пострадает. Теорема доказана.

На Рис. 1 приведена иллюстрация процедуры отбора моделей. В качестве $\mathbf{f}_k, \quad k = 1\ldots5$ используются линейные модели. Данные представляют собой ломаную из 4 сегментов и нормально распределенный шум. На верхнем графике приведены результаты описания данных мультимоделью $\bar{\mathbf{f}}$, являющейся смесью экспертов. На нижнем графике приведены правдоподобия (9) каждой из моделей в признаковом пространстве. Пятая модель $\mathbf{f}_5$ имеет близкое к нулю правдоподобие на всем рассматриваемом пространстве объектов, а значит, может быть исключена из мультимодели.

В качестве шлюзовой функции используется нейроная сеть с одним скрытым слоем из 50 нейронов:

$$\boldsymbol{\pi}(\mathbf{x},\mathbf{V}) = \frac{\exp(\mathbf{a}(\mathbf{x}))}{\sum_j \exp(\mathbf{a}_j(\mathbf{x}))}, \qquad \mathbf{a}(\mathbf{x}) = \mathbf{W}_2^{\mathrm{T}}\mathbf{tanh}(\mathbf{W}_1^{\mathrm{T}}\mathbf{x}), \tag{9}$$

$$\mathbf{h}_k(\mathbf{x}) = \boldsymbol{\sigma}(\mathbf{W}_k\mathbf{x} + \mathbf{b}_k),$$

$$\mathbf{V} = [\mathbf{W}_1,\ldots\mathbf{W}_k, \mathbf{b}_1,\ldots,\mathbf{b}_k, \mathbf{W}_{\text{hid}_1}, \mathbf{W}_{\text{hid}_2}].$$

Формально, процедура отбора моделей может быть записана следующим образом:

1) Инициализировать модели $\mathbf{f}_k, \quad k = 1, \ldots, K$ с учетом априорных знаний или на случайных подмножествах обучающей выборки.

2) Оптимизировать гиперпараметры мультимодели с помощью EM-алгоритма (Sec. 4.3).

3) Исключить незначимые модели и построить новую мультимодель без их участия.

В случае отсутствия априорной информации данную процедуру следует проводить многократно для нахождения оптимальной структуры мультимодели.

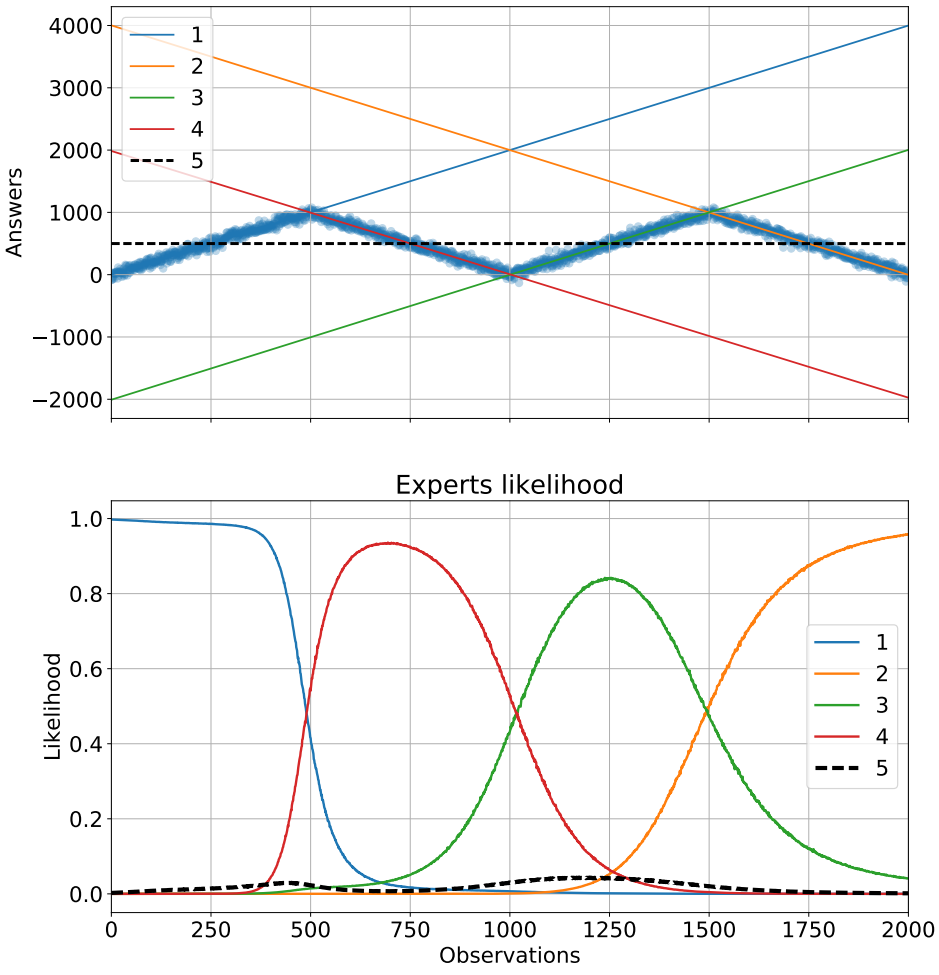

Рисунок 1: Пять моделей в составе мультимодели на синтетических данных. Пятая модель незначима и можешь быть исключена.

ПРИВИЛЕГИРОВАННОЕ ОБУЧЕНИЕ.

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

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

5.1 Контроль сходства (similarity control).

Данный подход был предложен В.Вапником в 2009 году [15] применительно к методу опорных векторов (SVM). Пусть для всех объектов доступно привилегированное описание $\mathbf{x}^*$. Функция учителя $\mathbf{f}_t : \mathbf{x}^* \to \mathbf{y}$ использует только привилегированное описание при обучении. Данная задача сводится к поиску седловой точки лагранжиана L (подробнее в п.4.1. [16]):

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \underbrace{\frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i \mathbf{y}_i \mathbf{f}_s(\mathbf{x}_i)}_{\text{функционал ошибки для случая разделимой выборки}} +$$

$$+ \underbrace{\frac{\gamma}{2}\|\mathbf{w}^*\|^2 + \sum_{i=1}^{n} (\alpha_i + \beta_i - C)\mathbf{f}_t(\mathbf{x}_i^*)}_{\text{дополнительные переменные}},$$

где $\mathbf{f}_s$ и $\mathbf{f}_t$ — линейные модели, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ — множители Лагранжа. Привлечение учителя $\mathbf{f}_t$ позволяет решать задачу в случае линейно неразделимой выборки (аналогично введению дополнительных переменных в стандартном решении SVM). Основная идея заключается в том, что используя привилегированное признаковое описание функция учителя будет вносить коррекцию в процесс обучения ученика, что позволит лучше решить задачу. В данном случае предполагается, что функция учителя обладает меньшей сложностью $\|\mathfrak{F}_t\|_C < \|\mathfrak{F}_s\|_C$, но использует более информативное признаковое описание.

5.2 Дистилляция (distillation).

Данный подход был предложен Д.Хинтоном [17] в 2015 году применительно к задаче многоклассовой классификации, но с легкостью может быть применен к задаче декодирования.

Семейство функций учителя $\mathfrak{F}_t$ содержит более сложные функции, чем семейство функций ученика $\mathfrak{F}_s$, $|\mathfrak{F}_s|_C << |\mathfrak{F}_t|_C$. Привилегированное описание данных отсутствует, $\mathbf{x}^* = \emptyset$.

Функция учителя определяется как решение задачи минимизации:

$$\mathbf{f}_t = \arg\min_{\mathbf{f}\in\mathfrak{F}_t} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(\mathbf{y}_i, \boldsymbol{\sigma}\big(\mathbf{f}(\mathbf{x}_i)\big)\Big) + \Omega\big(\|\mathbf{f}\|\big), \tag{10}$$

где $\boldsymbol{\sigma}$ — softmax, ℓ — кросс-энтропия, $\Omega : \mathbb{R} \to \mathbb{R}$ — некоторый регуляризатор.

Обучение учителя $\mathbf{f}_t$ происходит на всей доступной выборке. Затем предсказания учителя для обучающей выборки сглаживаются. Сглаженные предсказания $\mathbf{s}_i$:

$$\mathbf{s}_i = \boldsymbol{\sigma}\big(\mathbf{f}_t(\mathbf{x}_i)/T\big), \tag{11}$$

где T — температура сглаживания. Чем выше значение T, тем ближе распределение вероятностей вектора $\mathbf{s}_i$ к равномерному (иначе говоря, тем выше его энтропия). На Рис. 2 приведено сравнение исходных и сглаженных предсказаний $\mathbf{f}_t$ для датасета MNIST [18].

Функция ученика $\mathbf{f}_s$ обучается с учетом сглаженных предсказаний учителя $\mathbf{s}_i$:

$$\mathbf{f}_s = \arg\min_{\mathbf{f}\in\mathfrak{F}_s} \frac{1}{n}\sum_{i=1}^{n} \left[(1-\lambda)\ell\Big(\mathbf{y}_i, \boldsymbol{\sigma}\big(\mathbf{f}(\mathbf{x}_i)\big)\Big) + \lambda\ell\Big(\mathbf{s}_i, \boldsymbol{\sigma}\big(\mathbf{f}(\mathbf{x}_i)\big)\Big)\right], \tag{12}$$

где $\lambda \in [0; 1]$ — параметр имитации. Варьирование этого параметра позволяет балансировать между обучением на исходные векторы ответов $\mathbf{y}_i$ и на сглаженные предсказания учителя $\mathbf{s}_i$. Выгода заключается в том, что функция учителя ошибается на сложных и аномальных примерах, что приводит к меньшей ошибке при $\lambda > 0$, а значит и меньшему влиянию сложных и некорректных примеров на функцию ученика.

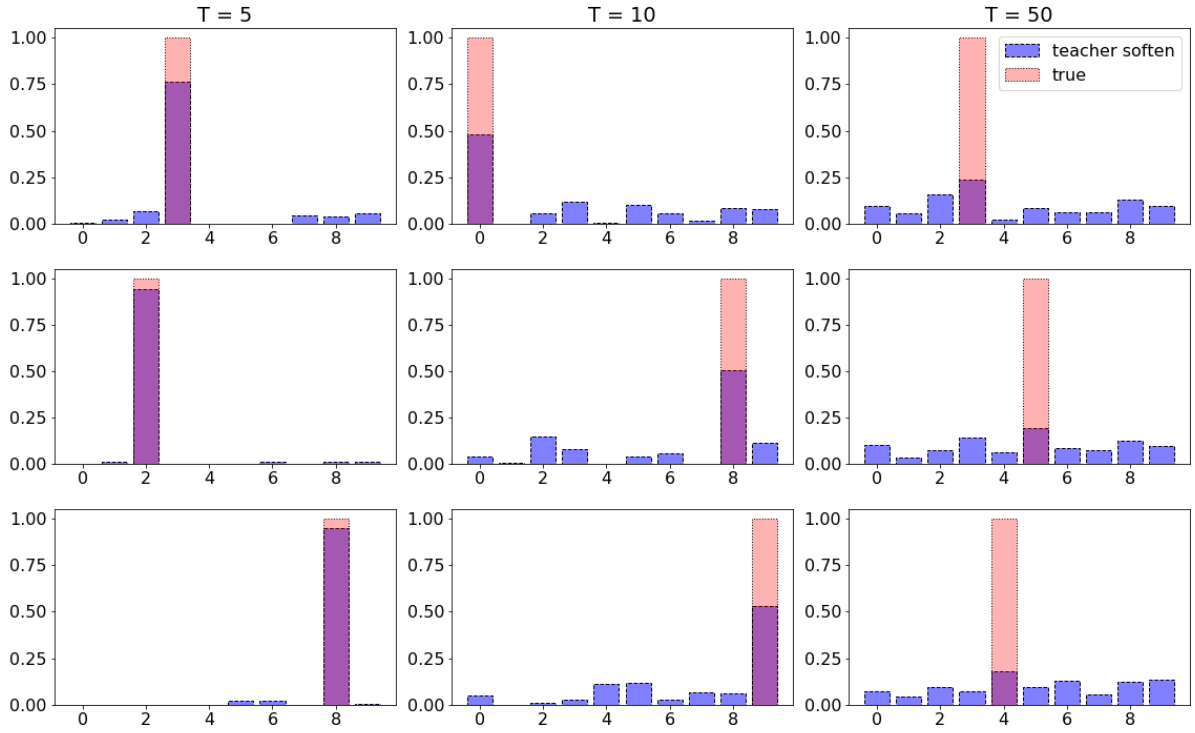

Рисунок 2: Пример сглаживания предсказанных вероятностей принадлежности классу в зависимости от температуры T.

5.3 Обобщенная дистилляция.

Данный подход объединяет предложенные выше идеи Вапника и Хинтона. Предполагается, что для некоторых объектов доступно привилегированное описание $\mathbf{x}^*$, и именно на них обучается функция учителя $\mathbf{f}_t$. Ограничения на сложность семейства $\mathfrak{F}_t$ при этом не накладывается. Это согласуется с интуитивными соображениями, согласно которым привилегированное описание может иметь гораздо более простую структуру, чем стандартное признаковое описание объектов, а значит требуются более простые модели для работы с ним. Рассмотрим данное утверждение на примере.

Пусть стандартное признаковое описание объектов — рентген-снимки внутренних органов, а привилегированное — медицинские заключения, написанные на их основе. Пространство всех возможных снимков (представляющих из себя не что иное, как набор пикселей) гораздо шире, чем пространство, состоящее из медицинских терминов, используемых при написании заключения. При этом

привилегированное пространство намного более информативно, и может быть использовано только на этапе обучения модели. В таком случае сложность семейства функций учителя $\mathfrak{F}_t$ может быть значительно ниже, чем у семейства функций ученика $\mathfrak{F}_s$, но при этом вносить существенный вклад в качество классификации.

Функция ученика находится в результате решения задачи минимизации 12. Общий алгоритм имеет вид:

1) Выбрать параметр имитации λ и температуру сглаживания T.

2) Выделить подмножество объектов, обладающих привилегированным описанием $\mathbf{x}^*$ и найти оптимальную функцию учителя $\mathbf{f}_t$ согласно 10.

3) Используя функцию учителя $\mathbf{f}_t$ построить сглаженные предсказания для всех объектов обучающей выборки с согласно 11.

4) Найти оптимальную функцию ученика согласно 12.

Данный подход включает в себя описанные выше методы переноса знаний и дистилляции. В случае

$$|\mathcal{F}_\mathrm{t}|_C \gg |\mathcal{F}_\mathrm{s}|_C, \quad \mathbf{x}^* = \emptyset$$

имеет место дистилляция, описанная Д.Хинтоном. В противоположном случае

$$|\mathcal{F}_\mathrm{t}|_C \ll |\mathcal{F}_\mathrm{s}|_C, \quad \mathbf{x}^* \neq \emptyset$$

обобщенная дистилляция переходит в метод, описанный В.Вапником.

Сформулируем общую идею, которую эксплуатирует обобщенная дистилляция: *ученик не может описать ничего из того, что не смог качественно описать учитель*. Благодаря подобному подходу ученик $\mathbf{f}_s$ "обращает" меньше внимания на сложные примеры, на которых ошибся учитель $\mathbf{f}_t$, что влечет лучшее описание более простых, стандартных объектов.

Данный подход к использованию априорной привилегированной информации и последовательному обучению интересен в первую очередь тем, что качество решения задач более простыми моделями (или же моделями, использующими менее эффективное признаковое описание) значительно повышается. Это особенно полезно в задачах, где присутствуют ограничения на сложность используемых

моделей. Вдобавок, он не требует наличия привилегированного описания у всех объектов обучающей выборки, что также играет роль в реальных задачах (т.к. частично обработанных данных зачастую гораздо больше).

ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ.

6.1 Использование априорной информации в смеси экспертов.

Используется синтетический набор данных, состоящий из зашумленной кусочно-линейной функции. Общее число точек $n = 2000$. В качестве шума используется нормально распределенная случайная величина с нулевым средним $\mu = 0$ и дисперсией $\sigma = 150$. На Рис. 3 приведена иллюстрация синтетических данных (серые точки). Отклонение оценивается с помощью среднеквадратичной ошибки.

В качестве шлюзовой функции $\boldsymbol{\pi}$ выступает нейронная сеть с одним скрытым слоем из 50 нейронов и ReLU-активациями. В роли экспертов используются линейные модели:

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

Сравнение мультимодели производится с нейронными сетями $\mathbf{f}_{\text{NN}}$, состоящими из двух скрытых слоев, содержащих 10, 25, 50 и 100 нейронов соответственно.

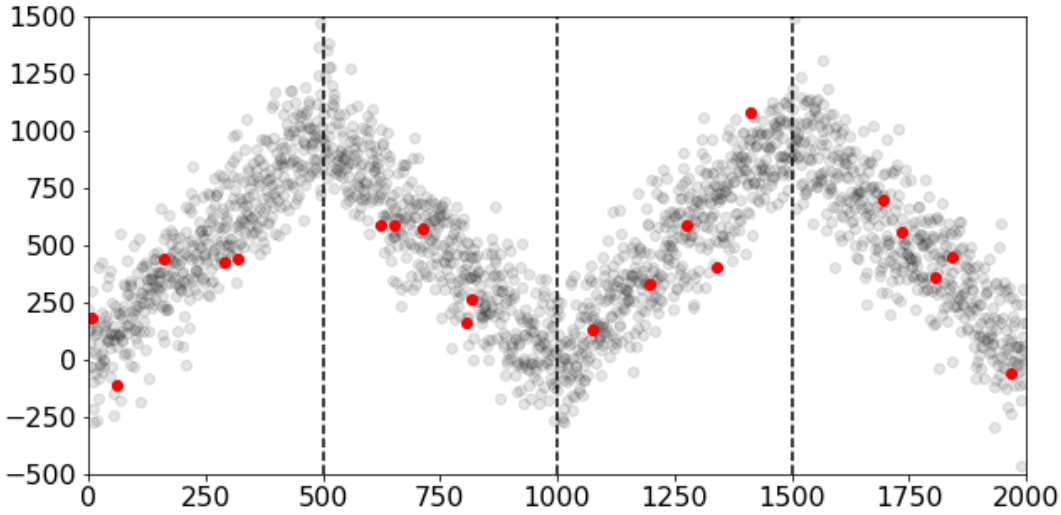

Рисунок 3: Зашумленные синтетические данные. Красным обозначены точки, для которых доступна априорная информация о принадлежности экспертам.

На Рис. 5 проиллюстрировано качество описания данных мультимоделью $\mathbf{f}_{\text{me}}$ (сверху) и правдоподобия каждого из экспертов на признаковом пространстве

(снизу). На Рис. 4 проиллюстрировано качество описания данных нейроными сетями в зависимости от размера скрытого слоя.

Как видно из иллюстрации, лишь нейронная сеть $\mathbf{f}_{\text{NN}}$ с размером скрытого слоя 100 смогла адекватно описать данные. При этом мультимодель $\mathbf{f}_{\text{me}}$ описывает данные адекватно, обладая значительно меньшим числом параметров,

$$S(\mathbf{f}_{\text{me}}) \approx S(\mathbf{f}_{\text{NN}}), \quad |\mathbf{f}_{\text{me}}|_C \ll |\mathbf{f}_{\text{NN}}|_C.$$

При этом сходимость ошибки $\mathbf{f}_{\text{me}}$ достигается лишь в 10% запусков. Без качественной начальной инициализации или же ограничений на искомый вектор гиперпараметорв использование мультимодели $\mathbf{f}_{\text{me}}$ не имеет смысла. Добавим априорную информации о принадлежности различным экспертам 20 точкам, по 5 случайных точек на каждый сегмент. Аналогичную разметку могут, например, выполнить асессоры в задаче нахождения точек смены режима временного ряда. На Рис. 3 точки, для которых доступна априорная информация, обозначены красным.

С использованием априорной информации о принадлежности некоторых точек различным экспертам сходимость функции ошибки достигается уже в 76% случаев, что уже является качественным результатом.

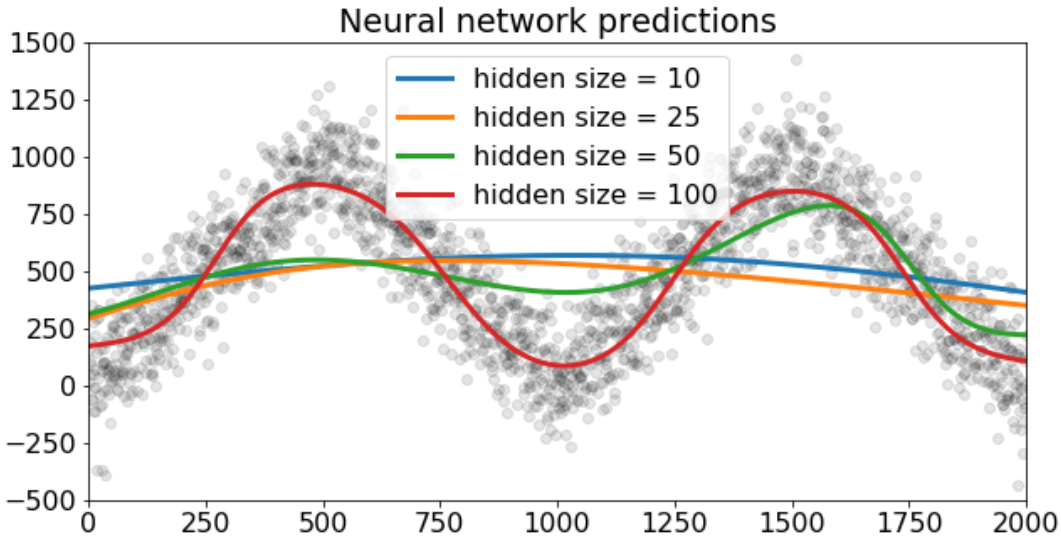

Рисунок 4: Описание зашумленных синтетических данных с помощью нейронных сетей в зависимости от размера скрытого слоя.

6.2 Дистилляция в задаче классификации изображений.

В данном эксперименте метод обобщенной дистилляции используется для повышения качества классификации простой модели $\mathbf{f}_s \in \mathfrak{F}_s$ за счет использования сложной модели $\mathbf{f}_t \in \mathfrak{F}_t$ на этапе обучения. Используются 500 изображений из датасета MNIST [18]. В качестве привилегированного признакового описания $\mathbf{x}^*$ используются стандартные изображения размера 28×28, в качестве стандартного описания изображения меньшего разрешения размерами 7×7. В качестве учителя и учения $\mathbf{f}_t$ выступают нейронные сети с двумя скрытыми слоями различной сложности: общее число параметров учителя $|\mathbf{f}_t|_C \approx 1.5 \cdot 10^4$, а ученика $|\mathbf{f}_s|_C \approx 1.5 \cdot 10^3$.

На Рис. 6 приведены результаты классификации изображений датасета MNIST в зависимости от значения параметра имитации λ и температуры сглаживания T. Видно, что обучение на сглаженных предсказаниях учителя благосклонно сказывается на качестве классификации ученика.

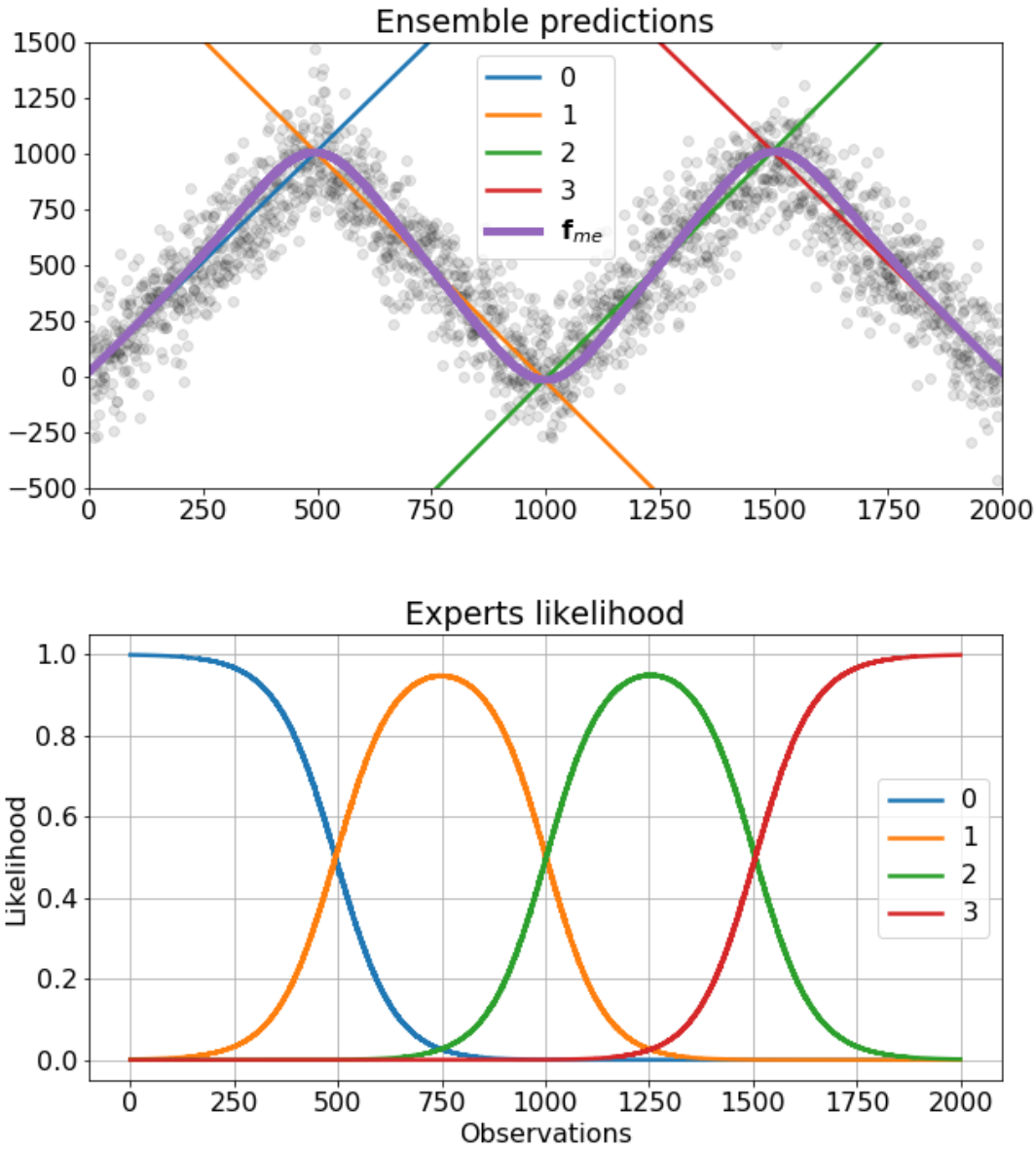

Рисунок 5: Описание зашумленных синтетических данных смесью экспертов.

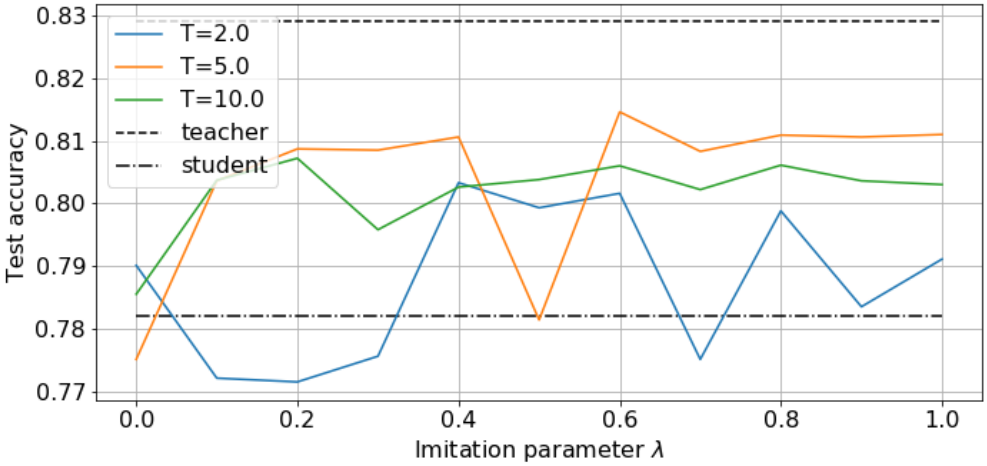

Рисунок 6: Точность классификации учителем $\mathbf{f}_t$ и учеником $\mathbf{f}_s$ датасета MNIST в зависимости от параметров дистилляции T и λ

.

ЗАКЛЮЧЕНИЕ

В данной работе предлагается несколько подходов к построению моделей оптимальной сложности. Использование мультимоделей предполагает разделение пространства объектов на подобласти,в каждой из которых данные описываются определенной моделью или их композицией. Привлечение априорной информации на этапе построения мультимодели приводит к значительным улучшениям в сходимости параметров мультимодели к оптимальным значениям. Мета-обучение предполагает использовать парадигму учителя-и-ученика, где на этапе обучения модели-ученика используются как истинные ответы на соответствующих объектах, так и предсказания модели-учителя. Использование привилегированной информации в мета-обучении повышает итоговое качество предсказаний модели и позволяет снижать ее сложность без значимых потерь в качестве предсказаний. Результаты данной работы были использованы при разработке системы автоматического прогнозирования энергопотребления датацентров компании Яндекс.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Comparison of correlation analysis techiques for irregularly sampled time series / K. Rehfeld, N. Marwan, J. Heitzig [и др.] // Nonlinear Processes in Geophysics. 2011. Т. 18. С. 389–404.

[2] Multi-dimensional function approximation and regression estimation. / F. P erez Cruz, G. Camps-Valls, E. Soria-Olivas [и др.] // Artificial Neural Networks. 2002. С. 757–762.

[3] Fung David S. Methods for the Estimation of Missing Values in TIme Series // Edith Cowan University Thesis. 2006.

[4] Outrageously large neural networks: the sparsely-gated mixture-of-experts layer / Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz [и др.].

[5] Barret Zoph Quoc Le. Neural Architecture Search with Reinforcement Learning // ICLR. 2017.

[6] Chen Tianqi, Guestrin Carlos. XGBoost: A Scalable Tree Boosting System // KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[7] Chen Xi, Ishwaran Hemant. Random Forests for Genomic Data Analysis // Genomics. 2012. Т. 99, № 6. С. 323–329.

[8] Yuksel Seniha Esen, Wilson Joseph N., Gader Paul D. Twenty Years of Mixture of Experts // IEEE Transactions on Neural Networks and Learning Systems. 2012. Т. 23, № 8. С. 1177–1193.

[9] Rasmussen Carl Edward, Ghahramani Zoubin. Infinite Mixtures of Gaussian Process Experts // Advances in Neural Information Processing Systems 14. 2002. С. 881–888.

[10] Scarpel Rodrigo Arnaldo. An integrated mixture of local experts model for demand forecasting // International Journal of Production Economics. 2015. T. 164, № C. C. 35–42.

[11] Chamroukhi F. Skew-normal Mixture of Experts // 2016 International Joint Conference on Neural Networks (IJCNN). 2016. July. C. 3000–3007.

[12] O.Bakhteev M.Popova V.Strijov. Systems and means of deep learning for classification problems. // Sistemy i Sredstva Inform. 2016. T. 26. C. 4–22.

[13] David Lopez-Paz Léon Bottou Bernhard Schölkopf Vladimir Vapnik. Unifying distillation and privileged information. 2016.

[14] Aduenko Alexander A., Motrenko Anastasia P., Strijov Vadim V. Object selection in credit scoring using covariance matrix of parameters estimations // Annals of Operations Research. 2018. January. T. 260, № 1. C. 3–21.

[15] Vapnik V., Vashist A. A new learning paradigm: Learning using privileged information. // Neural Networks. 2009.

[16] V.Vapnik R.Izmailov. Learning with Intelligent Teacher: Similarity Control and Knowledge Transfer. // Journal of Machine Learning Research. 2015. T. 16. C. 2023–2049.

[17] Hinton Geoffrey E., Vinyals Oriol, Dean Jeffrey. Distilling the Knowledge in a Neural Network // CoRR. 2015. T. abs/1503.02531. URL: http://arxiv.org/abs/1503.02531.

[18] LeCun Yann, Cortes Corinna. MNIST handwritten digit database. 2010. URL: http://yann.lecun.com/exdb/mnist/.