Анализ выбора априорного распределения для смеси экспертов *

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

Аннотация: Данная работа посвящен анализу свойств смеси экспертов в зависимости от выбора априорного распределения. Анализируется случай, когда выбрано информативное и неинформативное априорное распределения 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 соответсвует свой эксперт. В данной работе рассматривается случай, когда априорные распределения разных моделей являются независимы, а также случай, когда эти распределения являются зависиимыми. В качестве данных рассматриваются синтетически сгенерированные окружности с разным уровнем шума. Сравнивается устойчивость к шуму мультимомделей с задаными априорными распределениями на вектора параметров локальных моделей и без задания априорного распределения.

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

DOI: 00.00000/00000000000000

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.

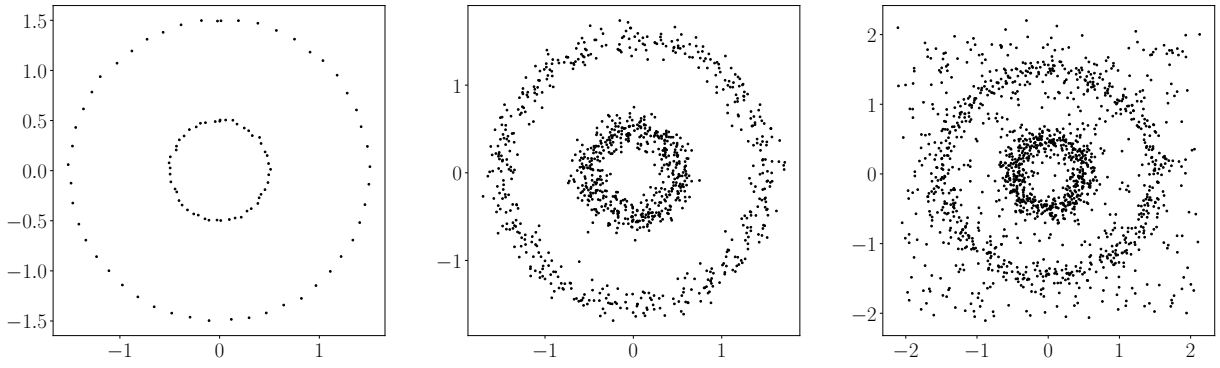
[1]Московский физико-технический институт, grabovoy.av@phystech.edu

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

Примерами мультимоделей 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 мультимодели и локальных моделей рассматривается вероятностная постановка задачи. В качестве функционала качества рассматривается логарифм правдоподобия модели. Для оптимизации данного функционала рассматривается EM-алгоритм [9].

Мультимоделей 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 распределения на веса локальных моделей, также предлагается использовать зависимость априорных распределений, для улучшения качества мультимодели.



(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. 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 посвящены выбору шлюзовой 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] 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

Задано бинарное изображение:

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

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:

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

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 окружности

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

Раскрыв скобки получим уравнение

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

Получаем задачу линейной регрессии для нахождения параметров окружности:

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

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

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 — 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 $\boldsymbol{\pi}$ рассматриваются следующие функции:

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

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

В качестве функционала качества рассматривается правдоподобие модели:

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

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

Получаем следующую задачу оптимизации:

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

5 EM–алгоритм для решения задачи смеси экспертов

Рассмотрим следующую вероятностную постановку задачи:

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

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

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

Тогда правдоподобия модели (4.4) переписывается в следующем виде:

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

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

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 правдоподобие (5.1) переписывается следующим образом:

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

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

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

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 $p\big(\mathbf{Z}, \mathbf{W}|\mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \hat{\boldsymbol{\alpha}}, \beta\big)$.

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

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

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

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

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

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

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

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

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

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

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

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

$$\frac{1}{\beta} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[y_i^2 - 2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right]\mathsf{E}z_{ik} \tag{5.10}$$

$$\frac{\partial \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)}{\partial \mathbf{w}_k^0} = \mathbf{A}_k^{-1}\left(\mathsf{E}\mathbf{w}_k - \mathbf{w}_k^0\right) + \hat{\boldsymbol{\alpha}}\sum_{k'=1}^{K}\left[\mathbf{w}_{k'}^0 - \mathbf{w}_k^0\right] = 0$$

$$\mathbf{w}_k^0 = \left[\mathbf{A}_k^{-1} + (K-1)\,\hat{\boldsymbol{\alpha}}\right]^{-1}\left(\mathbf{A}_k^{-1}\mathsf{E}\mathbf{w}_k + \hat{\boldsymbol{\alpha}}\sum_{k'=1,\ k'\neq k}^{K}\mathbf{w}_{k'}^0\right). \tag{5.11}$$

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 мультимоделей смеси экспертов с разными начальными предположениями был проведен вычислительный эксперимент на синтетических данных. Рассматривалась мультимодель в котором не было задано никаких априорных знаний. Рассматривалась мультимодель, 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.

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.

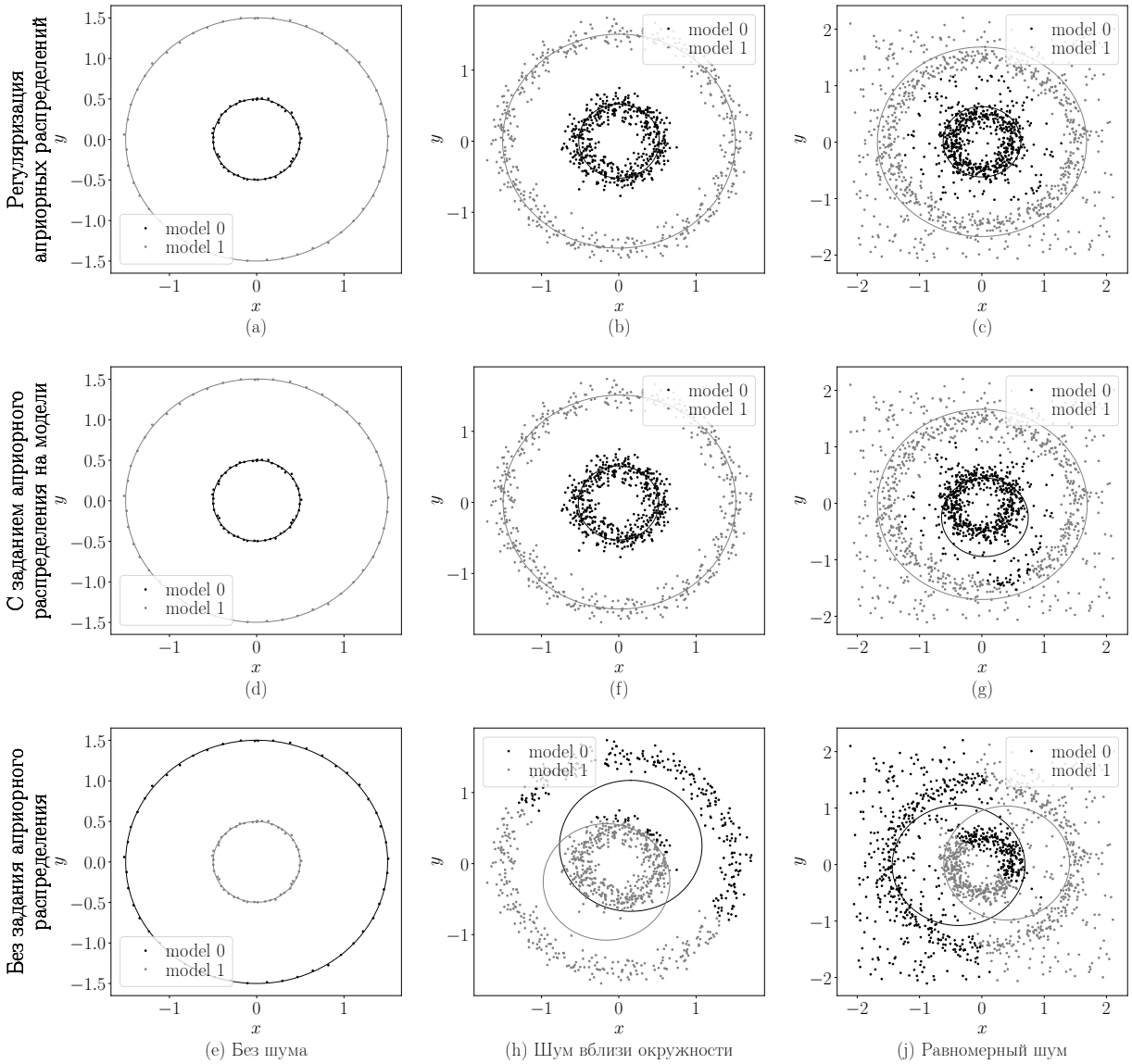


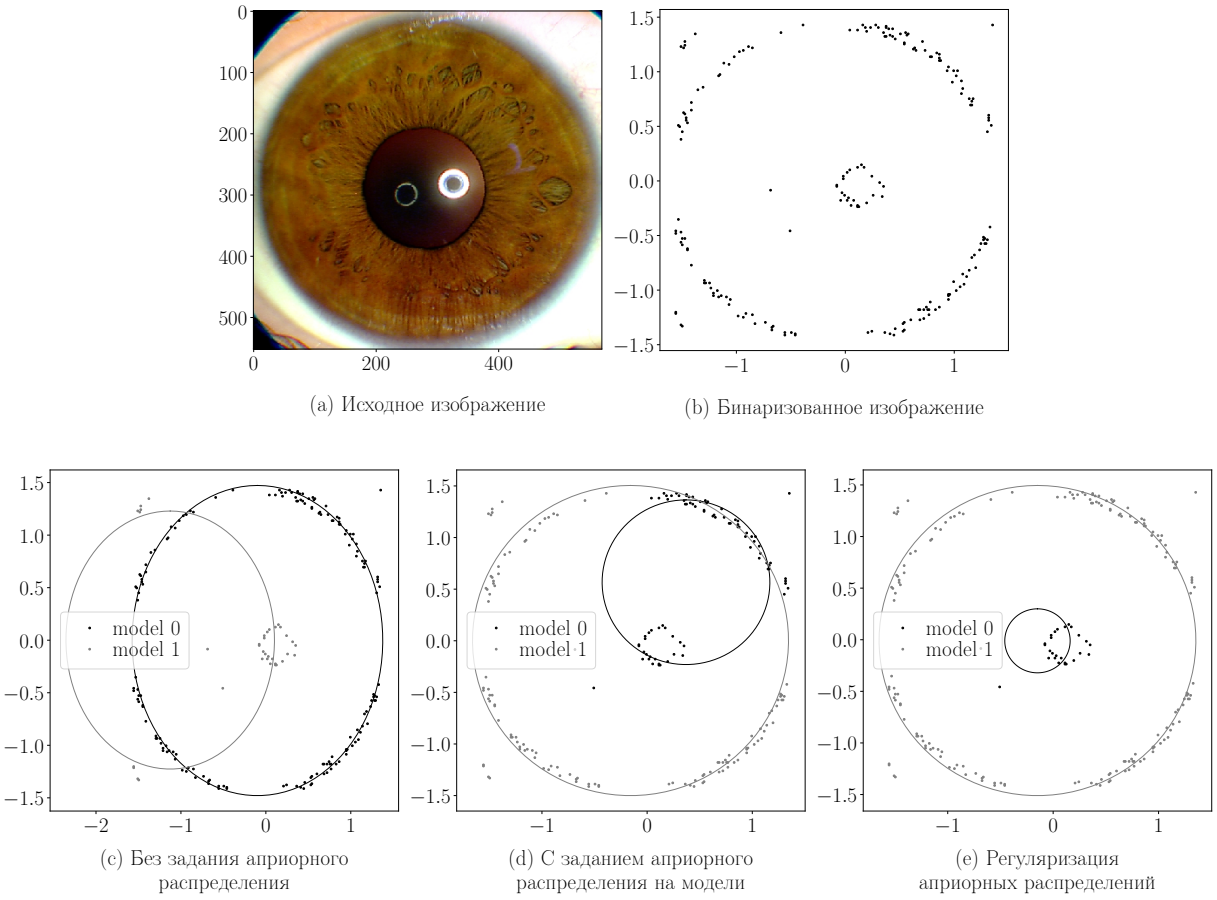
Рис. 2: Мультимодель в зависимости от разных априорных предположений и в зависимости от разного уровня шума: (a)–(c) модель с регуляризаций априорных распределений; (d)–(g) модель с заданными априорными распределениями на параметрах локальных моделей; (e)–(j) модель без заданных априорных предположений

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

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

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



(a) Исходное изображение

(b) Бинаризованное изображение

(c) Без задания априорного распределения

(d) С заданием априорного распределения на модели

(e) Регуляризация априорных распределений

Рис. 3: 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

Реальные данные. Проведен эксперимент на реальной выборке. В качестве данных рассматривались глаза, а точнее их предобработаное бинарное изображение с выделенным окружностями. Предлагается сравнить работу мультимоделей с разным количеством априорной информации об изображениях. Рассматривалась мультимодель в котором не было задано никаких априорных знаний. Рассматривалась мультимодель, 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 мультимоделей 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 мультимодели выросло, что позволило верно найти и внутреннюю окружность.

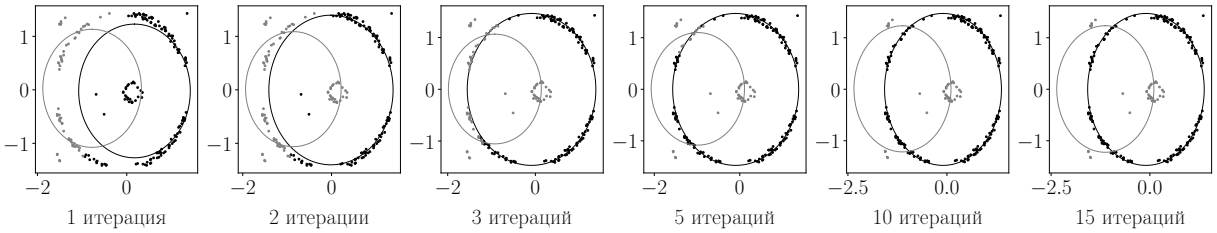


Рис. 4: Визуализация процесса обучения для мультимодели без априорных предположений: от 1й итерации до 15й итерации

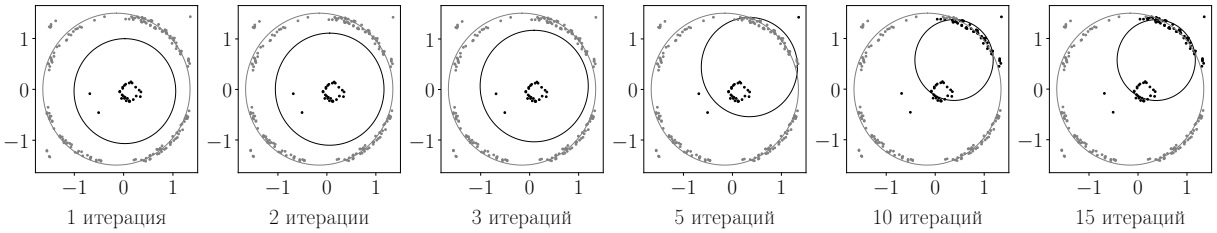


Рис. 5: Визуализация процесса обучения для мультимодели с априорным распределением на параметрах локальных моделей: от 1й итерации до 15й итерации

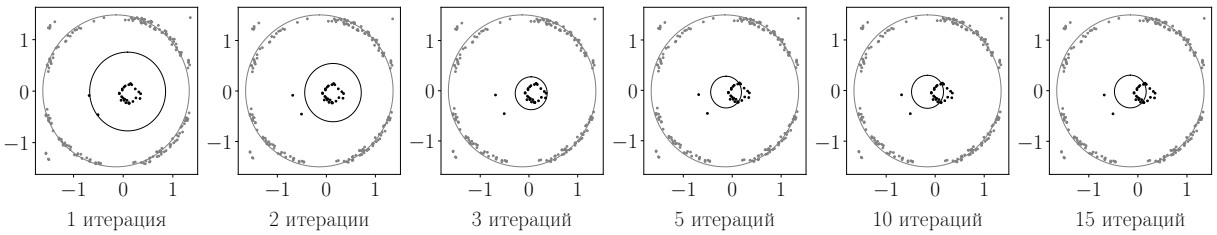


Рис. 6: Визуализация процесса обучения для мультимодели с заданной регуляризацией: от 1й итерации до 15й итерации

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

7 Заключение

В данной работе проведено сравнение мультимоделей в случае, когда было задано априорное распределение параметров каждой модели внутри мультимодели 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, мультимодель 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 мультимодель 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 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.