Анализ выбора априорного распределения для смеси экспертов *

А. В. Грабовой[1], **В. В. Стрижов**[2]

Аннотация: Данная работа посвящена анализу свойств смеси экспертов. Рассматриваются различные способы выбора априорного распределения. Анализируется случай, когда выбрано информативное и неинформативное априорные распределения параметров каждого эксперта. В качестве экспертов рассматриваются линейные модели, а в качестве их смеси рассматривается нейросеть с функцией softmax на последнем слое. В качестве базовой задачи рассматривается задача поиска окружностей на изображении. Предполагается, что каждой окружности на изображении соответствует свой эксперт. Рассматривается случай, когда априорные распределения локальных моделей являются независимы, а также случай, когда эти распределения являются зависимыми. В качестве данных рассматриваются синтетически сгенерированные окружности с разным уровнем шума. Сравнивается устойчивость к шуму смеси с заданными априорными распределениями на вектора параметров экспертов и без задания априорного распределения.

Ключевые слова: смесь экспертов; байесовский выбор модели; априорное распределение.

DOI: 00.00000/00000000000000

1 Введение

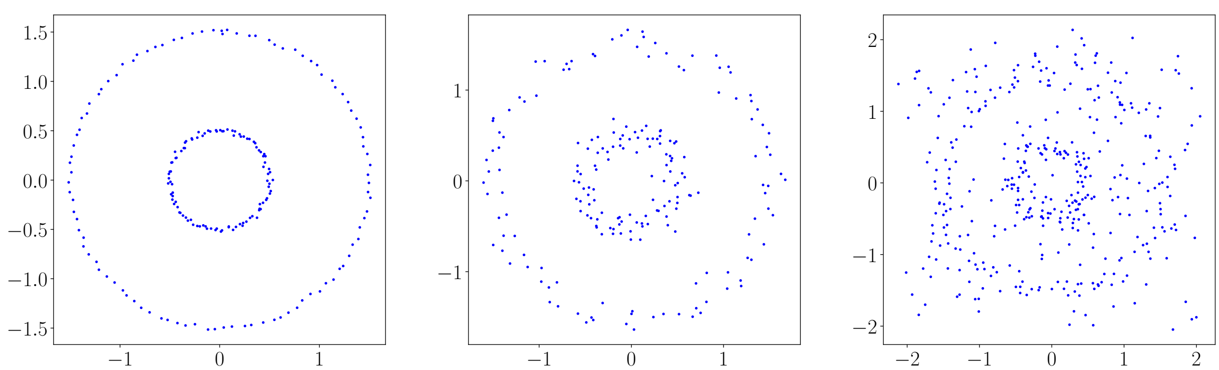
В данной работе исследуется проблема построения модели смеси экспертов. Смесь экспертов — это мультимодель, которая линейно взвешивает локальных моделей, которые аппроксимируют выборку. Значение весовых коэффициенты зависят от того объекта для которого производится предсказание.

*Работа выполнена при поддержке РФФИ и правительства РФ.
[1]Московский физико-технический институт, grabovoy.av@phystech.edu
[2]Московский физико-технический институт, strijov@ccas.ru

Примерами мультимоделей являются беггинг, градиентный бустинг [1] и случайный лес [2] решающих деревьев. Подход к мультимоделированию [3] предполагает, что вклад каждой модели в ответ зависит от рассматриваемого объекта. Смесь экспертов использует шлюзовую функцию, которая определяет значимость предсказания каждого эксперта — отдельной модели, входящей в смесь.

Для поиска оптимальных параметров смеси и локальных моделей рассматривается вероятностная постановка задачи. В качестве функционала качества рассматривается логарифм правдоподобия модели. Для оптимизации данного функционала используется EM-алгоритм [9].

Мультимодели имеют ряд недостатков, которые связаны с тем, что сходимость локальных моделей существенно зависит от их начальной инициализации. Для повышения скорости сходимости предлагается использовать априорные знания о распределении параметров и распределении весов экспертов. В данной работе задается априорное распределения на веса локальных моделей, также, для повышения качества мультимодели, предлагается использовать зависимость априорных распределений.



(a) Без шума (b) Шум вблизи окружности (c) Равномерный шум

Рис. 1: Пример изображений с окружностями с разным уровнем шума: (a) окружности без шума; (b) окружности с зашумленным радиусом; (c) окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению

Данная работа исследует зависимость качества модели в зависимости от выбора априорных распределений на веса локальных моеделей. Решается задача поиска окружностей на бинаризованном изображении. Предполагается, что радиусы окружностей различаются значимо, а также, что центры почти совпадают. Пример изображений показан на рис. 1. Предлагается рассмотреть как ведет себя модель с априорными знанием и без них в случае изображений с разным уровнем шума. В данной работе в качестве отдельных экспертов рассматриваются линейные модели — каждая модель отвечает своей окружности. В качестве шлюзовой функции рассматривается двухслойная нейронная сеть.

2 Работы по теме

Большое количество работ в области построения смеси экспертов посвящены выбору шлюзовой функции: используется softmax–регрессия, процесс Дирихле [5], нейронная сеть [4] с функцией softmax на последнем слое. Ряд работ посвящены выбору моделей в качестве отдельных экспертов. В работах [6, 10] в качестве модели эксперта рассматривается линейная модель. Работы [7, 8] рассматриваю модель SVM в качестве модели эксперта. В работа [3] представлен обзор методов и моделей в задачах смеси экспертов. В данной работе представлен обзор выше перечисленных шлюзовых функций. Также в данной работе проведен анализ разных моделей, которые могут выступать в качестве локальной модели.

3 Постановка задачи нахождения параметров окружностей

Задача аппроксимации окружности ставиться как задача линейной регрессии. Задано бинарное изображение:

$$\mathbf{M} \in \{0,1\}^{m_1 \times m_2}, \tag{3.1}$$

где 1 отвечает черной точке — изображения, 0 — белой точке фона.

По изображению $\mathbf{M}$ строится выборка $\mathbf{C}$, элементами которой являются координаты x_i, y_i черных точек на картинке:

$$\mathbf{C} \in \mathbb{R}^{N \times 2}, \tag{3.2}$$

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

Обозначим x_0, y_0 — центр окружности, которую требуется найти на бинарном изображении $\mathbf{M}$, а r ее радиус. Элементы выборки $(x_i, y_i) \in \mathbf{C}$ являются геометрическим местом точек, которое аппроксимируется уравнением окружности:

$$(x_i - x_0)^2 + (y_i - y_0)^2 = r^2. \tag{3.3}$$

Раскрыв скобки получим уравнение

$$(2x_0) \cdot x_i + (2y_0) \cdot y_i + (r^2 - x_0^2 - y_0^2) \cdot 1 = x_i^2 + y_i^2. \tag{3.4}$$

Поставим задачу линейной регрессии для нахождения параметров окружности:

$$\mathbf{X}\mathbf{w} \approx \mathbf{y}, \quad \mathbf{X} = [\mathbf{C}, \mathbf{1}], \quad \mathbf{y} = [x_1^2 + y_1^2, x_2^2 + y_2^2, \cdots, x_N^2 + y_N^2]^\mathsf{T}, \tag{3.5}$$

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

$$x_0 = \frac{w_1}{2}, \quad y_0 = \frac{w_2}{2}, \quad r = \sqrt{w_3 + x_0^2 + y_0^2}. \tag{3.6}$$

Решение уравнения (3.5) находит параметры единственной окружности на изображении. В случае, когда на изображении несколько окружностей, предлагается использовать мультимодель. В ее состав входят линейные модели. Каждая линейная модель описывает одну окружность на изображении. В качестве мультимодели рассматривается смесь экспертов. Данная постановка обобщается на поиск параметров эллипсов в приложении (A).

4 Постановка задачи построения смеси экспертов

Задана выборка из (3.5)

$$\mathbf{X} \in \mathbb{R}^{N \times n}, \tag{4.1}$$

где N — число объектов в выборке, а n — размерность признакового пространства.

Определение 4.1. *Смесь экспертов — мультимодель, определяющая правдоподобие веса π_k каждой локальной модели $\boldsymbol{f}_k$ на признаковом описании объекта $\boldsymbol{x}$.*

$$\hat{\mathbf{f}} = \sum_{k=1}^{K} \pi_k \mathbf{f}_k, \qquad \pi_k \left(\mathbf{x}, \mathbf{V}\right) : \mathbb{R}^{n \times |\mathbf{V}|} \to [0, 1], \qquad \sum_{k=1}^{K} \pi_k \left(\mathbf{x}, \mathbf{V}\right) = 1 \tag{4.2}$$

где $\hat{\mathbf{f}}$ — мультимодель, а $\mathbf{f}_k$ является некоторой моделью, π_k — шлюзовая функция, $\mathbf{w}_k$ — параметры k-й локальной модели, $\mathbf{V}$ — параметры шлюзовой функции.

В данной работе в качестве локальных моделей $\mathbf{f}_k$ и шлюзовой функции $\boldsymbol{\pi}$ рассматриваются следующие функции:

$$\mathbf{f}_k \left(\mathbf{x}\right) = \mathbf{w}_k^{\mathsf{T}} \mathbf{x}, \quad \boldsymbol{\pi} \left(\mathbf{x}, \mathbf{V}\right) = \text{softmax}\bigl(\mathbf{V}_1^{\mathsf{T}} \boldsymbol{\sigma} \left(\mathbf{V}_2^{\mathsf{T}} \mathbf{x}\right)\bigr), \tag{4.3}$$

где $\mathbf{V} = \bigl\{\mathbf{V}_1, \mathbf{V}_2\bigr\}$ — параметры шлюзовой функции.

Параметры локальных моделей оптимизируются согласно принципу максимального правдоподобия:

$$p\bigl(\mathbf{y}, \mathbf{W} | \mathbf{X}, \mathbf{V}\bigr) = \prod_{k=1}^{K} p^k\bigl(\mathbf{w}_k\bigr) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k p_k\bigl(y_i | \mathbf{w}_k, \mathbf{x}_i\bigr) \right), \tag{4.4}$$

где $\mathbf{W} = \bigl[\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K\bigr]^{\mathsf{T}}$.

Задача оптимизации параметров локальных моделей и параметров смеси:

$$\mathbf{W}, \mathbf{V} = \arg\max_{\mathbf{W}, \mathbf{V}} p\bigl(\mathbf{y}, \mathbf{W} | \mathbf{X}, \mathbf{V}\bigr). \tag{4.5}$$

5 EM–алгоритм для решения задачи смеси экспертов

Для построения смеси экспертов рассмотрим следующую вероятностную постановку 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 $p\left(\boldsymbol{\varepsilon}_{k,k'}|\boldsymbol{\alpha}\right) = \mathcal{N}\left(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\right)$, где $\boldsymbol{\Xi}$ — ковариационная матрица, $\boldsymbol{\varepsilon}_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0$.

Тогда правдоподобия модели (4.4) переписывается в следующем виде:

$$
\begin{aligned}
p\big(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) = &\prod_{k,k'=1}^{K} \mathcal{N}\left(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\right) \cdot \\
&\cdot \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)\right),
\end{aligned}
\tag{5.1}
$$

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

Для решения задачи (4.5) в предположении (5.1) введем матрицу скрытых переменных $\mathbf{Z}$, где $z_{ik} = 1$, если i-й объект порожден моделью k и $z_{ik} = 0$ иначе. Используя $\mathbf{Z}$, перепишем логарифм правдоподобия (5.1) следующим образом:

$$
\begin{aligned}
&\log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) = \\
&\quad= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\
&\quad+ \sum_{k=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}} \mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det \mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] + \\
&\quad+ \sum_{k=1}^{K}\sum_{k'=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}} \hat{\boldsymbol{\alpha}}^{-1}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2}\log\det\boldsymbol{\Xi} - \frac{n}{2}\log 2\pi\right].
\end{aligned}
\tag{5.2}
$$

С учетом (5.2) задача оптимизации (4.5) принимает вид:

$$
\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta = \arg\max_{\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta} \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big). \tag{5.3}
$$

Для поиска локального минимума в задаче оптимизации (5.3) воспользуемся вариационным EM–алгоритмом.

E–step. Найдем вариационной распределение в условиях априксимации среднего поля $q(\mathbf{Z},\mathbf{W}) = q(\mathbf{Z})\,q(\mathbf{W})$ наиболее близкое к $p\big(\mathbf{Z},\mathbf{W}|\mathbf{y},\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\boldsymbol{\Xi},\beta\big)$. Для упрощения будем искать логарифм с точностью до аддитивной константы, которую восстановим используя вид распределения.

$$
\begin{aligned}
\log q(\mathbf{Z}) &= \mathsf{E}_{q/\mathbf{Z}} \log p\big(\mathbf{y},\mathbf{Z},\mathbf{W}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\boldsymbol{\Xi},\beta\big) \propto \\
&\propto \sum_{i+1}^{N}\sum_{k=1}^{K} z_{ik}\left[\log \pi_k(\mathbf{x}_i,\mathbf{V}) - \frac{\beta}{2}\left(y_i^2 - \mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right) + \frac{1}{2}\log\frac{\beta}{2\pi}\right] \\
p(z_{ik}=1) &= \frac{\exp\left(\log \pi_k(\mathbf{x}_i,\mathbf{V}) - \frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i - \mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\right)\right)}{\sum_{k'=1}^{K}\exp\left(\log \pi_{k'}(\mathbf{x}_i,\mathbf{V}) - \frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_{k'}\mathbf{w}_{k'}^{\mathsf{T}}\mathbf{x}_i - \mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_{k'}\right)\right)}
\end{aligned}
\tag{5.4}
$$

Получаем, что распределениие $q\big(z_{ik}\big)$ является бернулевским с параметром z_{ik} из выражения (5.4).

$$
\begin{aligned}
\log q(\mathbf{W}) &= \mathsf{E}_{q/\mathbf{W}} \log p\big(\mathbf{y},\mathbf{Z},\mathbf{W}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\boldsymbol{\Xi},\beta\big) \propto \\
&\propto \sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{E}z_{ik}\left[\log \pi_k(\mathbf{x}_{i,\mathbf{V}}) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\
&+ \sum_{k=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] \\
&\propto \sum_{k=1}^{K}\left[\mathbf{w}_k^{\mathsf{T}}\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0 + \beta\sum_{i=1}^{N}\mathbf{x}_i y_i \mathsf{E}z_{ik}\right) - \frac{1}{2}\mathbf{w}_k^{\mathsf{T}}\left(\mathbf{A}_k^{-1} + \beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right)\mathbf{w}_k\right]
\end{aligned}
\tag{5.5}
$$

Из вида распределения (5.5) получаем, что распределение $q\big(\mathbf{w}_k\big) = \mathcal{N}(\mathbf{w}_k|\mathbf{m}_k,\mathbf{B}_k)$, является нормальным с параметрами $\mathbf{m}_k,\mathbf{B}_k$, которые определяются следующим образом:

$$
\mathbf{m}_k = \mathbf{B}_k\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0 + \beta\sum_{i=1}^{N}\mathbf{x}_i y_i \mathsf{E}z_{ik}\right) \qquad \mathbf{B}_k = \left(\mathbf{A}_k^{-1} + \beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}z_{ik}\right)^{-1} \tag{5.6}
$$

M–step. Найдем $\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta$ из максимизации $\mathsf{E}_q \log p\big(\mathbf{y},\mathbf{Z},\mathbf{W}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\boldsymbol{\Xi},\beta\big)$.

$$
\begin{aligned}
\mathcal{F}\left(\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta\right) &= \mathsf{E}_q \log p\big(\mathbf{y},\mathbf{Z},\mathbf{W}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\boldsymbol{\Xi},\beta\big) = \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{E}z_{ik}\left[\log \pi_k(\mathbf{x}_i,\mathbf{V}) - \frac{\beta}{2}\mathsf{E}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\
&+ \sum_{k=1}^{K}\left[-\frac{1}{2}\mathsf{E}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] + \\
&+ \sum_{k=1}^{K}\sum_{k'=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}}\boldsymbol{\Xi}^{-1}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2}\log\det\boldsymbol{\Xi} - \frac{n}{2}\log 2\pi\right].
\end{aligned}
\tag{5.7}
$$

Для нахождения оптимальных параметров $\mathbf{V}$ для, которые максимизируют функцию $\mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)$ воспользуемся градиентным методом оптимизации, который гарантирует сходимость к локальному экстремуму функции.

Оптимальное значение параметра A_k, которое максимизирует функцию $\mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)$, найдем из условия оптимума первого порядка:

$$\begin{aligned} \frac{\partial \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)}{\partial \mathbf{A}_k^{-1}} &= \frac{1}{2}\mathbf{A}_k - \frac{1}{2}\mathsf{E}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^T = 0, \\ \mathbf{A}_k &= \mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}} - \mathbf{w}_k^0\mathsf{E}\mathbf{w}_k^{\mathsf{T}} - \mathsf{E}\mathbf{w}_k\mathbf{w}_k^{0\mathsf{T}} + \mathbf{w}_k^0\mathbf{w}_k^{0\mathsf{T}}. \end{aligned} \tag{5.8}$$

Аналогично найдем оптимальные значения β и $\mathbf{w}_0^k$.

$$\begin{aligned} \frac{\partial \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)}{\partial \beta} &= \sum_{k=1}^{K}\sum_{i=1}^{N}\left(\frac{1}{\beta}\mathsf{E}z_{ik} - \frac{1}{2}\mathsf{E}z_{ik}\left[y_i^2 - 2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right]\right) = 0, \\ \frac{1}{\beta} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[y_i^2 - 2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right]\mathsf{E}z_{ik}. \end{aligned} \tag{5.9}$$

$$\begin{aligned} \frac{\partial \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)}{\partial \mathbf{w}_k^0} &= \mathbf{A}_k^{-1}\big(\mathsf{E}\mathbf{w}_k - \mathbf{w}_k^0\big) + \boldsymbol{\Xi}\sum_{k'=1}^{K}\big[\mathbf{w}_{k'}^0 - \mathbf{w}_k^0\big] = 0, \\ \mathbf{w}_k^0 &= \left[\mathbf{A}_k^{-1} + (K-1)\,\boldsymbol{\Xi}\right]^{-1}\left(\mathbf{A}_k^{-1}\mathsf{E}\mathbf{w}_k + \boldsymbol{\Xi}\sum_{k'=1,\ k'\neq k}^{K}\mathbf{w}_{k'}^0\right). \end{aligned} \tag{5.10}$$

Используя формулы (5.4-5.10) получаем итеративную процедуру, которая сходится к локальному решению (5.3). Если в списке вероятностных предположений оставить только пункт (1.) получим решение задачи оптимизации, когда не задано никаких априорных распределений на модели. В случае, когда рассматриваются пункты (1., 2.) получим задачу с заданными априорными распределениями на локальные модели. В случае, когда рассматриваются все пункты (1., 2., 3.) назовем решение с регуляризацией априорных распределений, так как в данном случае мы учитываем зависимость между локальными моделями.

6 Вычислительный эксперимент

Проводится вычислительный эксперимент для анализа качества моделей нахождения окружностей. В эксперименте рассматривается мультимодель без задания априорных распределений на параметры модели, которую обозначим $\mathfrak{M}_1$, мультимодель $\mathfrak{M}_2$ с заданным априорным распределением (6.3) на параметры локальных моделей, также рассматривается мультимодель $\mathfrak{M}_3$ с регуляризацией априорных распределений. Качество прогноза моделью окружности $\mathfrak{M}_i$ определяется функцией:

$$\mathcal{S}_{\mathfrak{M}_i} = \sum_{k=1}^{K}\big(x_0^k - x_{\mathrm{pr}}^k\big)^2 + \big(y_0^k - y_{\mathrm{pr}}^k\big)^2 + \big(r^k - r_{\mathrm{pr}}^k\big)^2, \tag{6.1}$$

где x_0^k, y_0^k, r^k — истинные значения центра и радиуса k-й окружности, $x_{\text{pr}}^k, y_{\text{pr}}^k, r_{\text{pr}}^k$ — предсказанные значения центра и радиуса k-й окружности.

Для сравнения качества моделей с разными априорными распределениями, качество модели оценивается правдоподобием модели без учета априорного распределения:

$$\log p\big(\mathbf{y}|\mathbf{W}, \mathbf{X}, \mathbf{V}, \beta\big) = \sum_{k=1}^{K}\sum_{i=1}^{N} \pi_k\big(\mathbf{x}_i, \mathbf{V}\big) \left[-\frac{\beta}{2}\left(y_i - \mathbf{w}^{\mathsf{T}}\mathbf{x}_i\right)^2 - \frac{1}{2}\log 2\pi + \frac{1}{2}\log\beta\right]. \quad (6.2)$$

Априорные распределения на параметры локальных моделей в эксперименте было задано следующим образом:

$$p^1\left(\mathbf{w}_1\right) \sim \mathcal{N}\left(\mathbf{w}_1^0, \mathbf{I}\right), \quad p^2\left(\mathbf{w}_2\right) \sim \mathcal{N}\left(\mathbf{w}_2^0, \mathbf{I}\right), \quad (6.3)$$

где $\mathbf{w}_1^0 = [0, 0, 0.1]$, $\mathbf{w}_2^0 = [0, 0, 2]$, что указывает на концентричность окружностей и на различность радиусов.

Синтетические данные с разным типом шума в изображении. Для сравнения качества работы мультимоделей $\mathfrak{M}_1, \mathfrak{M}_2, \mathfrak{M}_3$ смеси экспертов с разными начальными предположениями был проведен вычислительный эксперимент на синтетических данных.

Вычислительный эксперимент проводится на синтетических выборках, которые получена при помощи генерации двух концентрических окружностей с разным уровнем шума. Выборка Synthetic 1 — выборка без шума, Synthetic 2 — выборка с шумом вблизи окружностей, Synthetic 3 — выборка с шумом вблизи окружности, а также с равномерным шумом по всему изображению.

На рис. 2 показан случайный результаты работы мультимоделей $\mathfrak{M}_1, \mathfrak{M}_2, \mathfrak{M}_3$. На всех изображениях обе модели обучались 50 итераций EM–алгоритма. Мультимодели $\mathfrak{M}_2, \mathfrak{M}_3$ работают лучше мультимодели $\mathfrak{M}_1$, так как они восстанавливают окружности лучше. Качество прогноза посчитанное по формуле (6.1) представлены в табл. 1.

Таблица 1: Результаты работы мультимоделей на синтетических выборках

Выборка	$\mathcal{S}_{\mathfrak{M}_1}$	$\mathcal{S}_{\mathfrak{M}_2}$	$\mathcal{S}_{\mathfrak{M}_3}$
Synthetic 1	10^{-5}	10^{-5}	10^{-5}
Synthetic 2	0.6	10^{-3}	10^{-3}
Synthetic 3	0.6	10^{-3}	10^{-3}

Процесс обучения на синтетических данных. Для анализа свойств мультимоделей $\mathfrak{M}_1, \mathfrak{M}_2, \mathfrak{M}_3$ во время обучения проведен вычислительный эксперимент. В качестве данных рассматривалась синтетическая выборка Synthetic 3.

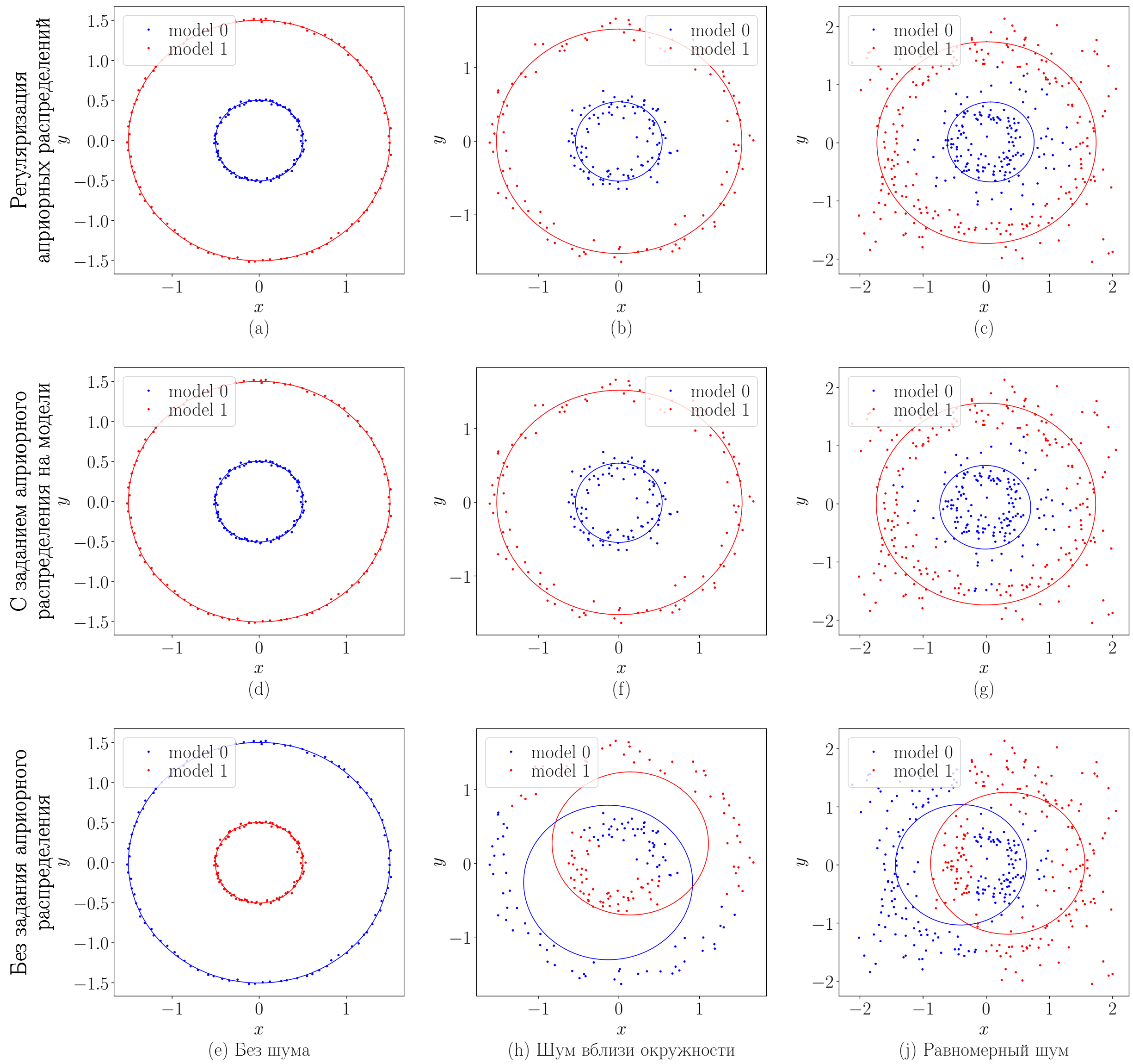


Рис. 2: Мультимодель в зависимости от разных априорных предположений и в зависимости от разного уровня шума: (a)–(c) модель с регуляризаций априорных распределений; (d)–(g) модель с заданными априорными распределениями на параметрах локальных моделей; (e)–(j) модель без заданных априорных предположений

На рис. 3 показана зависимость радиуса и центра окружности от номера итерации. Мультимодель $\mathfrak{M}_2$ с априорным распределением находит центры и радиусы окружностей в среднем лучше, чем мультимодель $\mathfrak{M}_1$ без задания априорного распределения. Мультимодель $\mathfrak{M}_3$ с заданием регуляризатора является более устойчивой, чем мультимодель $\mathfrak{M}_2$, так как дисперсия восстановленных центров и радиуса окружностей меньше.

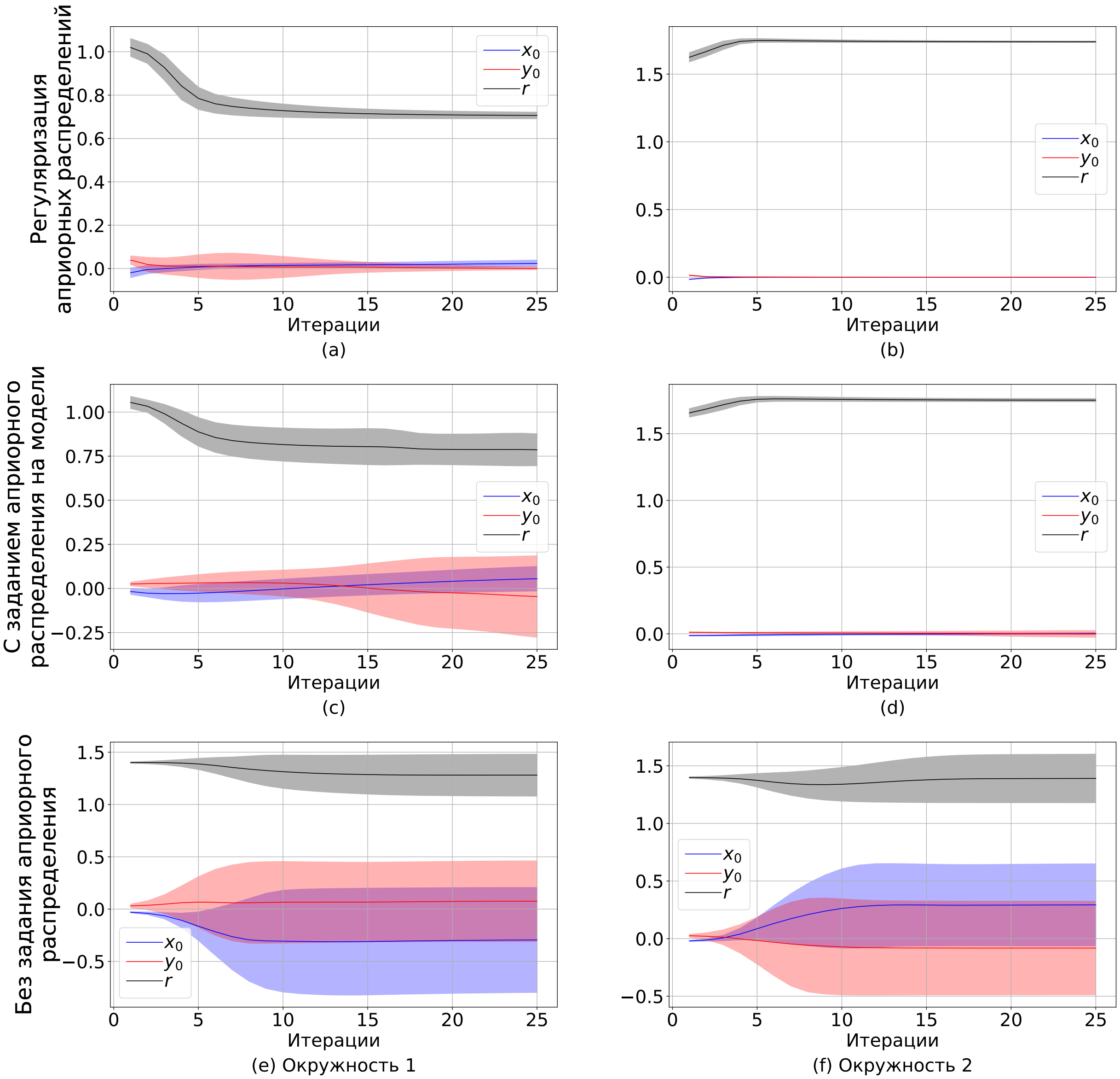


Рис. 3: График зависимости центра и радиуса окружностей от номера итерации: (a)–(b) модель с регуляризацией априорных распределений; (c)–(d) модель с заданными априорными распределениями на параметры моделей; (e)–(f) модель без задания априорных распределений

На рис. 4 показана зависимость правдоподобия мультимодели (6.2) от номера итерации EM–алгоритма. Правдоподобие модели на начальных этапах EM–алгоритма растет быстрее в случае мультимоделей $\mathfrak{M}_2, \mathfrak{M}_3$ чем в мультимодели $\mathfrak{M}_1$. После 20-й итерации все три мультимодели имеют одинаковое правдоподобие.

На рис. 5-7 показан процесс обучения смеси экспертов для разных мультимоделей. На рис. 7 проиллюстрирована работа EM–алгоритма для мультимодели $\mathfrak{M}_1$, которая

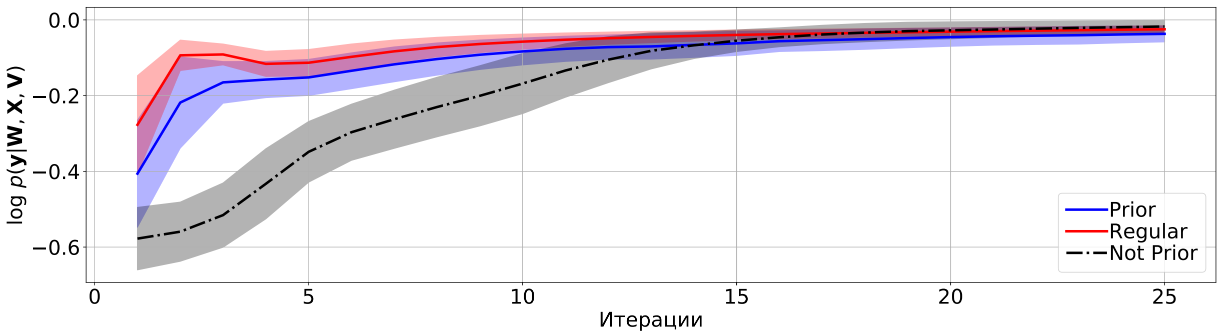


Рис. 4: График зависимости логарифма правдоподобия модели от номера итерации EM-алгорима

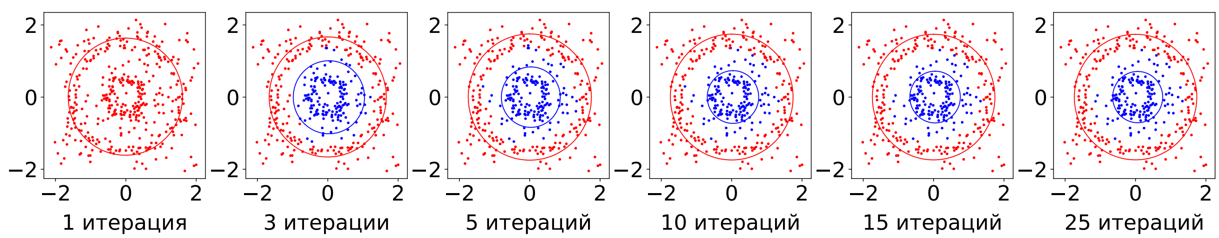


Рис. 5: Визуализация процесса обучения для мультимодели с заданной регуляризацией: от 1й итерации до 25й итерации

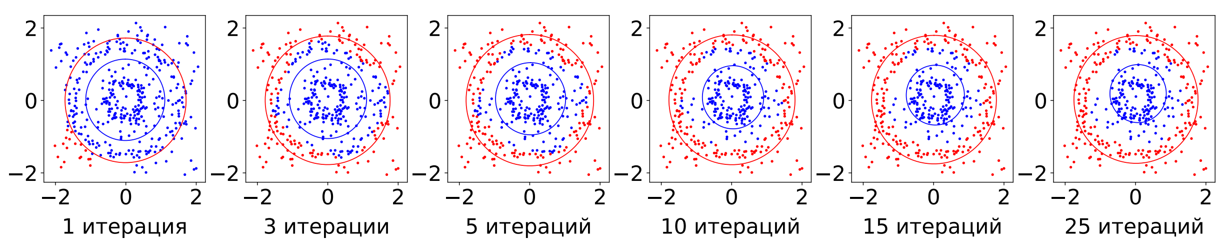


Рис. 6: Визуализация процесса обучения для мультимодели с заданным априорным распределением на параметрах локальных моделей: от 1й итерации до 25й итерации

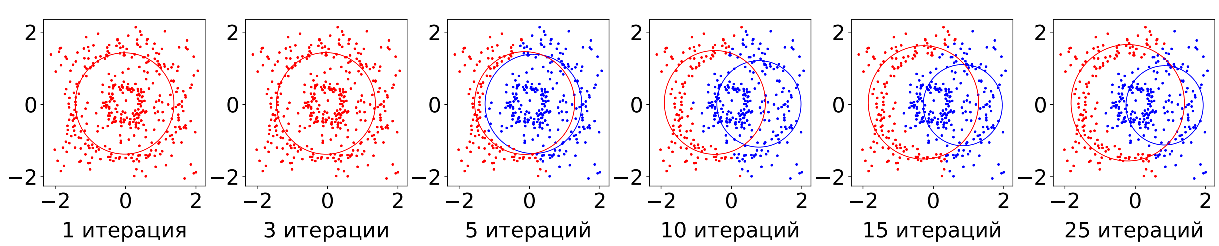


Рис. 7: Визуализация процесса обучения для мультимодели баз заданных априорных распределений: от 1й итерации до 25й итерации

не находит окружности верно. Иллюстрация работы EM–алгоритма для мультимоделей $\mathfrak{M}_2, \mathfrak{M}_3$ показана на рис. 5-6. Мультимодели $\mathfrak{M}_2, \mathfrak{M}_3$ находят обе окружности на изображении.

В ходе данного эксперимента показано, что задание априорных распределений улучшает качество мультимодели, позволяя находить нужные окружности в среднем лучше, чем мультимодель без заданого априорного распределения на параметрах локальных моделей. Задание регуляризации априорных распределений позволяет улучшить устойчивость мультимодели, так как дисперсия центра и радиуса окружностей становится меньше. Также в эксперименте показано, что не смотря на примерное равенство правдоподобий (4) различных мультимоделей, качество предсказание окружностей для разных мультимоделей существенно различается. В случае задания априорного распределения качество нахождения окружностей выше.

Анализ мультимоделей в зависимости от уровня шума. Проведен вычислительный эксперимент, для анализа свойств мультимоделей $\mathfrak{M}_1, \mathfrak{M}_2, \mathfrak{M}_3$ от уровня зашумленности. В качестве данных рассматривалась синтетическая выборка Synthetic 1 с добавлением к ней разного уровня шума. Минимальный уровень шума равен 0, когда нету шумовых точек, а максимальный уровень шума равен 1, когда число шумовых точек равно числу точек обоих окружностей.

На рис. 8 показан график зависимости центра (x_0, y_0) и радиуса r окружностей от уровня шума. Видно, что радиус окружностей растет при увеличении уровня шума. Центры окружностей модели $\mathfrak{M}_2, \mathfrak{M}_3$ в среднем находят верно, но модель с регуляризацией $\mathfrak{M}_3$ имеет меньшую дисперсию. Модель $\mathfrak{M}_1$ имеет худший результат, так как имеет большую дисперсию по всем элементам: x_0, y_0, r.

На рис. 9 показан график зависимости логарифма правдоподобия модели (6.2). Видно, что все модели имеют одинаковое правдоподобие модели, но как показано на рис. 8 качество предсказание окружностей у разных моделей различается.

В данной части эксперимента показано, что наиболее устойчивой является модель $\mathfrak{M}_3$ с регуляризацией априорных распределений.

Реальные данные. Проведен эксперимент на реальной выборке. В качестве данных рассматривались глаза, а точнее их предобработаное бинарное изображение с выделенными границами радужки и роговицы. Проводится анализ качества предсказания моделей $\mathfrak{M}_1, \mathfrak{M}_2, \mathfrak{M}_3$.

На рис. 10 показан результат работы разных мультимоделей. Мультимодель $\mathfrak{M}_1$ не верно находит меньшую окружность. Мультимодели $\mathfrak{M}_2, \mathfrak{M}_3$ одинаково хорошо находят обе окружности.

На рис. 11-13 показан процесс оптимизации мультимоделей. Показано изменение предсказания окружностей мультимоделями в процессе обучения. На рис. 11 показан процесс оптимизации параметров для мультимодели $\mathfrak{M}_1$ без априорных знаний. На рис. 12 показан процесс оптимизации параметров для мультимодели $\mathfrak{M}_2$, в которой задано априорное распределение на параметры локальных моделей. На рис. 13

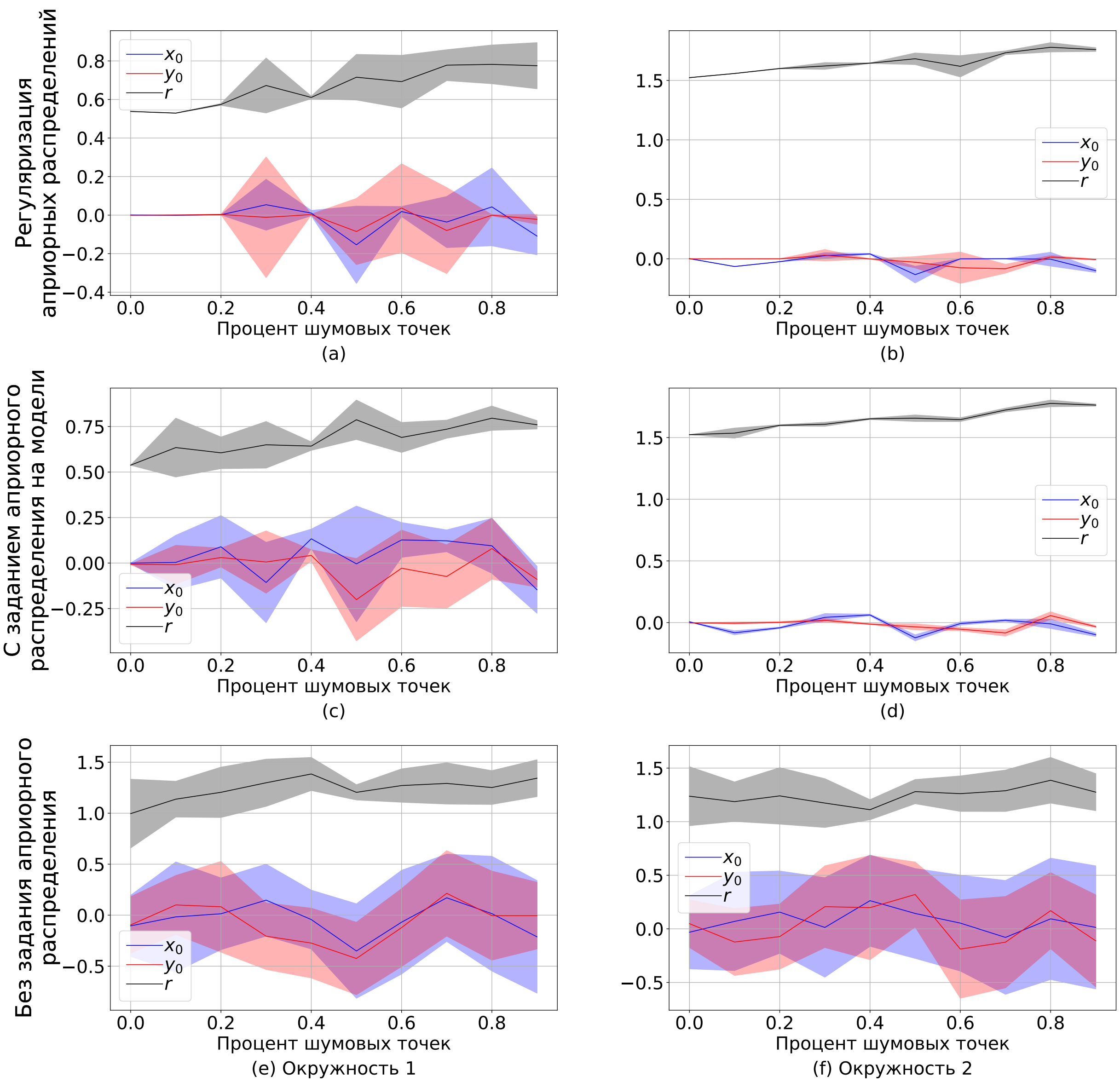


Рис. 8: График зависимости центра и радиуса окружностей от номера итерации: (a)–(b) модель с регуляризацией априорных распределений; (c)–(d) модель с заданными априорными распределениями на параметры моделей; (e)–(f) модель без задания априорных распределений

показан процесс оптимизации параметров для мультимодели $\mathfrak{M}_3$ с регуляризацией априорных распределений на параметрах локальных моделей.

В данной части эксперимента показано, что на реальных данных мультимодели $\mathfrak{M}_2, \mathfrak{M}_3$ с заданными априорными распределениями и регуляризацией являются более точными в определении окружностей чем мультимодель $\mathfrak{M}_2$ без априорных распределений.

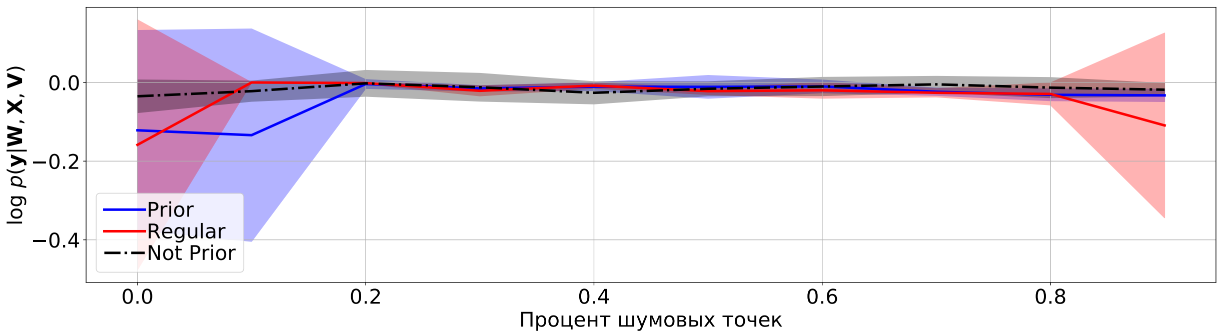
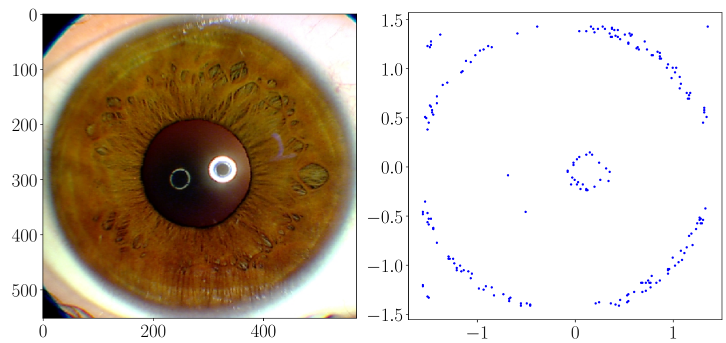
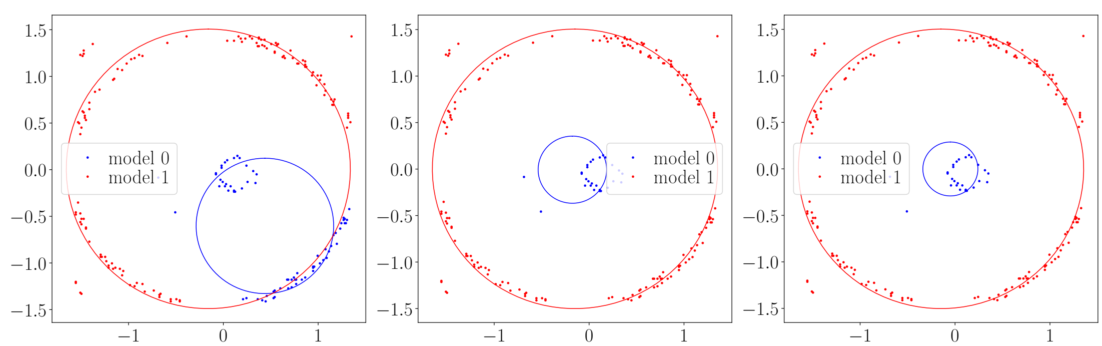


Рис. 9: График зависимости логарифма правдоподобия от уровня шума



(a) Исходное изображение

(b) Бинаризованное изображение



(c) Без задания априорного распределения

(d) С заданием априорного распределения на модели

(e) Регуляризация априорных распределений

Рис. 10: Мультимодель в зависимости от разных априорных предположений на реальном изображении: (a) исходное изображение; (b) бинаризованое изображение; (c) мультимодель без априорных предположений; (d) мультимодель с априорными распределениями на параметрах локальных моделей; (e) мультимодель с регуляризаций на априорных распределениях параметров локальных моделей

7 Заключение

В данной работе проведено сравнение мультимоделей при различных априорных распределениях параметров локальной модели смеси и в случае, когда априорного рас-

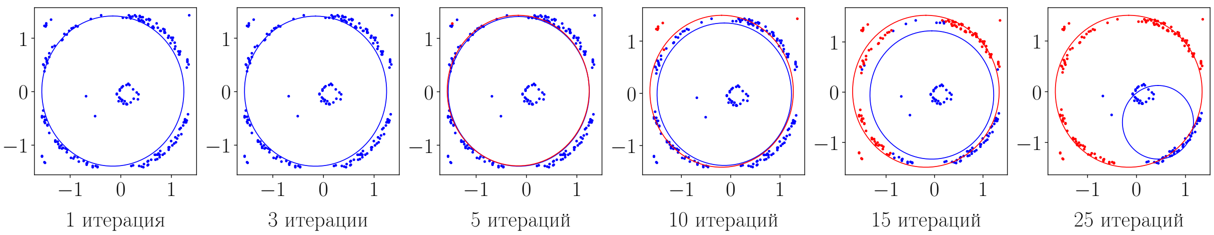


Рис. 11: Визуализация процесса обучения для мультимодели без априорных предположений: от 1й итерации до 15й итерации

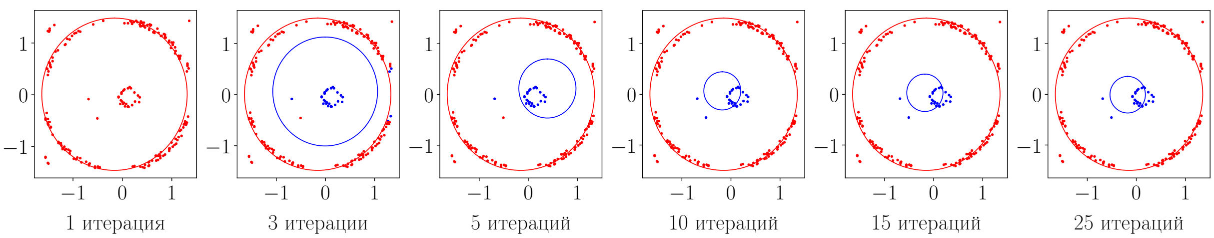


Рис. 12: Визуализация процесса обучения для мультимодели с априорным распределением на параметрах локальных моделей: от 1й итерации до 15й итерации

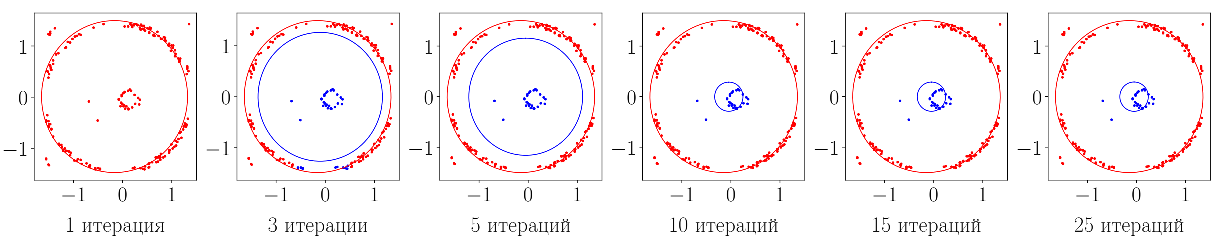


Рис. 13: Визуализация процесса обучения для мультимодели с заданной регуляризацией: от 1й итерации до 15й итерации

пределения не было задано. В качестве данных использовались изображения концентрических окружностей с разным уровнем шума. Для поиска окружностей использовались линейные модели. В качестве шлюзовой функции использовалась двухслойная нейросеть.

Как показано в эксперименте, в случае, когда введены априорные знания на линейные модели, мультимодель является более точной, так как вернее находит окружности на изображениях.

Также был проведен эксперимент по исследованию различных способов регуляризации априорных распределений параметров локальных моделей. В эксперименте показано, что в случае, когда регуляризация задана, мультимодель находит окружности более устойчиво.

В ходе эксперимента было показано, что модели, которыу рассматриваются в работе, является чувствительными к выбросам. Для решения данной проблемы предлагается рассматривать не только локальные модели, которые описывают окружности, но также и модели, которые описывают шум.

В дальнейшем планируется улучшить мультимодель при помощи задания априорного распределения на шлюзовую функцию. Планируется рассмотреть в качестве моделей не только модели, которые описывают данные, а также модель, которая отвечает за шум в данных. Предполагается, что вероятность шума мала, поэтому важно задать априорного распределение, которое учитывало бы этот факт.

Список литературы

[1] *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.

[2] *Chen Xi, Ishwaran Hemant* Random Forests for Genomic Data Analysis // Genomics. 2012. Issues. 99, No 6. pp. 323–329.

[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. Issues. 23, No 8. pp. 1177–1193.

[4] *Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz* Outrageously large neural networks: the sparsely-gated mixture-of-experts layer // ICLR, 2017.

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

[6] *M. I. Jordan* Hierarchical mixtures of experts and the EM algorithm // Neural Comput., vol. 6, no. 2, pp. 181–214, 1994.

[7] *C. A. M. Lima, A. L. V. Coelho, F. J. Von Zuben* Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification // Inf. Sci., vol. 177, no. 10, pp. 2049–2074, 2007.

[8] *L. Cao* Support vector machines experts for time series forecasting // Neurocomputing, vol. 51, pp. 321–339, Apr. 2003.

[9] *A. P. Dempster, N. M. Laird and D. B. Rubin* Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1 pp. 1-38, 1977.

[10] *M. I. Jordan, R. A. Jacobs* Hierarchies of adaptive experts // in Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1991, pp. 985–992.

A Постановка задачи нахождения параметров эллипса

Задано бинарное изображение:

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

где 0 отвечает черной точке — изображения, 1 — белой точке фона.

По изображению $\mathbf{M}$ строится выборка $\mathbf{C}$, элементами которой являются координаты x_i, y_i белых точек на картинке:

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

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

Обозначим x_0, y_0 — центр эллипса, который требуется найти на бинарном изображении $\mathbf{M}$, а a, b его коэффициенты вдоль координат. Элементы выборки $(x_i, y_i) \in \mathbf{C}$ являются геометрическим местом точек, которое заданно уравнением эллипса:

$$(x_i - x_0)^2 + \frac{a^2}{b^2}(y_i - y_0)^2 = a^2. \tag{A.3}$$

Раскрыв скобки получим уравнение

$$(2x_0) \cdot x_i + \left(\frac{2y_0 a^2}{b^2}\right) \cdot y_i + \left(-\frac{a^2}{b^2}\right) y_i^2 + \left(a^2 - x_0^2 - \frac{a^2}{b^2} y_0^2\right) \cdot 1 = x_i^2. \tag{A.4}$$

Получаем задачу линейной регрессии для нахождения параметров окружности:

$$\mathbf{X}\mathbf{w} \approx \mathbf{y}, \quad \mathbf{X} = \mathrm{concat}\left[\mathbf{C}, \left[y_1^2, y_2^2, \cdots, y_N^2\right]^\mathsf{T}, \mathbf{1}\right], \quad \mathbf{y} = [x_1^2, x_2^2, \cdots, x_N^2]^\mathsf{T}, \tag{A.5}$$

где найденые оптимальные параметры линейной регрессии $\mathbf{w} = \left[w_1, w_2, w_3, w_4\right]^\mathsf{T}$ восстанавливают параметры окружности:

$$x_0 = \frac{w_1}{2}, \quad y_0 = -\frac{w_2}{2w_3}, \quad a^2 = w_4 - \frac{w_1}{2} - \frac{w_2^2}{4w_3}, \quad b^2 = -\frac{1}{w_3}\left(w_4 - \frac{w_1}{2} - \frac{w_2^2}{4w_3}\right). \tag{A.6}$$

Решение уравнения (A.5) находит параметры единственного эллипса на изображении.