

Анализ выбора априорного распределения для смеси экспертов *

А. В. Грабовой¹, В. В. Стрижов²

Аннотация: Данная работа посвящен анализу свойств смеси экспертов в зависимости от выбора априорного распределения. Анализируется случай, когда выбрано информативное и неинформативное априорное распределение весов параметров каждого эксперта. В качестве экспертов рассматриваются линейные модели, а в качестве гипермодели 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. В данной работе рассматривается случай, когда априорные распределения разных 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 мультимоделей с заданными априорными распределениями на вектора параметров локальных моделей и без задания априорного распределения.

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

DOI: 00.00000/0000000000000000

1 Введение

В данной работе рассматривается задача смеси экспертов. Смесь экспертов — это мультимодель, которая линейно взвешивает локальных моделей аппроксимирующих выборку. Значение весовых коэффициенты зависят от того объекта для которого производится предсказание.

*Работа выполнена при поддержке РФФИ и правительства РФ.

¹Московский физико-технический институт, grabovoy.av@phystech.edu

²Московский физико-технический институт, strijov@ccas.ru

Примерами мультимodelей являются беггинг и градиентный бустинг [1], случайный лес [2]. 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 мультимodelи и локальных моделей рассматривается вероятностная постановка задачи. В качестве функционала качества рассматривается логарифм правдоподобия модели. Для оптимизации данного функционала рассматривается ЕМ-алгоритм [9].

Мультимodelей имеют ряд недостатков, которые связаны с тем, что сходимость локальных моделей сильно зависит от начальной инициализации векторов параметров. Для улучшения сходимости предлагается использовать априорные знания. В данной работе задается априорное распределение на веса локальных моделей, также предлагается использовать зависимость априорных распределений, для улучшения качества мультимodelи.

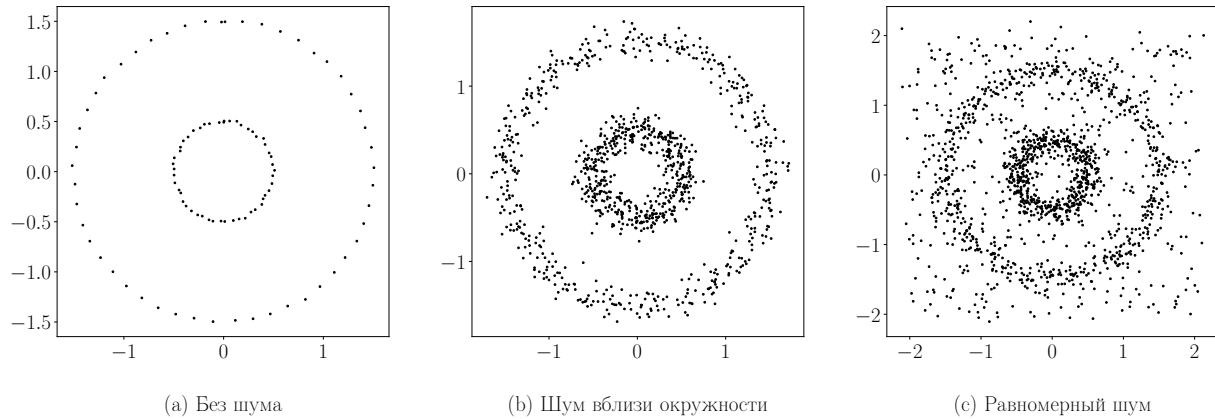


Рис. 1: Пример изображений с окружностями с разным уровнем шума: (а) окружности без шума; (b) окружности с зашумленным радиусом; (с) окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению

Данная работа исследует зависимость качества модели в зависимости от выбора априорных распределений. 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 знаниями и без них в случае изображений с разным уровнем шума. В данной работе в качестве отдельных экспертов рассматриваются линейные модели — каждая модель отвечает своей окружностью. В качестве шлюзовой функции рассматривается двухслойная нейронная сеть.

2 Работы по теме

Большое количество работ посвящены выбору плюсовой функции: 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] рассматривают 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 плюсовых функций. Также в данной работе проведен анализ разных моделей, которые могут выступать в качестве локальной модели.

3 Постановка задачи нахождения параметров окружностей

Задано бинарное изображение:

$$\mathbf{M} \in \{0, 1\}^{m_1 \times m_2}, \quad (3.1)$$

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

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

$$\mathbf{C} \in \mathbb{R}^{N \times 2}, \quad (3.2)$$

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 заданное уравнение окружности

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

Раскрыв скобки получим уравнение

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

Получаем задачу линейной регрессии для нахождения параметров окружности:

$$\mathbf{X}\mathbf{w} \approx \mathbf{y}, \quad \mathbf{X} = \mathbf{C} \times \mathbf{1}, \quad \mathbf{y} = [x_1^2 + y_1^2, x_2^2 + y_2^2, \dots, x_N^2 + y_N^2]^T, \quad (3.5)$$

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

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

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.

4 Постановка задачи построения смеси экспертов

Задана выборка из (3.5)

$$\mathbf{X} \in \mathbb{R}^{N \times n}, \quad (4.1)$$

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 \mathbf{f}_k на признаковом описании объекта \mathbf{x} .*

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

где $\hat{\mathbf{f}}$ — мультимодель, а \mathbf{f}_k является некоторой моделью, π_k — параметрическая модель, \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 π рассматриваются следующие функции:

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

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

В качестве функционала качества рассматривается правдоподобие модели:

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

где $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_K]^T$

Получаем следующую задачу оптимизации:

$$\mathbf{W}, \mathbf{V} = \arg \max_{\mathbf{W}, \mathbf{V}} p(\mathbf{y}, \mathbf{W} | \mathbf{X}, \mathbf{V}) \quad (4.5)$$

5 ЕМ-алгоритм для решения задачи смеси экспертов

Рассмотрим следующую вероятностную постановку задачи:

$$1. p_k(y_i | \mathbf{w}_k, \mathbf{x}_i) = \mathcal{N}(y_i | \mathbf{w}_k^T \mathbf{x}_i, \beta^{-1})$$

$$2. p^k(\mathbf{w}_k) = \mathcal{N}(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k)$$

$$3. p(\boldsymbol{\varepsilon}_{k,k'} | \boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\varepsilon}_{k,k'} | \mathbf{0}, \hat{\boldsymbol{\alpha}}), \text{ где } \boldsymbol{\varepsilon}_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0, \hat{\boldsymbol{\alpha}} = \text{diag}(\boldsymbol{\alpha}) \text{ — диагональная матрица, диагональ которой равняется } \boldsymbol{\alpha}.$$

Тогда правдоподобия модели (4.4) переписывается в следующем виде:

$$p(\mathbf{y}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \hat{\boldsymbol{\alpha}}, \beta) = \prod_{k,k'=1}^K \mathcal{N}(\boldsymbol{\varepsilon}_{k,k'} | \mathbf{0}, \hat{\boldsymbol{\alpha}}) \cdot \prod_{k=1}^K \mathcal{N}(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k) \prod_{i=1}^N \left(\sum_{k=1}^K \pi_k \mathcal{N}(y_i | \mathbf{w}_k^\top \mathbf{x}_i, \beta^{-1}) \right), \quad (5.1)$$

где $\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_K]$

Для решения задачи (4.5) в условиях правдоподобия (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} , логарифм правдоподобия (5.1) переписывается следующим образом:

$$\begin{aligned} \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \hat{\boldsymbol{\alpha}}, \beta) &= \\ &= \sum_{i=1}^N \sum_{k=1}^K z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (y_i - \mathbf{w}_k^\top \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\ &+ \sum_{k=1}^K \left[-\frac{1}{2} (\mathbf{w}_k - \mathbf{w}_k^0)^\top \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] + \\ &+ \sum_{k=1}^K \sum_{k'=1}^K \left[-\frac{1}{2} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0)^\top \hat{\boldsymbol{\alpha}}^{-1} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0) + \frac{1}{2} \log \det \hat{\boldsymbol{\alpha}} - \frac{n}{2} \log 2\pi \right]. \end{aligned} \quad (5.2)$$

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

$$\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta = \arg \max_{\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta} \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \hat{\boldsymbol{\alpha}}, \beta) \quad (5.3)$$

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

E-step. Найдем $q(\mathbf{Z}, \mathbf{W}) = q(\mathbf{Z}) q(\mathbf{W})$ наиболее близкое к $p(\mathbf{Z}, \mathbf{W} | \mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \hat{\boldsymbol{\alpha}}, \beta)$.

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

Получаем, что параметр $q(z_{ik})$ является Бернулевской случайной величиной с параметров заданным в выражении (5.4).

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

Получаем, что распределение $q(\mathbf{w}_k) = \mathcal{N}(\mathbf{w}_k | \mathbf{m}_k, \mathbf{B}_k)$, где параметры $\mathbf{m}_k, \mathbf{B}_k$ определяются следующим образом:

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

M-step. Найдем $\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta$ из максимизации $\mathbb{E}_q \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \hat{\alpha}, \beta)$.

$$\begin{aligned}
\mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta) &= \mathbb{E}_q \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \hat{\alpha}, \beta) = \\
&= \sum_{i=1}^N \sum_{k=1}^K \mathbb{E} z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} \mathbb{E} (y_i - \mathbf{w}_k^\top \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\
&+ \sum_{k=1}^K \left[-\frac{1}{2} \mathbb{E} (\mathbf{w}_k - \mathbf{w}_k^0)^\top \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] + \\
&+ \sum_{k=1}^K \sum_{k'=1}^K \left[-\frac{1}{2} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0)^\top \hat{\alpha}^{-1} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0) + \frac{1}{2} \log \det \hat{\alpha} - \frac{n}{2} \log 2\pi \right].
\end{aligned} \tag{5.7}$$

Так-как $\mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)$ является вогнутой, то для нахождения оптимальных параметров воспользуемся условием первого порядка.

$$\frac{\partial \mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \mathbf{V}} = 0 \tag{5.8}$$

Решение уравнения (5.8) найдем при помощи градиентного метода оптимизации, которое гарантирует сходимость к локальному экстремуму функции $\mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)$.

$$\begin{aligned}
\frac{\partial \mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \mathbf{A}_k^{-1}} &= \frac{1}{2} \mathbf{A}_k - \frac{1}{2} \mathbb{E} (\mathbf{w}_k - \mathbf{w}_k^0) (\mathbf{w}_k - \mathbf{w}_k^0)^\top = 0 \\
\mathbf{A}_k &= \mathbb{E} \mathbf{w}_k \mathbf{w}_k^\top - \mathbf{w}_k^0 \mathbb{E} \mathbf{w}_k^\top - \mathbb{E} \mathbf{w}_k \mathbf{w}_k^{0\top} + \mathbf{w}_k^0 \mathbf{w}_k^{0\top}
\end{aligned} \tag{5.9}$$

$$\begin{aligned}\frac{\partial \mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \beta} &= \sum_{k=1}^K \sum_{i=1}^N \left(\frac{1}{\beta} E z_{ik} - \frac{1}{2} E z_{ik} [y_i^2 - 2y_i \mathbf{x}_i^T \mathbf{E} \mathbf{w}_k + \mathbf{x}_i^T \mathbf{w}_k \mathbf{w}_k^T \mathbf{x}_i] \right) = 0 \\ \frac{1}{\beta} &= \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K [y_i^2 - 2y_i \mathbf{x}_i^T \mathbf{E} \mathbf{w}_k + \mathbf{x}_i^T \mathbf{E} \mathbf{w}_k \mathbf{w}_k^T \mathbf{x}_i] E z_{ik}\end{aligned}\tag{5.10}$$

$$\begin{aligned}\frac{\partial \mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \mathbf{w}_k^0} &= \mathbf{A}_k^{-1} (\mathbf{E} \mathbf{w}_k - \mathbf{w}_k^0) + \hat{\alpha} \sum_{k'=1}^K [\mathbf{w}_{k'}^0 - \mathbf{w}_k^0] = 0 \\ \mathbf{w}_k^0 &= [\mathbf{A}_k^{-1} + (K-1) \hat{\alpha}]^{-1} \left(\mathbf{A}_k^{-1} \mathbf{E} \mathbf{w}_k + \hat{\alpha} \sum_{k'=1, k' \neq k}^K \mathbf{w}_{k'}^0 \right).\end{aligned}\tag{5.11}$$

Используя формулы (5.4-5.11) получаем итеративный процесс, который 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

Синтетические данные. Для сравнения качества работы мультимodelей смеси экспертов с разными начальными предположениями был проведен вычислительный эксперимент на синтетических данных. Рассматривалась мультимodelь в котором не было задано никаких априорных знаний. Рассматривалась мультимodelь, где в качестве априорных знаний задавались априорные распределения на вектора параметров локальных моделей, также рассматривалась мультимodelь, где в качестве априорных знаний дополнительно была введена регуляризация априорных распределений.

Вычислительный эксперимент проводится на синтетической выборке, которая получена при помощи генерации двух концентрических окружностей с разным уровнем шума.

Предлагается сравнить две постановки задачи смеси экспертов: в случае когда используются априорные знания об изображении и в случае когда априорные знания отсутствуют. Причем априорные знания задаются двумя разными способами: с регуляризацией априорных распределений и без нее.

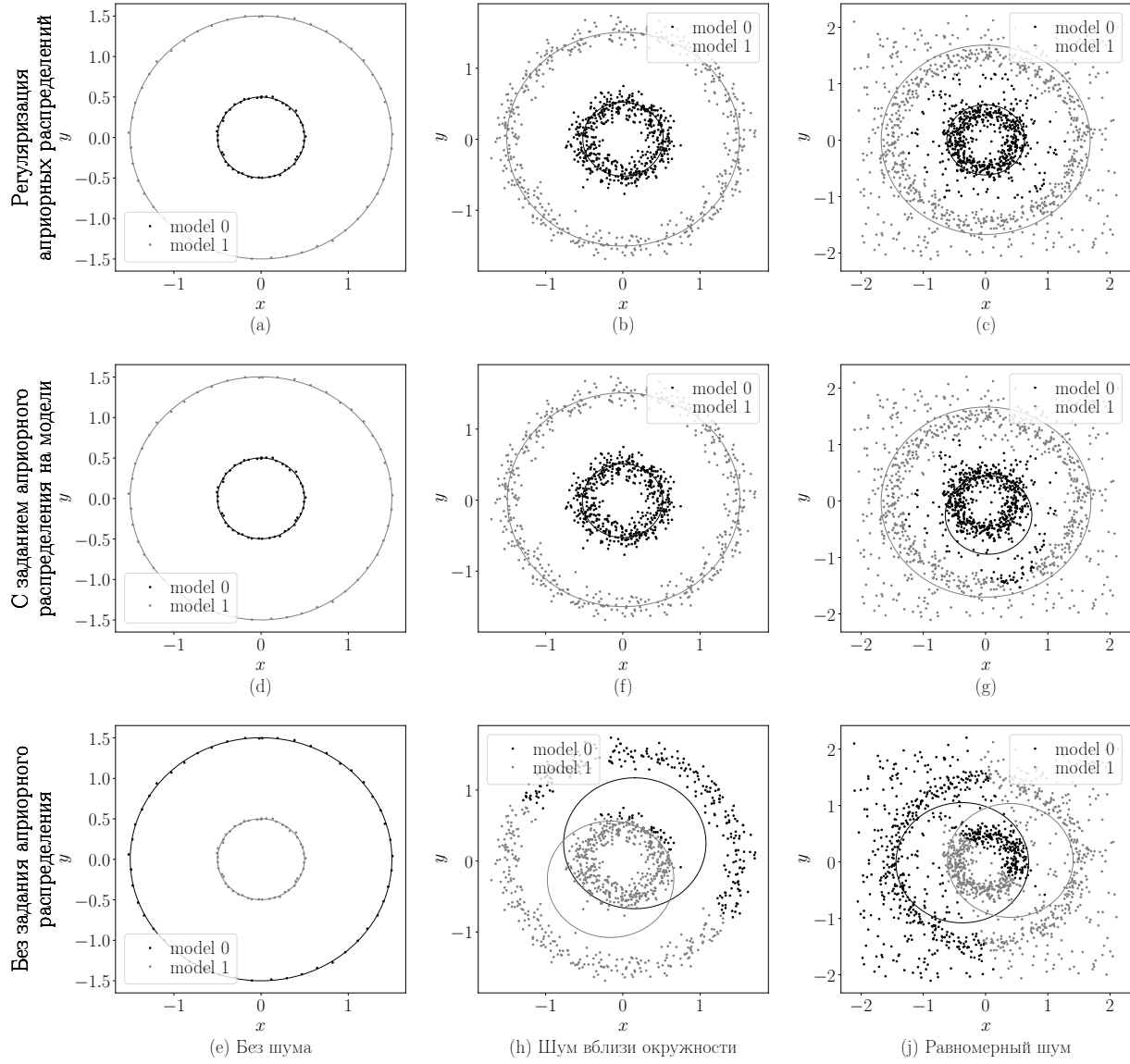


Рис. 2: Мультимодель в зависимости от разных априорных предположений и в зависимости от разного уровня шума: (a)–(c) модель с регуляризацией априорных распределений; (d)–(g) модель с заданными априорными распределениями на параметрах локальных моделей; (e)–(j) модель без заданных априорных предположений

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

$$\mathcal{N}(\mathbf{w}_1|\mathbf{w}_1^0, \mathbf{I}), \quad \mathcal{N}(\mathbf{w}_2|\mathbf{w}_2^0, \mathbf{I}), \quad (6.1)$$

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.

На рис. 2 показан случайный результаты работы мультимodelей с априорными знаниями и без них. На всех картинках обе модели работала 50 итераций. Так как сходимость мультимodelей очень сильно зависит от начальной инициализации, также был проведен эксперимент с множественным запуском мультимodelей на одном и том же изображении, после чего результаты были усреднены.

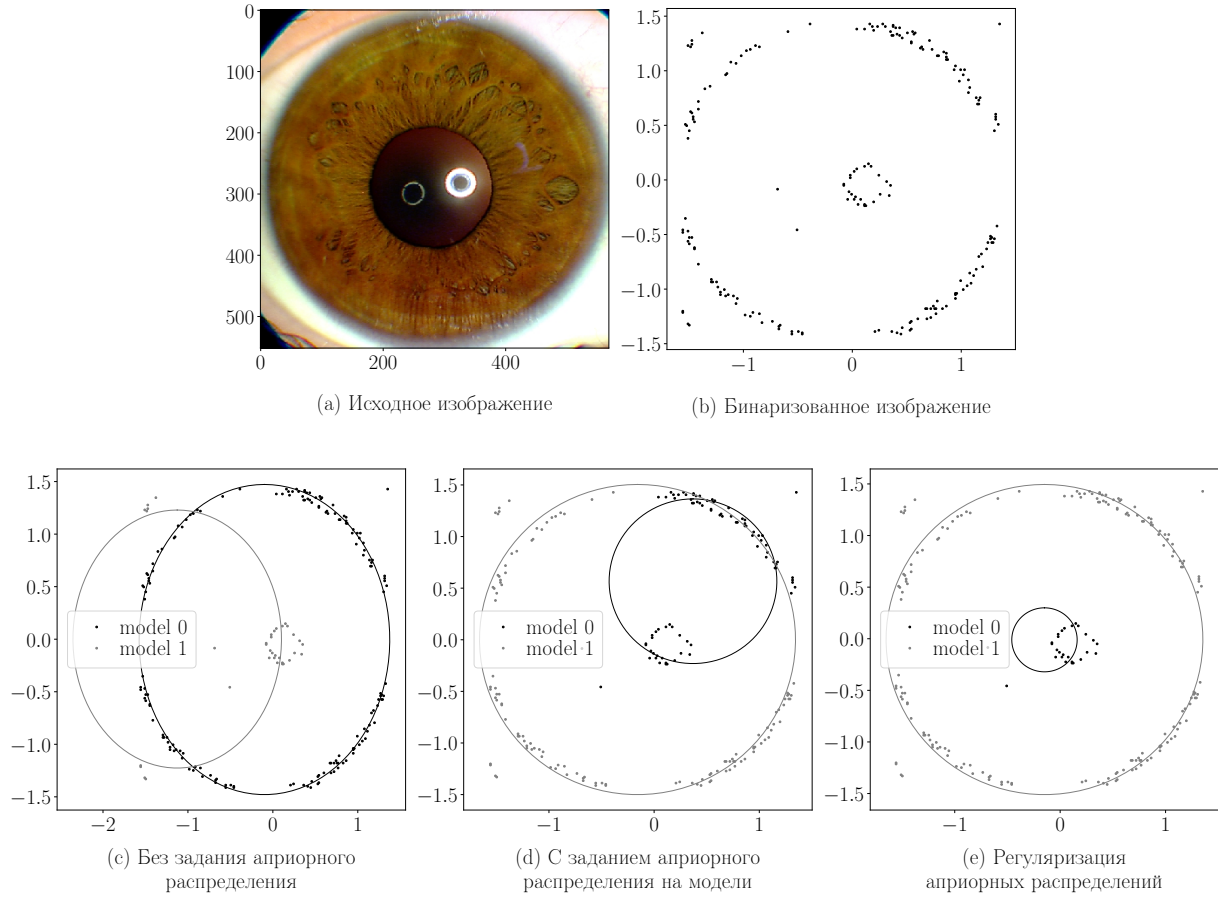


Рис. 3: Мультимodel в зависимости от разных априорных предположений на реальном изображении: (a) исходное изображение; (b) бинаризованное изображение; (c) мультимodel без априорных предположений; (d) мультимodel с априорными распределениями на параметрах локальных моделей; (e) мультимodel с регуляризацией на априорных распределениях параметров локальных моделей

Реальные данные. Проведен эксперимент на реальной выборке. В качестве данных рассматривались глаза, а точнее их предобработанное бинарное изображение с выделенным окружностями. Предлагается сравнить работу мультимodelей с разным количеством априорной информации об изображениях. Рассматривалась мультимodel в котором не было задано никаких априорных знаний. Рассматривалась мультимodel, где в качестве априорных знаний задавались априорные распределения на

вектора параметров локальных моделей, также рассматривалась мультимодель, где в качестве априорных знаний дополнительно была введена регуляризация априорных распределений.

На рис. 3 показан результат работы мультимodelей с разным количеством априорной информации. В случае, когда априорные распределения не заданы мультимодель верно находит внешнюю окружность, но не находит внутреннюю окружность. В случае, когда задано априорное распределение на параметры локальных моделей качество нахождение окружностей улучшилось, но внутренняя окружность все еще найдена не верно. После добавления к заданным априорным распределениям параметров регуляризации качество мультимodelи выросло, что позволило верно найти и внутреннюю окружность.

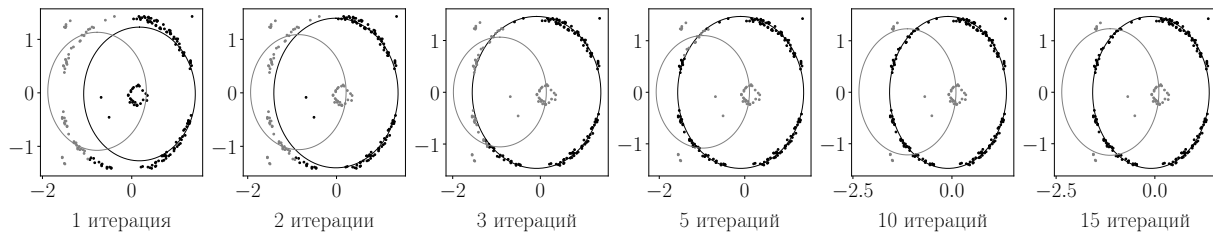


Рис. 4: Визуализация процесса обучения для мультимodelи без априорных предположений: от 1й итерации до 15й итерации

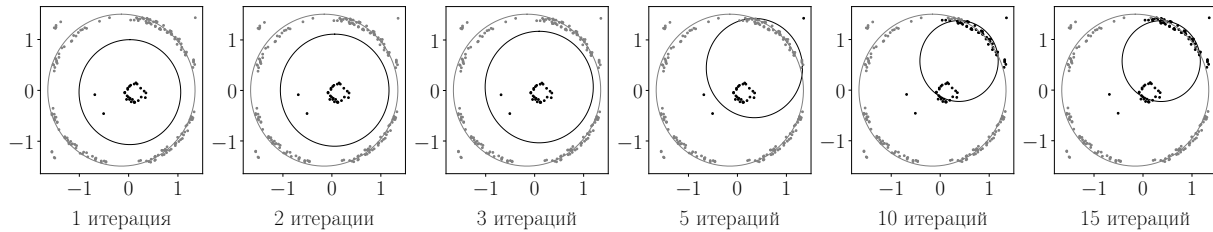


Рис. 5: Визуализация процесса обучения для мультимodelи с априорным распределением на параметрах локальных моделей: от 1й итерации до 15й итерации

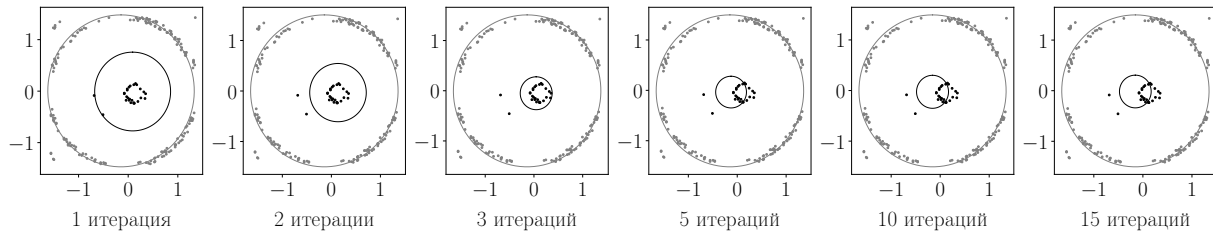


Рис. 6: Визуализация процесса обучения для мультимodelи с заданной регуляризацией: от 1й итерации до 15й итерации

На рис. 4-6 показан процесс оптимизации мультимodelей и как менялись их предсказания в процессе обучения. На рис. 4 показан процесс оптимизации для мультимodelей без априорных знаний. На рис. 5 показан процесс оптимизации для мультимodelей, в которой на параметры локальных моделей задано априорное распределение. На рис. 6 процесс оптимизации для мультимodelей с априорным распределением на параметры локальных моделей, а также регуляризация априорных распределений.

7 Заключение

В данной работе проведено сравнение мультимodelей в случае, когда было задано априорное распределение параметров каждой модели внутри мультимodelей и в случае, когда априорного распределения не было. 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, мультимodelь является более точной, так как вернее находит окружности на изображениях.

Также был проведен эксперимент по заданию регуляризации на априорные распределения параметров локальных моделей. В эксперименте показано, что в случае, когда регуляризация задана мультимodelь находит окружности точнее.

В ходе эксперимента было показано, что модель, которая рассматривалась в работе является чувствительной 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 мультимodelь 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 априорное 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.