Анализ выбора априорного распределения для смеси экспертов *

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

Аннотация: Данная работа исследует свойства смеси экспертов. Смесь экспертов это ансамбля локальных аппроксимирующих моделей, которые являются экспертами и шлюзовой функции, которая взвешивает данных экспертов. В данной работе в качестве экспертов рассматриваются линейные модели, а в качестве шлюзовую функции рассматривается нейронная сеть с функций softmax на последнем слое. В работе анализируются разные априорные распределения для каждого эксперта. Авторы предлагают метод, который учитывает связь между априорными распределениями разных экспертов. Для поиска оптимальных параметров локальных моделей и шлюзовой функции в работе используется EM алгоритм. В работе рассматривается задача распознавания окружностей на изображении. Каждый эксперт аппроксимирует одну окружность на изображении: находит координаты центра окружности и радиус окружности. Для анализа предложенного метода проводится вычислительный эксперимент на синтетических и реальных данных. В качестве реальных данных используются изображения радужки глаза, которые используются в задачах распознавания радужки глаза.

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

1 Введение

В статье исследуется проблема построения смеси экспертов. Смесь экспертов - это мультимодель, которая состоит из множества локальных моделей, которые называ-

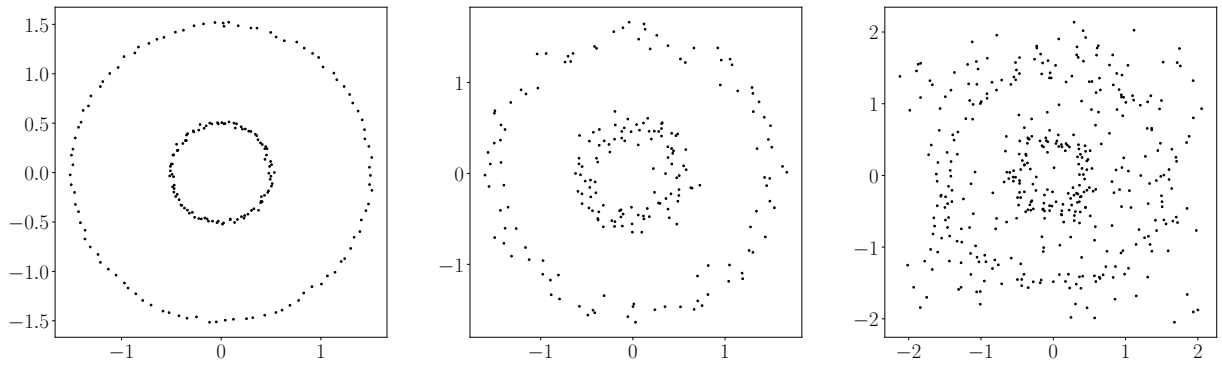
*Работа выполнена при поддержке РФФИ (проекты 19-07-01155, 19-07-0875), НТИ (проект 13/1251/2018) и правительства РФ (соглашение 5.Y09.21.0018).

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

[2]Московский физико-технический институт, Вычислительный центр имени А. А. Дородницына ФИЦ ИУ РАН, strijov@phystech.edu

ются экспертами и шлюзовой функции. Смест экспертов использует шлюзовую функцию для взвешивания прогнозов каждого эксперта. Весовые коэффициенты шлюзовую функции зависят от объекта, для которого производится прогноз. Примерами мультимоделей являются бэггинг, градиентный бустинг [1] и случайный лес [2]. В статье [3] предполагается, что вклад каждого эксперта в ответ зависит от объекта из набора данных.

Основной проблемой построения мультимоделей является то, что ансамбль зависит от начальной инициализации параметров. Для улучшения устойчивости мультимодели предлагается использовать вероятностную постановку задачи для поиска оптимальных параметров шлюзовой функции и параметров локальной модели. В данной работе задается априорное распределения на параметры локальных моделей, также, для повышения, предлагается учесть зависимость априорных распределений для разных моделей.



(a) Without noise (b) Noise near circle (c) Uniform noise

Рис. 1: Пример окружностей 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

В данной работе 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. В данной работе в качестве экспертов рассматриваются линейные модели — каждая модель аппроксимирует одну окружности. В качестве шлюзовой функции рассматривается двухслойная нейронная сеть.

Работы по теме Большое количество работ в области построения смеси экспертов посвящены выбору шлюзовой функции: используется softmax, процесс Дирихле [4], нейронная сеть [5] с функцией softmax на последнем слое. Ряд работ посвящены выбору моделей в качестве отдельных экспертов. В работах [6, 7] в качестве модели эксперта рассматривается линейная модель. Работы [8, 9] рассматриваю модель SVM

в качестве модели эксперта. В работе [3] представлен обзор методов и моделей в задачах смеси экспертов.

Смесь экспертов имеет множество приложений в прикладных задачах. Работы [10, 11, 12] посвящены применению смеси экспертов в задачах прогнозирования временных рядов. В работе [13] предложен метод распознавания рукописных цифр. Метод распознания текстов при помощи смеси экспертов иследуется в работах [14], распознание речи [15, 16, 17]. В работе [18] исследуется смесь экспертов для задачи распознавания трехмерных движений человека. В [19] описаны работы по исследованию обнаружения радужки глаза на изображении. В работах [20, 21] в частности описаны методы выделения границ радужки и зрачка.

2 Постановка задачи аппроксимации параметров окружности

Задано бинарное изображение

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

где 1 — это черный пиксель, который принадлежит рассматриваемой фигуре на изображении, а 0 — белый пиксель, который является фоном изображения. Пример изображения показан на рис. 1. Изображение $\mathbf{M}$ отображается в множество координат $\mathbf{C} = \{x_i, y_i\}_{i=1}^{N}$. Координата (x_i, y_i) является координатой i-го черного пикселя на изображении $\mathbf{M}$:

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

где N — число черных пикселей.

Обозначим точку (x_0, y_0) центром окружности, а r радиусом окружности. Координаты $(x_i, y_i) \in \mathbf{C}$ это геометрическое место точек, которое удовлетворяет системе уравнений:

$$\big(x_i - x_0\big)^2 + \big(y_i - y_0\big)^2 = r^2 + \varepsilon_i, \qquad \forall i \in \{1, \cdots, N\},$$

где $\varepsilon_i \in \mathcal{N}\big(0, \beta^{-1}\big)$ является невязкой i-го уравнения, которая является следствием шума на изображении.

Раскрыв скобки получаем:

$$(2x_0) \cdot x_i + (2y_0) \cdot y_i + \big(r^2 - x_0^2 - y_0^2\big) \cdot 1 = x_i^2 + y_i^2 + \varepsilon_i. \tag{2.1}$$

Выражение (2.1) переписывается в задачу линейной регрессии следующим образом:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbf{R}^n} ||\mathbf{X}\mathbf{w} - \mathbf{y}||, \quad \mathbf{X} = [\mathbf{C}, \mathbf{1}], \quad \mathbf{y} = \left[x_1^2 + y_1^2, x_2^2 + y_2^2, \cdots, x_N^2 + y_N^2\right]^{\mathsf{T}}. \tag{2.2}$$

Используя вектор параметров $\hat{\mathbf{w}} = [w_1, w_2, w_3]^{\mathsf{T}}$ получим параметры окружности x_0, y_0, r:

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

Решение уравнения (2.2) находит параметры единственной окружности на изображении. В случае, когда на изображении несколько окружностей, предлагается использовать смесь экспертов, которая состоит из линейных модели — экспертов. Каждый эксперт описывает одну окружность на изображении.

3 Постановка задачи построения смеси экспертов

Обобщим подход аппроксимации одной окружности на изображении на случай, когда на изображении несколько окружностей. Пусть изображение состоит из K окружностей, тогда множество черных пикселей $\mathbf{C}$ представляется в виде:

$$\mathbf{C} = \sqcup_{k=1}^{K} \mathbf{C}'_k,$$

где $\mathbf{C}'_k$ множество точек принадлежащих k-й окружности. Множеству точек $\mathbf{C}'_k \subset \mathbf{C}$ соответсвует задача линейной регрессии для выборки $\mathbf{X}'_k \subset \mathbf{X}, \mathbf{y}'_k \subset \mathbf{y}$. Модель $\mathbf{g}_k$ аппроксимирующая выборку $\mathbf{X}'_k, \mathbf{y}'_k$ является локальной моделью для выборки $\mathbf{X}$, $\mathbf{y}$.

Определение 3.1. *Модель* $\mathbf{g}$ *называется локальной моделью для выборки* $\boldsymbol{U}$, *если* $\mathbf{g}$ *аппроксимирует некоторое не пустое подмножество* $\boldsymbol{U}' \subset \boldsymbol{U}$.

Определение 3.2. *Мультимодель* $\mathbf{f}$ *называется смесью экспертов, если:*

$$\mathbf{f} = \sum_{k=1}^{K} \pi_k \mathbf{g}_k\big(\mathbf{w}_k\big), \qquad \pi_k\big(\mathbf{x}, \mathbf{V}\big) : \mathbb{R}^{n \times |\mathbf{V}|} \to [0,1], \qquad \sum_{k=1}^{K} \pi_k\big(\mathbf{x}, \mathbf{V}\big) = 1, \tag{3.1}$$

где $\mathbf{g}_k$ *является* k*-й локальной моделью,* π_k *— шлюзовая функция, вектор* $\mathbf{w}_k$ *является параметрами* k*-й локальной моделью, а* $\mathbf{V}$ *— параметры шлюзовой функции.*

В данной работе в качестве локальных моделей рассматриваются линейные модели. В качестве шлюзовой функции рассматривается двухслойный перцептрон:

$$\mathbf{g}_k\big(\mathbf{x}\big) = \mathbf{w}_k^{\mathsf{T}}\mathbf{x}, \quad \boldsymbol{\pi}\big(\mathbf{x}, \mathbf{V}\big) = \operatorname{softmax}\big(\mathbf{V}_1^{\mathsf{T}}\boldsymbol{\sigma}\big(\mathbf{V}_2^{\mathsf{T}}\mathbf{x}\big)\big), \tag{3.2}$$

где $\mathbf{V} = \big\{\mathbf{V}_1, \mathbf{V}_2\big\}$ — множество параметров шлюзовой функции.

В статье предлагается использовать вероятностный подход для описания смеси экспертов. Вводиться предположение, что $\mathbf{y}$ является случайным вектором, который задается плотностью распределения $p\big(\mathbf{y}|\mathbf{X}\big)$. Предполагается, что плотность распределения $p\big(\mathbf{y}|\mathbf{X}, \mathbf{f}\big)$ аппроксимирует истинную плотность распределения $p\big(\mathbf{y}|\mathbf{X}\big)$:

$$p\big(\mathbf{y}|\mathbf{X}, \mathbf{f}\big) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k p_k\big(y_i|\mathbf{g}_k\big(\mathbf{x}_i\big)\big) \right), \tag{3.3}$$

где $\mathbf{f}$ — это смесь экспертов, а $\mathbf{g}_k, \boldsymbol{\pi}$ определяются выражением (3.2).

Пусть $\mathbf{w}_k$ является случайным вектором, который задается плотностью распределения $p^k\big(\mathbf{w}_k\big)$. Получим совместное распределения параметров локальных моделей и вектора ответов:

$$p\big(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\big) = \prod_{k=1}^{K} p^k\big(\mathbf{w}_k\big) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k p_k\big(y_i|\mathbf{w}_k, \mathbf{x}_i\big) \right), \tag{3.4}$$

где $\mathbf{W} = \big\{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K\big\}$. Оптимальные параметры находятся при помощи максимизации правдоподобия:

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

4 Вероятностная постановка смеси экспертов

Для построения смеси экспертов (3.1, 3.4), введем следующие вероятностные предположения о данных (2.2):

1) правдоподобие $p_k\big(y_i|\mathbf{w}_k, \mathbf{x}_i\big) = \mathcal{N}\big(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\big)$, где параметр β является уровнем шума,

2) априорное распределение параметров $p^k\big(\mathbf{w}_k\big) = \mathcal{N}\big(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\big)$, где $\mathbf{w}_k^0$ — вектор размерности $n \times 1$, а $\mathbf{A}_k$ — ковариационная матрица размерности $n \times n$,

3) регуляризация априорного распределения $p\big(\boldsymbol{\varepsilon}_{k,k'}|\boldsymbol{\Xi}\big) = \mathcal{N}\big(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\big)$, где $\boldsymbol{\Xi}$ — ковариационная матрица, а $\boldsymbol{\varepsilon}_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0$.

Предположение 2) задает априорное предположения на вектора параметров локальных модели $\mathbf{w}_k$. Априорное распределение задает ограничения на локальную модель. Например, если $\mathbf{w}_k^0 = [0, 0, 1]$, то k-я локальная модель аппроксимирует окружность с параметрами $x_0 = 0, y_0 = 0, r = 1$ с большей вероятностью.

Предположения 3) задает регулярицию априорных распределений. Данная регулярицая учитывает связь между априорными ограничениями разных локальных моделей. Например, если $\operatorname{diag}\big(\boldsymbol{\Xi}\big) = [0.001, 0.001, 1]$, то центры разных окружностей совпадают.

Используя предположения $1), 2), 3)$ и выражение (3.4) получаем полное правдоподобие:

$$\begin{aligned} p\big(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) = &\prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\big(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\big) \right) \cdot \\ &\cdot \prod_{k=1}^{K} \mathcal{N}\big(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\big) \cdot \prod_{k,k'=1}^{K} \mathcal{N}\big(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\big), \end{aligned} \tag{4.1}$$

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

Введем бинарную матрицу $\mathbf{Z}$. Элемент матрицы z_{ik} равно 1 тогда и только тогда, когда i-й объект аппроксимируется k-й локальной моделью. Подставляя бинарную матрицу $\mathbf{Z}$ в выражении (4.1), а также взяв логарифм получаем:

$$\begin{aligned}
\log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) = \\
= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[\log \pi_k\big(\mathbf{x}_i, \mathbf{V}\big) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\
+ \sum_{k=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] + \\
+ \sum_{k=1}^{K}\sum_{k'=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}}\boldsymbol{\Xi}^{-1}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2}\log\det\boldsymbol{\Xi} - \frac{n}{2}\log 2\pi\right].
\end{aligned} \tag{4.2}$$

Получаем новую задачу оптимизации обоснованности. Функция обоснованности получается при интегрировании выражения (4.2) по параметрам $\mathbf{W}, \mathbf{Z}$:

$$\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta = \arg\max_{\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta}\int_{\mathbf{W},\mathbf{Z}} \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big)d\mathbf{W}d\mathbf{Z}. \tag{4.3}$$

5 EM–алгоритм для решения задачи опимизации

Рассмотрим вариационную плотность $q\big(\mathbf{W}, \mathbf{Z}\big)$ для параметров $\mathbf{W}, \mathbf{Z}$. Тогда функция обоснованности принимает следующий вид:

$$\begin{aligned}
\log p\big(\mathbf{y}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) &= \int_{\mathbf{W},\mathbf{Z}} q\big(\mathbf{W}, \mathbf{Z}\big)\log p\big(\mathbf{y}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big)d\mathbf{W}d\mathbf{Z} = \\
&= \int_{\mathbf{W},\mathbf{Z}} q\big(\mathbf{W}, \mathbf{Z}\big)\log\frac{p\big(\mathbf{y}, \mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big)}{p\big(\mathbf{W}, \mathbf{Z}|\mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big)}d\mathbf{W}d\mathbf{Z} = \\
&= \int_{\mathbf{W},\mathbf{Z}} q\big(\mathbf{W}, \mathbf{Z}\big)\log\frac{p\big(\mathbf{y}, \mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big)q\big(\mathbf{W}, \mathbf{Z}\big)}{p\big(\mathbf{W}, \mathbf{Z}|\mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big)q\big(\mathbf{W}, \mathbf{Z}\big)}d\mathbf{W}d\mathbf{Z} = \\
&= \int_{\mathbf{W},\mathbf{Z}} q\big(\mathbf{W}, \mathbf{Z}\big)\frac{p\big(\mathbf{y}, \mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big)}{q\big(\mathbf{W}, \mathbf{Z}\big)}d\mathbf{W}d\mathbf{Z} + \\
&+ \int_{\mathbf{W},\mathbf{Z}} q\big(\mathbf{W}, \mathbf{Z}\big)\frac{q\big(\mathbf{W}, \mathbf{Z}\big)}{p\big(\mathbf{W}, \mathbf{Z}|\mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big)}d\mathbf{W}d\mathbf{Z} = \\
&= \mathcal{L}\big(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\big) + \mathsf{D}_{KL}\left(q\big(\mathbf{W}, \mathbf{Z}\big)||p\big(\mathbf{W}, \mathbf{Z}|\mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big)\right)
\end{aligned} \tag{5.1}$$

Используя (5.1) получаем нижнюю оценку обоснованности:

$$\log p\big(\mathbf{y}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) \geq \mathcal{L}\big(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\big),$$

где $\mathcal{L}\big(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\big)$ называется нижней оценкой обоснованности.

Используем EM–алгоритм [22, 23] для решения оптимизационной задачи (4.3). Заметим, что EM–алгоритм вместо оптимизации $\log p\big(\mathbf{y}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\boldsymbol{\Xi},\beta\big)$ оптимизирует нижнюю оценку $\mathcal{L}\left(q,\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta\right)$.

E-шаг. E-шаг решает следующую оптимизационую задачу:

$$\mathcal{L}\big(q,\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta\big) \to \max_{q\left(\mathbf{w},\mathbf{z}\right)},$$

где параметры $\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta$ являются зафиксированными.

Пусть совместное распределение $q\big(\mathbf{Z},\mathbf{W}\big)$ удовлетворяет условию независимости $q\big(\mathbf{Z},\mathbf{W}\big) = q\big(\mathbf{Z}\big)q\big(\mathbf{W}\big)$ [23]. Далее символом $\propto$ обозначим то, что обе стороны выражения равны с точностью до аддитивной константы. Сначала найдем распределение $q\big(\mathbf{Z}\big)$:

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

Используя выражения (5.2) получаем, что распределение $q\big(z_{ik}\big)$ является бернулевским распределением с параметром z_{ik}, которое задается выражением (5.2). Далее найдем распределение $q\big(\mathbf{W}\big)$:

$$\begin{aligned}
\log q\big(\mathbf{W}\big) &= \mathsf{E}_{q/\mathbf{W}} \log p\big(\mathbf{y},\mathbf{Z},\mathbf{W}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\boldsymbol{\Xi},\beta\big) \propto \\
&\propto \sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{E}z_{ik}\left[\log\pi_k\big(\mathbf{x}_{i,\mathbf{V}}\big) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^\mathsf{T}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\
&+ \sum_{k=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^\mathsf{T}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] \\
&\propto \sum_{k=1}^{K}\left[\mathbf{w}_k^\mathsf{T}\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0 + \beta\sum_{i=1}^{N}\mathbf{x}_i y_i \mathsf{E}z_{ik}\right) - \frac{1}{2}\mathbf{w}_k^\mathsf{T}\left(\mathbf{A}_k^{-1} + \beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right)\mathbf{w}_k\right].
\end{aligned} \tag{5.3}$$

Используя выражение (5.3) получаем, что распределение $q\big(\mathbf{w}_k\big)$ является нормальным распределением со средним $\mathbf{m}_k$ и ковариационной матрицей $\mathbf{B}_k$:

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

M-шаг. M-шаг решает следующую оптимизационную задачу:

$$\mathcal{L}\big(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\big) \to \max_{\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta},$$

где $q\big(\mathbf{W}, \mathbf{Z}\big)$ является известной плотностью распределения. Распределение $q\big(\mathbf{Z}, \mathbf{W}\big)$ является фиксированным, в то время как вариацонная нижняя оценка $\mathcal{L}\big(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\big)$ максимизируется по параметрам $\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta$:

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

Во-первых, для нахождения оптимального параметра $\mathbf{V}$ используется градиентный метод оптимизации, который сходится к некоторому локальному экстремуму. Во вторых, используя выражения (5.4) получаем оптимальное значения параметра $\mathbf{A}_k$

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

Аналогично получаем оптимальные значения для параметра β и для параметров $\mathbf{w}_k^0$

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

$$\begin{aligned}
\frac{\partial \mathcal{L}\big(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\big)}{\partial \mathbf{w}_k^0} &= \mathbf{A}_k^{-1} \left(\mathsf{E}\mathbf{w}_k - \mathbf{w}_k^0 \right) + \boldsymbol{\Xi} \sum_{k'=1}^{K} \left[\mathbf{w}_{k'}^0 - \mathbf{w}_k^0 \right] = 0, \\
\mathbf{w}_k^0 &= \left[\mathbf{A}_k^{-1} + (K-1)\, \boldsymbol{\Xi} \right]^{-1} \left(\mathbf{A}_k^{-1} \mathsf{E}\mathbf{w}_k + \boldsymbol{\Xi} \sum_{k'=1,\ k' \neq k}^{K} \mathbf{w}_{k'}^0 \right).
\end{aligned} \tag{5.5}$$

Выражения (5.2–5.5) задают итеративную процедуру, которая сходится к некоторому локальному максимуму оптимизационной задачи (4.3).

6 Вычислительный эксперимент

Для анализа качества различных мультимоделей для аппроксимации окружности проводится вычислительный эксперимент. В эксперимент рассматриваются следующие мультимодели: мультимодель $\mathbf{f}_1$ без использования априорных распределений, мультимодель $\mathbf{f}_2$, которая использует априорные распределения (6.2) для параметров и мультимодель $\mathbf{f}_3$, которая использует регуляризацию априорных распределений. Точность аппроксимации мультимодели $\mathbf{f}_i$ задается следующим образом:

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

где x_0^k, y_0^k, r^k является истинным центром и радиусом для k-й окружности, $x_{\text{pr}}^k, y_{\text{pr}}^k, r_{\text{pr}}^k$ является предсказанным центром и радиусом для k-й окружности.

Для сравнение модель с разными вероятностными предположениями используется правдоподобие (3.3). В вычислительном эксперименте используется следующее априорное распределение:

$$p^1\big(\mathbf{w}_1\big) \sim \mathcal{N}\big(\mathbf{w}_1^0, \mathbf{I}\big), \quad p^2\big(\mathbf{w}_2\big) \sim \mathcal{N}\big(\mathbf{w}_2^0, \mathbf{I}\big), \tag{6.2}$$

где $\mathbf{w}_1^0 = [0, 0, 0.1],\ \mathbf{w}_2^0 = [0, 0, 2]$.

Синтетические данные с разным типом шума в изображении. В вычислительном эксперименте сравнивается качество следующих мультимоделей $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$ на синтетических данных. Синтетические данные являются двумя концентрическими окружностями с разным уровнем шума. Выборка Synthetic 1 является изображением без шума, выборка Synthetic 2 изображение с зачумлённым радиусом окружности, а выборка Synthetic 3 — изображение с равномерным шумом. На рис. 2 показаны результаты для мельтимоделей $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$. Все модели оптимизировались при помощи 50 итераций EM-алгоритма. Мультимодели $\mathbf{f}_2, \mathbf{f}_3$ аппроксимируют окружности лучше чем мультимодель $\mathbf{f}_1$. В табл. 1 показано качество аппрроксимации (6.1) для всех мультимоделей.

Таблица 1: Качество аппроксимации (6.1) для всех мультимоделей

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

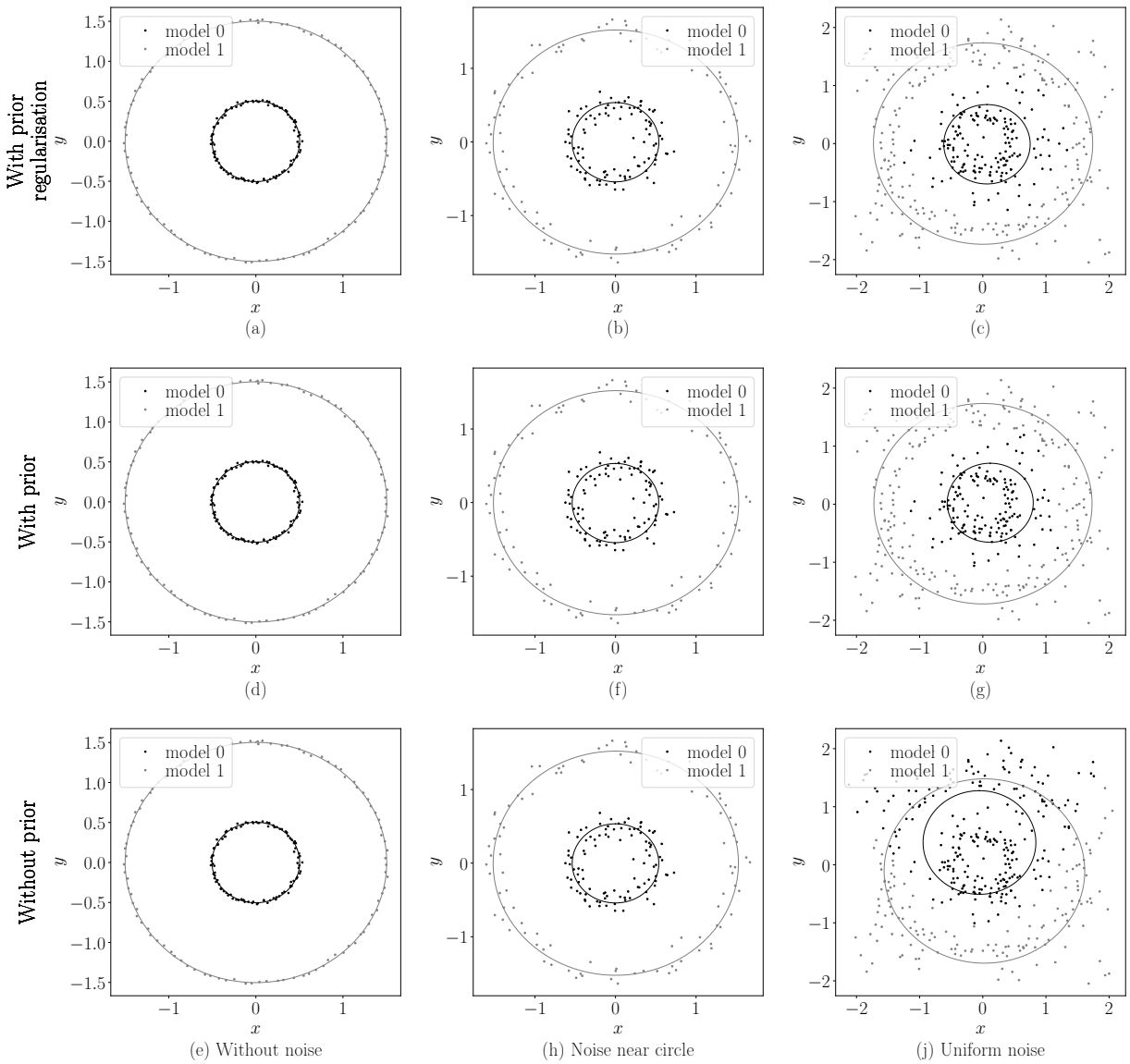


Рис. 2: Мультимодель в зависимости от разных априорных предположений и в зависимости от разного уровня шума: (a)–(c) модель с регуляризаций априорных распределений; (d)–(g) модель с заданными априорными распределениями на параметрах локальных моделей; (e)–(j) модель без заданных априорных предположений

Анализ сходимости на синтетической выборке. Данная часть эксперимента анализирует качество сходимости EM-алгоритма для разных мультимоделей $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$. Анализ всех мультимоделей проводиться на выборке Synthetic 3.

На рис. 3 показана зависимость предсказано центра и радуса в зависимости от номера итерации EM-алгоритма. Мультимодель $\mathbf{f}_2$, которая использует априорное распределение аппроксимирует окружность лучше мультимодели $\mathbf{f}_1$, которая не ис-

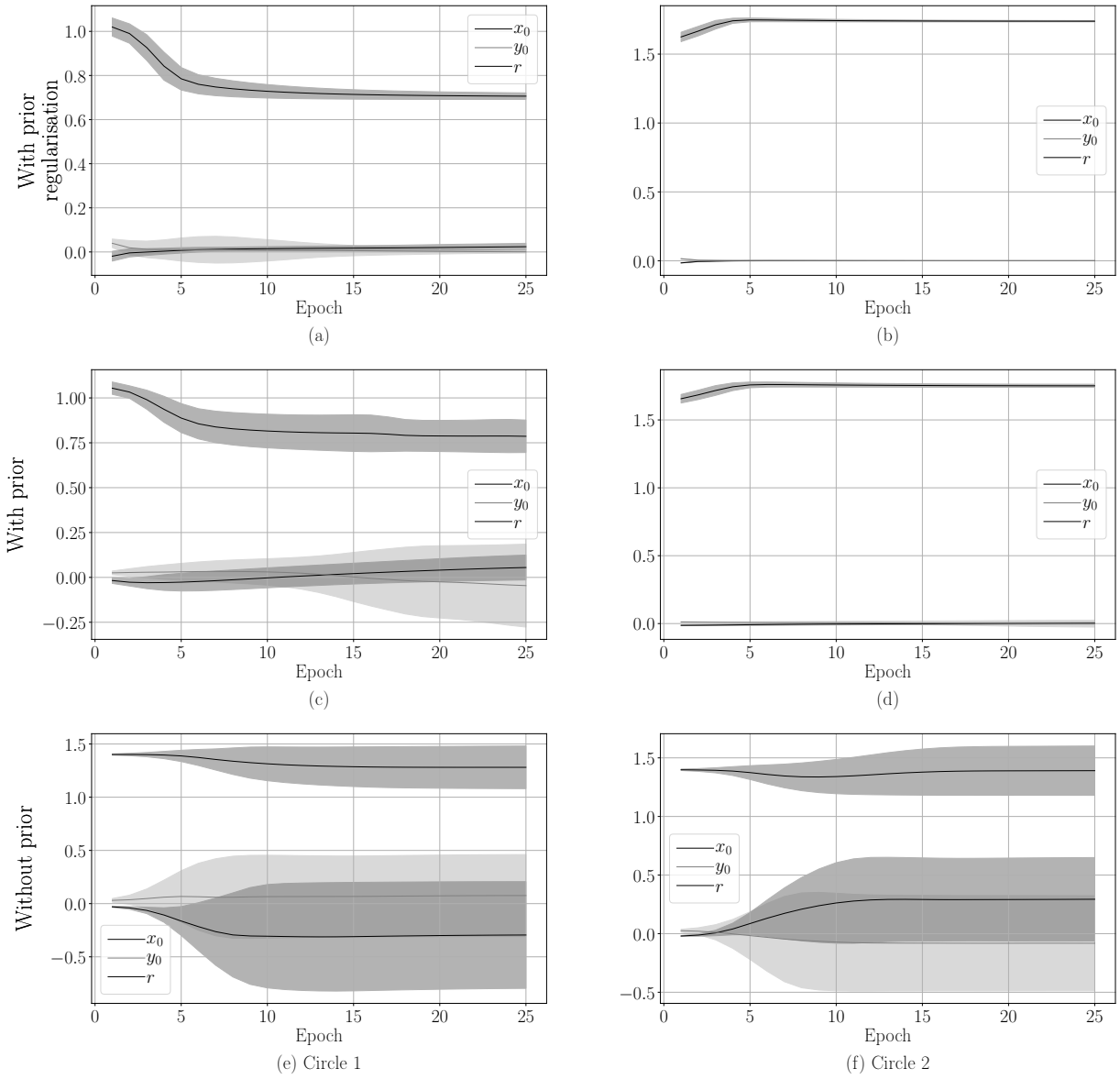


Рис. 3: График зависимости центра и радиуса окружностей от номера итерации: (a)–(b) модель с регуляризацией априорных распределений; (c)–(d) модель с заданными априорными распределениями на параметры моделей; (e)–(f) модель без задания априорных распределений

пользует никакого априорного распределения. Мультимодель $\mathbf{f}_3$, которая использует регуляризатор априорных распределений является более стабильной, чем мультимодель $\mathbf{f}_2$.

На рис. 4 показана зависимость логарифма правдоподобия (3.3) от номера итерации EM-алгоритма. Логарифм правдоподобия мультимодели $\mathbf{f}_2, \mathbf{f}_3$ растет быстрее чем логарифм правдоподобия мультимодели $\mathbf{f}_1$. После 20-й итерации все мультимодели

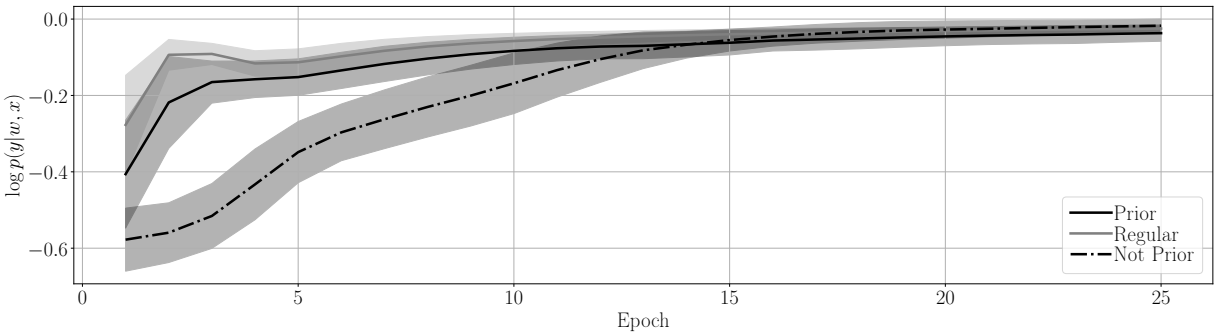


Рис. 4: График зависимости логарифма правдоподобия (3.3) от номера итерации.

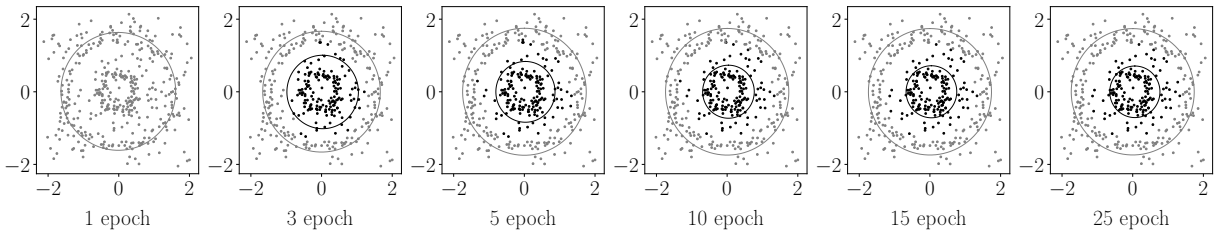


Рис. 5: Визуализации процесса сходимости мультимодели с использованием априорной регуляризации.

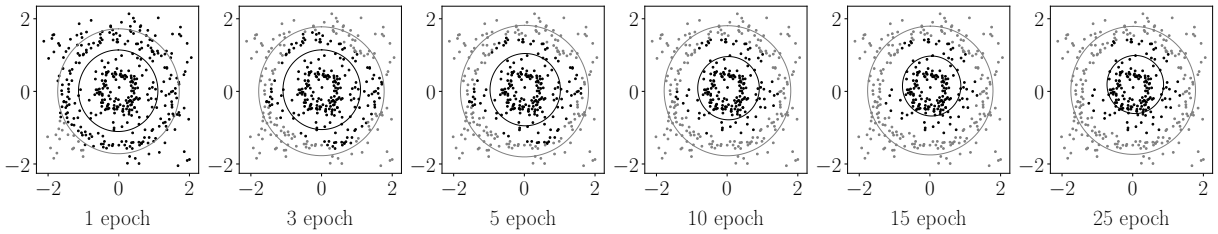


Рис. 6: Визуализации процесса сходимости мультимодели с использованием априорного распределением.

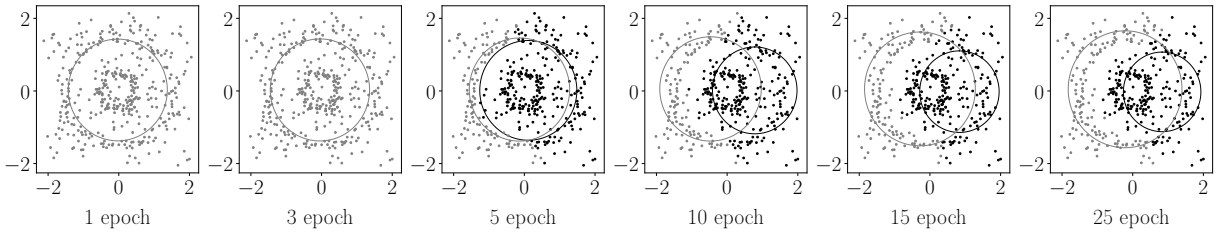


Рис. 7: Визуализации процесса сходимости мультимодели без использования априорного распределения.

имеют одинаковое правдоподобие.

На рис. 5-7 показан процесс сходимости для разных мультимоделей $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$. На рис. 7 показана мультимодель $\mathbf{f}_1$, которая аппроксимирует окружности не верно. На рис. 5-6 показаны мультимодели $\mathbf{f}_2, \mathbf{f}_3$, которые аппроксимируют окружности верно.

Вычислительный эксперимент показывает, что мультимодели $\mathbf{f}_2, \mathbf{f}_3$, которые используют априорные распределения на параметры экспертов аппроксимируют окружности лучше чем мультимодель $\mathbf{f}_1$, которая работает без априорных распределений.

Анализ мультимоделей в зависимости от уровня шума. Данная часть эксперимента анализирует зависимость разных мультимоделей $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$ от уровня шума. Анализ всех мультимоделей проводиться на выборке Synthetic 1, с добавлением разного уровня шума. Минимальный уровень шума равен 0, когда числа шумовых точек равно 0. Максимальный уровень шума равен 1, когда число шумовых точек равно числу точек на изображении. На рис. 8 показан график зависимости центра окружности и ее радиус в зависимости от уровня шума. Из графика видно, что радиус окружности увеличивается при увеличении уровня шума. Мультимодели $\mathbf{f}_2, \mathbf{f}_3$ аппроксимируют центр окружности верно, но мультимодель $\mathbf{f}_3$ более устойчива к шуму . На рис. 9 показана зависимость логарифма правдоподобия (3.3) от уровня шума. Из графика видно, что логарифм правдоподобия (3.3) эквивалентный для всех мультимоделей, но на рис. 8 видно, что качество аппроксимации (6.1) зависит от мультимодели. Данная часть вычислительного эксперимента показывает, что мультимодель $\mathbf{f}_3$ с регуляризацией априорного распределения является более устойчива к шуму, чем остальные.

Реальные данные. Данная часть эксперимента анализирует разные мультимодели $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$ на реальной выборке. На рис. 10 показан результат работы разных мультимоделей. Мультимодель $\mathbf{f}_1$ не верно аппроксимирует меньшую окружность. Мультимодели $\mathbf{f}_2, \mathbf{f}_3$ аппроксимируют обе окружности верно.

На рис. 11-13 показан процесс аппроксимации для разных мультимоделей $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$.

Данная часть эксперимента показывает, что мультимодели $\mathbf{f}_2, \mathbf{f}_3$ аппроксимируют окружности на реальных изображениях лучше, чем мультимодель $\mathbf{f}_1$.

7 Заключение

В данной работе сравниваются мультимодели, которые используют различные априорные предположения. Для анализа проводился вычислительный эксперимент на концентрических окружностях с разным уровнем шума. Для аппроксимации окружности на изображении использовалась линейная модель. Для взвешивания ответов разных линейных моделей использовалась шлюзовая функция, которая является двухслойным перцептрон с функцией softmax на последнем слое. В вычислительном эксперименте сравниваются мультимодели, которые используют априорное распределение и которые не используют. Мультимодели, которые используют априорные распределения имеют большую точностью аппроксимации чем мультимодель, которая не использует априорные распределения.

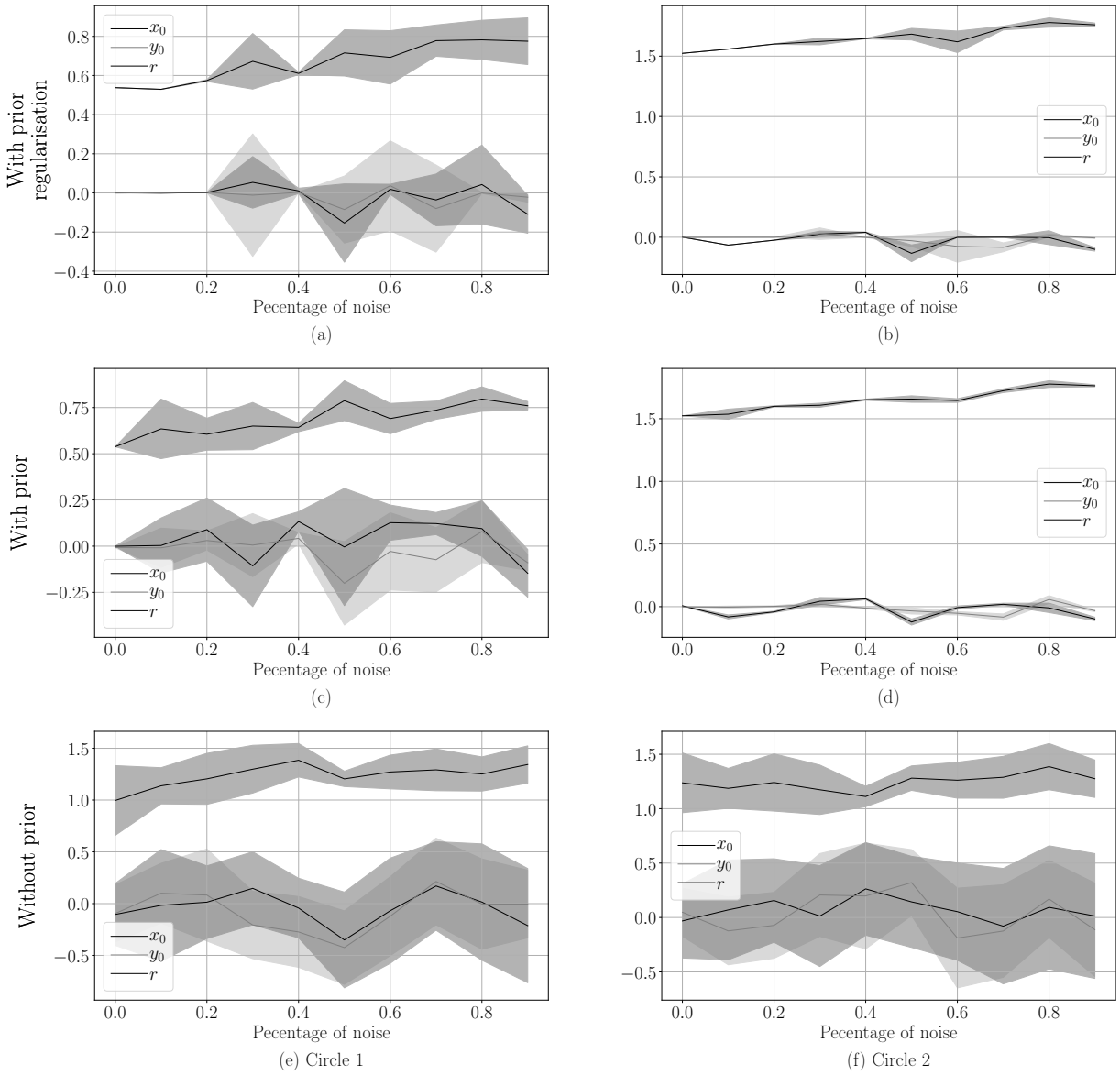


Рис. 8: рафик зависимости центра и радиуса окружностей от номера итерации: (a)–(b) модель с регуляризацией априорных распределений; (c)–(d) модель с заданными априорными распределениями на параметры моделей; (e)–(f) модель без задания априорных распределений

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 эксперименте было показано, что все мультимодели являются чувствительными к выбросам. Для решения данной задачи предлагается использовать еще одну локальную модель, которая будет аппроксимировать шум.

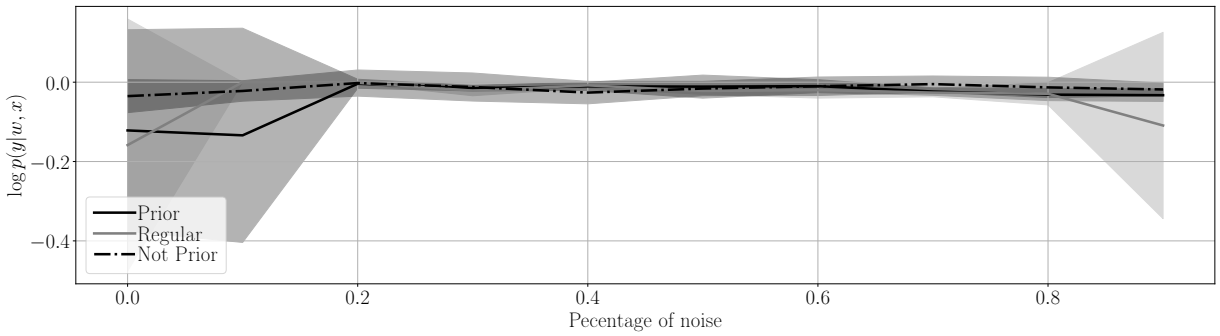


Рис. 9: График зависимости логарифма правдоподобия (3.3) от уровня шума.

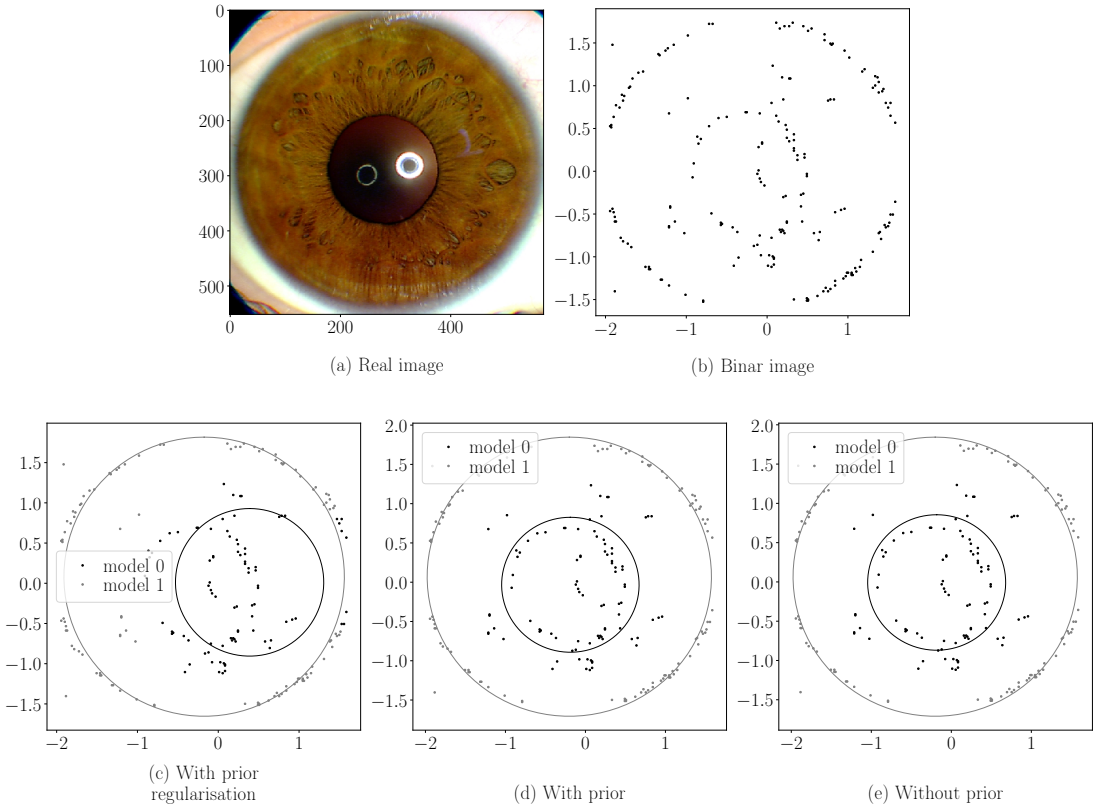


Рис. 10: Мультимодель в зависимости от разных априорных предположений на реальном изображении: (a) исходное изображение; (b) бинаризованое изображение; (c) мультимодель без априорных предположений; (d) мультимодель с априорными распределениями на параметрах локальных моделей; (e) мультимодель с регуляризаций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

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 аппрок-

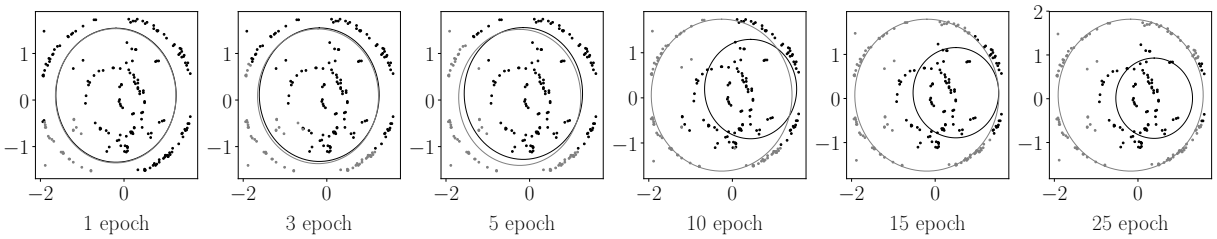


Рис. 11: Визуализации процесса сходимости мультимодели без использования априорного распределения.

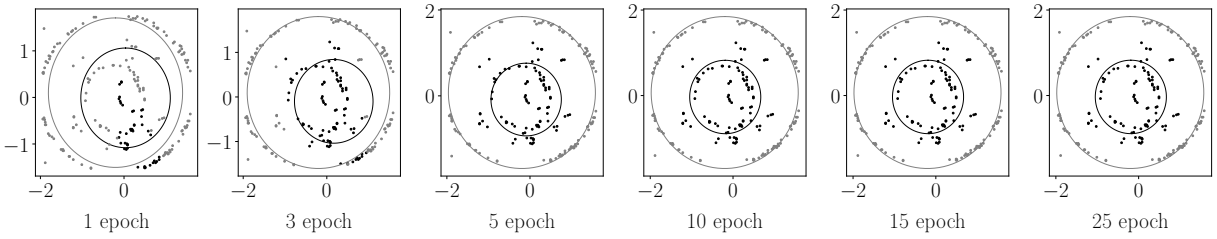


Рис. 12: Визуализации процесса сходимости мультимодели с использованием априорного распределением.

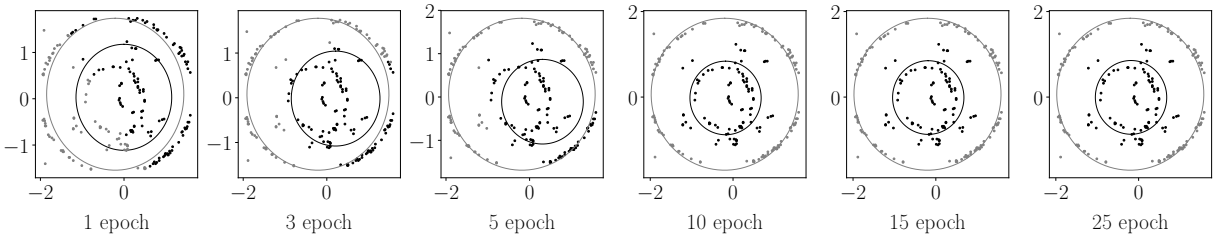


Рис. 13: Визуализации процесса сходимости мультимодели с использованием априорной регулярициии.

симирует шум в данных. Предполагается, что число шумовых точек мало, поэтому требуется задать априорное распределение, которое учитывает данную информацию.

Список литературы

[1] *Tianqi C., Carlos G.* XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

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

[3] *Esen Y. S., Wilson J., Gader P. D.* Twenty Years of Mixture of Experts // IEEE Transactions on Neural Networks and Learning Systems. 2012. Issues. 23. No. 8. P. 1177–1193.

[4] *Rasmussen C. E., Ghahramani Z.* Infinite Mixtures of Gaussian Process Experts // Advances in Neural Information Processing Systems 14. 2002. P. 881–888.

[5] *Shazeer N., Mirhoseini A., Maziarz K.* Outrageously large neural networks: the sparsely-gated mixture-of-experts layer // International Conference on Learning Representations. 2017.

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

[7] *Jordan M. I., Jacobs R. A.* Hierarchies of adaptive experts // In Advances in Neural Information Processing Systems. 1991. P. 985–992.

[8] *Lima C., Coelho A., Zuben F. J.* Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification // Inf. Sci. 2007. Vol. 177. No. 10. P. 2049–2074.

[9] *Cao L.* Support vector machines experts for time series forecasting // Neurocomputing. 2003. Vol. 51. P. 321–339.

[10] *Yumlu M. S., Gurgen F. S., Okay N.* Financial time series prediction using mixture of experts // In Proc. 18th Int. Symp. Comput. Inf. Sci. 2003. P. 553–560.

[11] *Cheung Y. M., Leung W. M., Xu L.* Application of mixture of experts model to financial time series forecasting // On Proc. Int. Conf. Neural Netw. Signal Process. 1995. P. 1–4.

[12] *Weigend A. S., Shi S.* Predicting daily probability distributions of S&P500 returns // J. Forecast. 2000. Vol. 19. No. 4. P. 375–392.

[13] *Ebrahimpour R., Moradian M. R., Esmkhani A., Jafarlou F. M.* Recognition of Persian handwritten digits using characterization loci and mixture of experts // J. Digital Content Technol. Appl. 2009. Vol. 3. No. 3. P. 42–46.

[14] *Estabrooks A., Japkowicz N.* A mixture-of-experts framework for text classification //In Proc. Workshop Comput. Natural Lang. Learn., Assoc. Comput. Linguist. 2001. P. 1–8.

[15] *Mossavat S., Amft O., Petkov Vries B., Kleijn W.* A Bayesian hierarchical mixture of experts approach to estimate speech quality // In Proc. 2nd Int. Workshop Qual. Multimedia Exper. 2010. P. 200–205.

[16] *Peng F., Jacobs R. A., Tanner M. A.* Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition // J. Amer. Stat. Assoc. 1996. Vol. 91. No. 435. P. 953–960.

[17] *Tuerk A.* The state based mixture of experts HMM with applications to the recognition of spontaneous speech. Ph.D. thesis. Cambridge: Univ. Cambridge, 2001.

[18] *Sminchisescu C., Kanaujia A., Metaxas D.* Discriminative density propagation for visual tracking // IEEE Trans. Pattern Anal. Mach. Intell. 2007. Vol. 29. No. 11. P. 2030–2044.

[19] *Bowyer K., Hollingsworth K., Flynn P.* A Survey of Iris Biometrics Research: 2008–2010.

[20] *Matveev I.* Detection of iris in image by interrelated maxima of brightness gradient projections // Appl.Comput. Math. 2010. Vol. 9. No. 2. P. 252–257.

[21] *Matveev I., Simonenko I.*. Detecting precise iris boundaries by circular shortest path method // Pattern Recognition and Image Analysis. 2014. Vol. 24. P. 304–309.

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

[23] *Bishop C.* Pattern Recognition and Machine Learning. Berlin: Springer, 2006. P. 758.