Задача обучения с экспертом для построение интерпретируемых моделей машинного обучения

Грабовой Андрей Валериевич

Московский физико-технический институт

МФТИ, г. Москва

Вероятностная интерпретация дистилляции моделей

Цель
Предложить постановку задачи обучения с *экспертной информацией* для обучения интерпретируемых моделей машинного обучения.

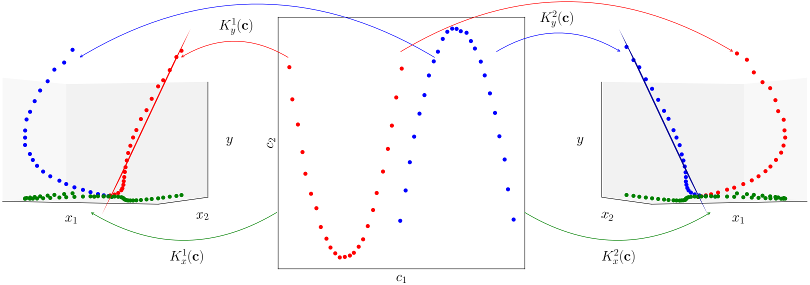
Задачи
1. Поставить задачу обучения с экспертной информацией.
2. Предложить метод решения предложенной задачи.
3. Провести анализ предложенного метода для задачи аппроксимации кривых второго порядка.

Исследуемая проблема
Построение интерпретируемых моделей глубокого обучения.

Список литературы

1. *Грабовой А.В., Стрижов В.В.* Анализ свойств вероятностных моделей обучения с экспертом // в процессе подачи.
2. *Грабовой А.В., Стрижов В.В.* Анализ выбора априорного распределения для смеси экспертов // Журнал Вычислительной математики и математической физики, 2021. Т. 61. № 5.
3. *Yuksel Seniha Esen, Wilson Joseph N., Gader Paul D* Twenty Years of Mixture of Experts // IEEE Transactions on Neural Networks and Learning Systems, 2012. Vol. 23. No 8. Pp. 1177–1193.

Обучение с экспертной информацией



(a) (b) (c)

1. Учитель $\mathbf{f}$ влияет на выбор ученика $\mathbf{g}$ в пространстве x.
2. Учитель $\mathbf{f}_1$ корректирует шумные данные в x.
3. Модель учителя $\mathbf{f}_2$ более сложная, поэтому она аппроксимирует также и шум.

Постановка задачи: кривые второго порядка

Изображение
$$\mathbf{M} \in \{0, 1\}^{m_1 \times m_2},$$
где 1 — точка изображение, а 0 — точка фона.
Точки изображения — кривая второго порядка Ω.
Координаты точек изображения $\mathbf{C} \in \mathbb{R}^{N \times 2}$.
Экспертная информация о фигуре Ω обозначается $E\bigl(\Omega\bigr)$.
Построение новой задачи:
$$K_x\bigl(E\bigl(\Omega\bigr)\bigr) : \mathbb{R}^2 \to \mathbb{R}^n, \quad K_y\bigl(E\bigl(\Omega\bigr)\bigr) : \mathbb{R}^2 \to \mathbb{R},$$
где K_x, K_y отображения объектов в признаковое описание и пространство ответов.
Выборка для аппроксимации:
$$\mathfrak{D} = \bigl\{(\mathbf{x}, y) \mid \forall \mathbf{c} \in \mathbf{C}\ \mathbf{x} = K_x(\mathbf{c}),\ y = K_y(\mathbf{c})\bigr\}.$$
В данной работе предполагается, что выборка $\mathfrak{D}$ аппроксимируется линейной моделью:
$$g(\mathbf{x}, \mathbf{w}) = \mathbf{x}^{\mathsf{T}}\mathbf{w},$$
где $\mathbf{w}$ вектор, параметр, который требуется найти.
Требуется решить следующую оптимизационную задачу:
$$\hat{\mathbf{w}} = \arg\min \sum \|g(\mathbf{x}, \mathbf{w}) - y\|_2^2.$$

Постановка задачи: признаковое описание кривых

Произвольная кривая второго порядка, главная ось которой не параллельна оси ординат, задается следующим выражением:

$$x^2 = B'xy + C'y^2 + D'x + E'y + F',$$

где на коэффициенты B', C' накладываются ограничения, которые зависят от вида кривой.
Также получаем:

$$K_x\left(\mathbf{c}_i\right) = \left[x_i y_i, y_i^2, x_i, y_i, 1\right], \quad K_y\left(\mathbf{c}_i\right) = x_i^2,$$

откуда получаем задачу линейной регрессии для восстановления параметров:

$$B', C', D', E', F'$$

по выборке $\mathfrak{D}$.

Постановка задачи нахождения параметров окружностей

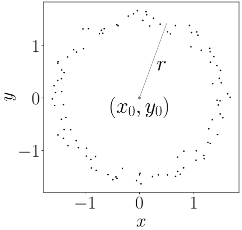
Задано бинарное изображение:

$$\mathbf{M} \in \{0, 1\}^{m_1 \times m_2},$$

где 1 — черная точка, 0 — белая точка фона.
По изображению $\mathbf{M}$ строится выборка $\mathbf{C}$:

$$\mathbf{C} \in \mathbb{R}^{N \times 2},$$

где N — число черных точек на изображении $\mathbf{M}$.



Пусть x_0, y_0 — центр окружности, которую требуется найти, а r ее радиус.
Точки $(x_i, y_i) \in \mathbf{C}$ должны удовлетворять уравнению окружности:

$$(x_i - x_0)^2 + (y_i - y_0)^2 = r^2 \Rightarrow (2x_0) \cdot x_i + (2y_0) \cdot y_i + (r^2 - x_0^2 - y_0^2) \cdot 1 = x_i^2 + y_i^2.$$

Задачу линейной регрессии для нахождения окружности:

$$\mathbf{X}\mathbf{w} \approx \mathbf{y}, \quad \mathbf{X} = [\mathbf{C}, \mathbf{1}], \quad \mathbf{y} = [x_1^2 + y_1^2, x_2^2 + y_2^2, \cdots, x_N^2 + y_N^2]^\mathsf{T},$$

где найденые оптимальные параметры линейной регрессии $\mathbf{w} = \bigl[w_1, w_2, w_3\bigr]^\mathsf{T}$ восстанавливают параметры окружности:

$$x_0 = \frac{w_1}{2}, \quad y_0 = \frac{w_2}{2}, \quad r = \sqrt{w_3 + x_0^2 + y_0^2}.$$

Постановка задачи: мультимодель

Заданы K кривых второго порядка $\Omega_1, \cdots, \Omega_K$ с $E_k = E\bigl(\Omega_k\bigr)$.

Definition

Функция f называется смесью K экспертов, если:

$$f = \sum_{k=1}^{K} \pi_k(\mathbf{x}, \mathbf{V}) g_k(\mathbf{w}_k), \quad \pi_k(\mathbf{x}, \mathbf{V}) : \mathbb{R}^{n \times |\mathbf{V}|} \to [0,\, 1], \quad \sum_{k=1}^{K} \pi_k(\mathbf{x}, \mathbf{V}) = 1,$$

где g_k — локальная модель, $\mathbf{x}$ — признаки, π_k — шлюзовая функция, $\mathbf{w}_k$ — параметры, $\mathbf{V}$ — параметры шлюзовой функции.

Мультимодель, описывающую кривые $\Omega_1, \ldots, \Omega_K$ на изображении $\mathbf{M}$:

$$f = \sum_{\mathbf{c} \in \mathbf{C}} \sum_{k=1}^{K} \pi_k(\mathbf{c}, \mathbf{V}) g_k(K_x^k(\mathbf{c}), \mathbf{w}_k), \quad \mathbf{x} = K_x^1(\mathbf{c}) = \cdots = K_x^K(\mathbf{c}).$$

Решается задача оптимизации:

$$\mathcal{L} = \sum_{(\mathbf{x}, y) \in \mathfrak{D}} \sum_{k=1}^{K} \pi_k(\mathbf{x}, \mathbf{V}) (y - \mathbf{w}_k^{\mathsf{T}} \mathbf{x})^2 + R\bigl(\mathbf{V}, \mathbf{W}, E(\Omega)\bigr) \to \min_{\mathbf{V}, \mathbf{W}},$$

где $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$ — параметры локальных моделей, $R\bigl(\mathbf{V}, \mathbf{W}, E(\Omega)\bigr)$ — регуляризация параметров, основанная на экспертной информации.

Вероятностная задача

Рассматривается вероятностная постановка задачи:

1) правдоподобие выборки $p_k\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right) = \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)$, где β уровень шума,
2) априорное распределение параметров $p^k\left(\mathbf{w}_k\right) = \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right)$, где $\mathbf{w}_k^0$ — вектор размера $n \times 1$, $\mathbf{A}_k$ — ковариационная матрица параметров,
3) регуляризация априорного распределения $p\left(\varepsilon_{k,k'}|\boldsymbol{\alpha}\right) = \mathcal{N}\left(\varepsilon_{k,k'}|\mathbf{0}, \Xi\right)$, где Ξ — ковариационная матрица общего вида, $\varepsilon_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0$.

Правдоподобие модели включает правдоподобие выборки, априорное распределение параметров, а также их регуляризацию

$$p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta\bigr) = \prod_{k,k'=1}^{K} \mathcal{N}\left(\varepsilon_{k,k'}|\mathbf{0}, \Xi\right) \cdot$$
$$\cdot \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)\right),$$

где $\mathbf{A} = \bigl\{\mathbf{A}_1, \cdots, \mathbf{A}_K\bigr\}$.

EM–алгоритм решения задачи

Введем скрытые переменные $\mathbf{Z} = [z_{ik}]$, где $z_{ik} = 1$ тогда и только тогда, когда $k_i = k$:

$$\log p\bigl(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr) =$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] +$$

$$+ \sum_{k=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] +$$

$$+ \sum_{k=1}^{K}\sum_{k'=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}}\hat{\boldsymbol{\alpha}}^{-1}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2}\log\det\boldsymbol{\Xi} - \frac{n}{2}\log 2\pi\right]$$

Задача оптимизации параметров локальных моделей и параметров смеси принимает следующий вид:

$$\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta = \arg\max_{\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta} \log p\bigl(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr).$$

Для оптимизации используется вариационный EM–алгоритм с предположением $q\left(\mathbf{Z}, \mathbf{W}\right) = q\left(\mathbf{Z}\right)q\left(\mathbf{W}\right)$.

EM–алгоритм для решения задачи смеси экспертов

Итерационные формулы EM–алгоритма:

1. E–шаг:

$$p\left(z_{ik}=1\right)=\frac{\exp\left(\log\pi_k\left(\mathbf{x}_i,\mathbf{V}\right)-\frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i-\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\right)\right)}{\sum_{k'=1}^{K}\exp\left(\log\pi_{k'}\left(\mathbf{x}_i,\mathbf{V}\right)-\frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_{k'}\mathbf{w}_{k'}^{\mathsf{T}}\mathbf{x}_i-\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_{k'}\right)\right)},$$

$$q(\mathbf{w}_k)=\mathcal{N}(\mathbf{w}_k|\mathbf{m}_k,\mathbf{B}_k),$$

$$\mathbf{m}_k=\mathbf{B}_k\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0+\beta\sum_{i=1}^{N}\mathbf{x}_iy_i\mathsf{E}z_{ik}\right),\quad \mathbf{B}_k=\left(\mathbf{A}_k^{-1}+\beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}z_{ik}\right)^{-1}.$$

2. M–шаг:

$$\mathbf{A}_k=\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}-\mathbf{w}_k^0\mathsf{E}\mathbf{w}_k^{\mathsf{T}}-\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{0\mathsf{T}}+\mathbf{w}_k^0\mathbf{w}_k^{0\mathsf{T}},$$

$$\frac{1}{\beta}=\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[y_i^2-2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k+\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right]\mathsf{E}z_{ik},$$

$$\mathbf{w}_k^0=\left[\mathbf{A}_k^{-1}+(K-1)\,\boldsymbol{\Xi}\right]^{-1}\left(\mathbf{A}_k^{-1}\mathsf{E}\mathbf{w}_k+\boldsymbol{\Xi}\sum_{k'=1,\ k'\neq k}^{K}\mathbf{w}_{k'}^0\right),$$

$$\mathbf{V}=\arg\max_{\mathbf{V}}\mathsf{E}_{q^s}\log p\big(\mathbf{y},\mathbf{Z},\mathbf{W}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\boldsymbol{\Xi},\beta\big).$$

Вычислительный эксперимент

Эксперимент с окружностями:
Информация

Эксперимент с разным уровнем шума в данных:
Информация

Аппроксимация радужки глаза
Информация

Эксперимент с окружностями

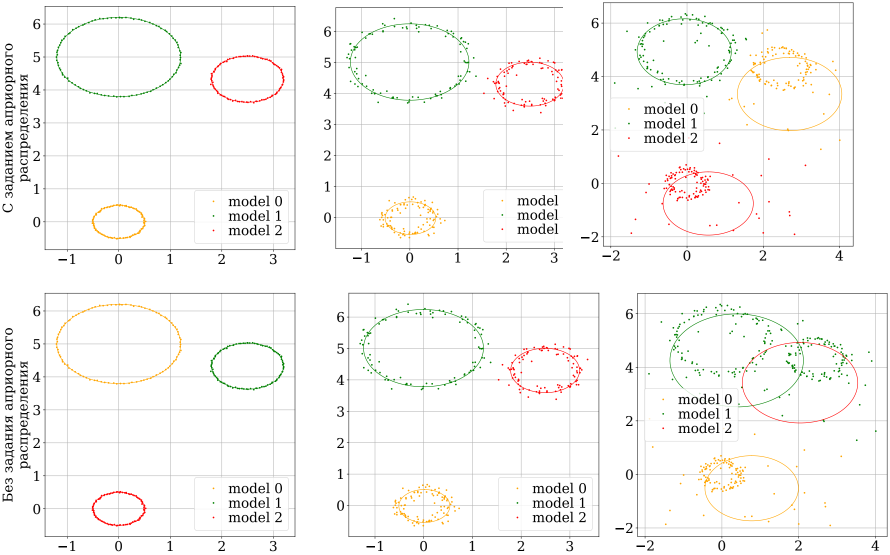


Рис.: Мультимодель в зависимости от разных априорных предположений и уровня шума. Сверху вниз: построение с заданием априорного распределения; без задания

Эксперимент с разным уровнем шума

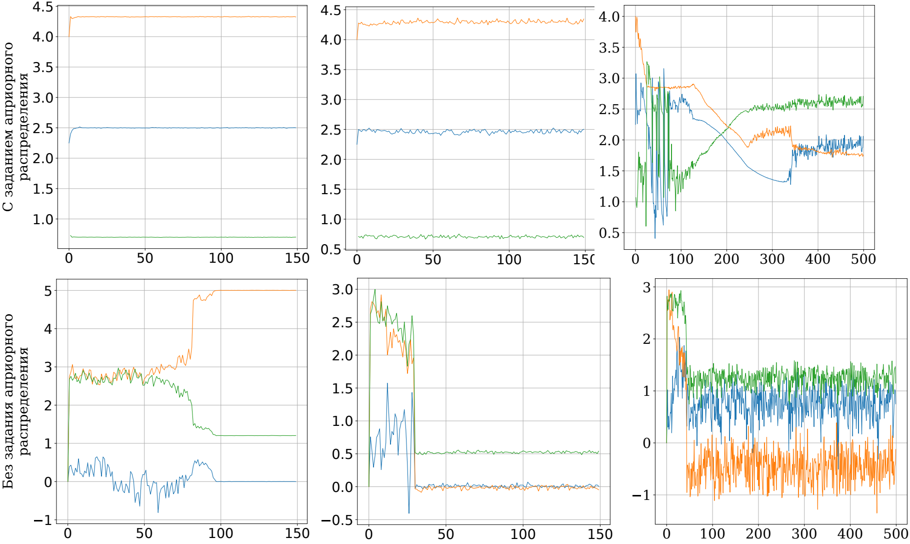


Рис.: Зависимость параметров r, x_0 и y_0 от номера итерации при разных априорных распределениях. Сверху вниз: построение с заданием априорного распределения; без задания априорного распределения. Слева на право: окружности без шума; шум в

Эксперимент с разным уровнем шума

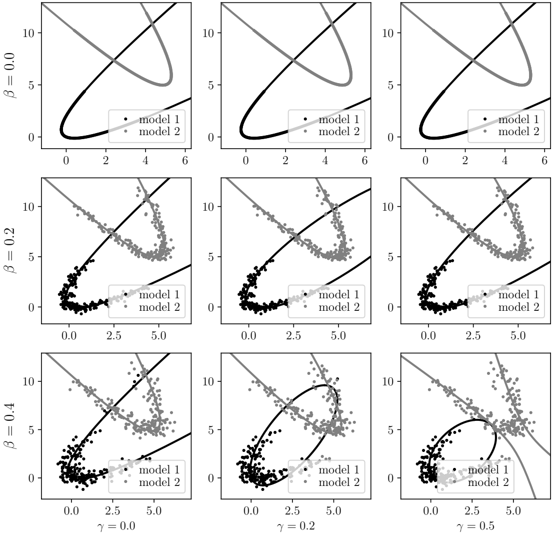


Рис.: Результат аппроксимации для данных с разным уровнем шума β и от дисперсии априорного распределения γ

Эксперимент с разным уровнем шума

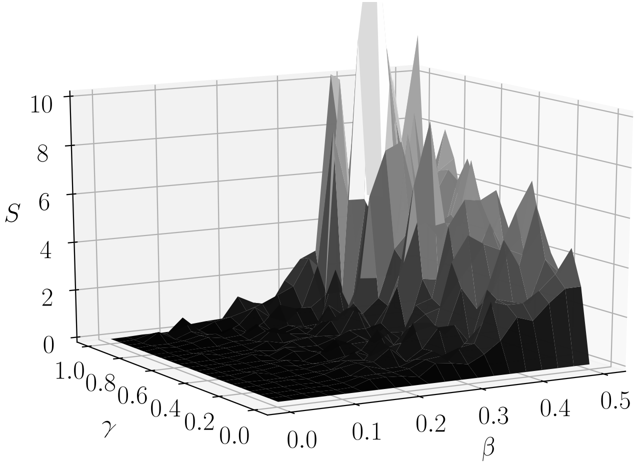
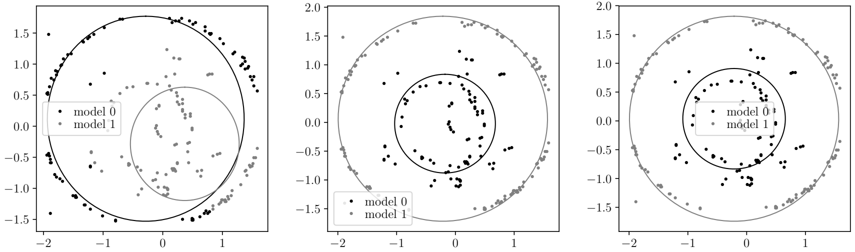


Рис.: Результат аппроксимации для данных с разным уровнем шума β и от дисперсии априорного распределения γ

Эксперимент с аппроксимаций радужки глаза



(a) (b) (c)

Рис.: Визуализация аппроксимации радужки глаза: a) в случае, если задан регуляризатор R_0; b) в случае, если задан регуляризатор R_1; b) в случае, если задан регуляризатор R_2.

Заключение

1. Поставлена задача обучения с экспертной информацией.
2. Предложен метод решения задачи обучения с экспертной информацией.
3. Приведен частный случай обучения с экспертной информацией для решения задачи поиска кривых второго порядка.
4. Проведен вычислительный эксперимент для анализа предложенной модели.

Планируется:

1. Провести адаптация предложенного метода для методов глубокого обучения.

Публикации ВАК по теме

1. *Грабовой А.В., Бахтеев О.Ю., Стрижов В.В.* Определение релевантности параметров нейросети // Информатика и ее применения, 2019, 13(2).
2. *Грабовой А.В., Бахтеев О. Ю., Стрижов В.В.* Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения, 2020, 14(2).
3. *A. Grabovoy, V. Strijov.* Quasi-periodic time series clustering for human. Lobachevskii Journal of Mathematics, 2020, 41(3).
4. *Грабовой А.В., Стрижов В.В.* Анализ выбора априорного распределения для смеси экспертов // Журнал Вычислительной математики и математической физики, 2021. 61(5).
5. *Грабовой А.В., Стрижов В.В.* Анализ моделей привилегированного обучения и дистилляции // Автоматика и телемеханика, 2021 (текущая работа, на рецензировании).
6. *T. Gadaev, A. Grabovoy, A. Motrenko, V. Strijov* Numerical methods of minimum sufficient sample size estimation for linear models // in progress.
7. *Базарова А.И., Грабовой А.В., Стрижов В.В.* Анализ свойств вероятностных моделей в задачах обучения с экспертом // подано.