

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 зависимиыми. 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 результаты других моделей для конечного результата, причем коэффициенты взвешивания зависят от 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

Мультимодели показывают отличные результаты во многих задачах. Классическими методами являются беггинг и градиентный бустинг [1], случайный лес [2]. Подход к мультимоделированию [3] предполагает, что вклад каждой модели в ответ должен зависеть от конкретного объекта. Смесь экспертов базируется на понятии шлюзовой функции, которая определяет значимость предсказания каждого эксперта — отдельной модели.

Для поиска оптимальных параметров рассматривается вероятностная постановка задачи. В качестве функционала качества рассматривается логарифм правдоподобия модели. Для оптимизации данного функционала рассматривается ЕМ-алгоритм.

Несмотря на значимые успехи мультимodelей, они имеют ряд недостатков. Данные недостатки связаны с тем, что сходимость локальных моделей сильно зависит от начальной инициализации векторов параметров. Для улучшения сходимости предлагается использовать априорные знание о данных.

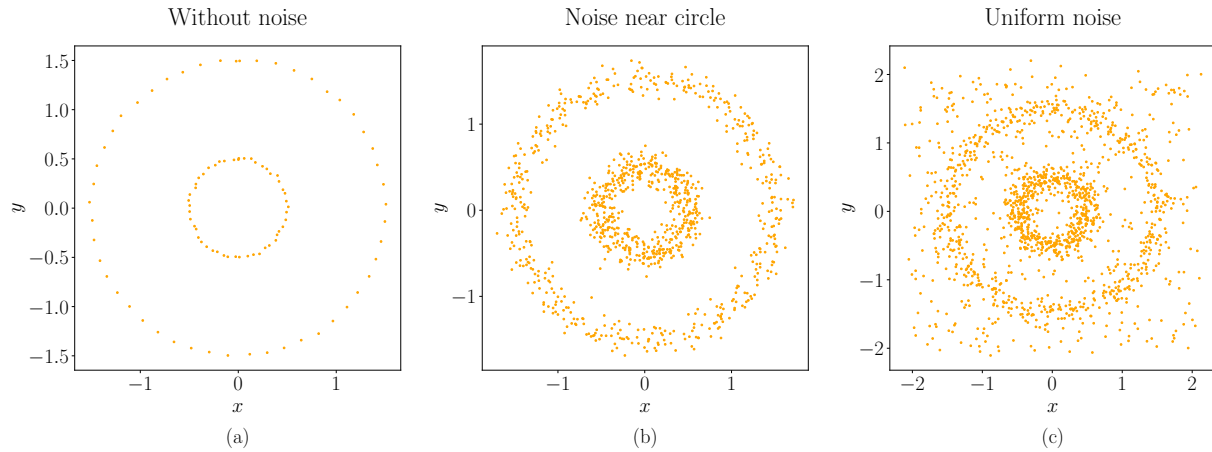


Рис. 1: Пример изображений с окружностями с разным уровнем шума: а) окружности без шума; б) окружности с зашумленным радиусом; в) окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению

Данная работа исследует, влияние выбора априорного распределения параметров моделей на сходимость мультимodelей. Решается задача поиска окружностей на би-наризованном изображении. Предполагается, что радиусы окружностей различаются значимо. 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 softmax на последнем слое.

Ряд работ посвящены выбору моделей в качестве отдельных экспертов. В работах [6, 9] в качестве модели эксперта рассматривается линейная модель — гауссиана. Работы [7, 8] рассматривают модель SVM в качестве модели эксперта.

В работе [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:

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

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

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 — количество 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. Тогда элементы выборки \mathbf{C} должны удовлетворять уравнению окружности:

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

Раскрыв скобки получим следующие 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:

$$\mathbf{X}\mathbf{w} \approx \mathbf{Y}, \quad \mathbf{X} = \mathbf{C} \times \mathbf{1}, \quad \mathbf{Y} = \{x^2 + y^2 \mid \forall x, y \in \mathbf{C}\}, \quad (3.5)$$

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

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

Данное решение позволяет искать параметры единственной окружности на рисунке. В случае, когда на картинке несколько окружностей предлагается использовать мультимодель, где в качестве каждой модели рассматривается единственная линейная модель, которая отвечает одной окружности на рисунке. В качестве мультимодели рассматривается смесь экспертов.

4 Постановка задачи построения смеси экспертов

Задана выборка

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

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

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

$$\hat{\mathbf{f}} = \sum_{k=1}^K \pi_k \mathbf{f}_k, \quad \pi_k(\mathbf{x}, \mathbf{V}) : \mathbb{R}^{2 \times n} \rightarrow [0, 1], \quad \sum_{k=1}^K \pi_k(\mathbf{x}, \mathbf{V}) = 1 \quad (4.2)$$

где $\hat{\mathbf{f}}$ — мультимодель, а \mathbf{f}_k является некоторой моделью, π_k — параметрическая модель.

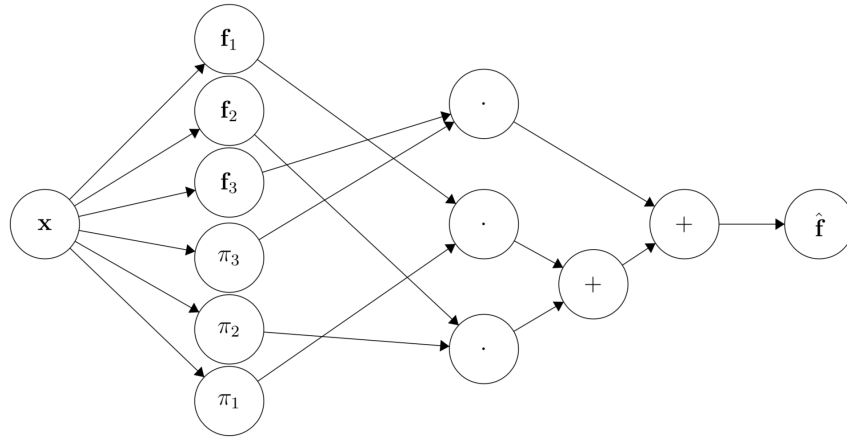


Рис. 2: Схема смеси экспертов, в случае трех локальных моделей: \mathbf{f}_k — локальные модели, π_k — плюсовая функция, $\hat{\mathbf{f}}$ — мультимодель. Каждая стрелка обозначает то, что объект от которого исходит стрелка подставляется в ту функция, где стрелка заканчивается.

На рис. 2 показан пример мультимодели, которая изображена при помощи графовой нотации. В данном примере число локальных моделей \mathbf{f}_k равняется трем.

4.1 Смесь экспертов с prior на моделях

В данной задаче рассматривается случай линейных локальных моделей:

1. $p_k(y_i | \mathbf{w}_k, \mathbf{x}_i) = N(y_i | \mathbf{w}_k^\top \mathbf{x}_i, \beta^{-1})$
2. $p^k(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k) = N(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k)$
3. $\boldsymbol{\pi}(\mathbf{x}_i, \mathbf{V}) = \text{softmax}(\mathbf{F}(\mathbf{x}_i, \mathbf{V}))$, где $F : \mathbb{R}^n \times \mathbb{R}^V \rightarrow \mathbb{R}^K$ — нейросеть, V — число параметров нейросети.

С учетом наших предположений логарифм правдоподобия переписывается в следующем виде:

$$\begin{aligned} \log p(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}) &= \sum_{i=1}^N \sum_{k=1}^K z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (y_i - \mathbf{w}_k^\top \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\ &+ \sum_{k=1}^K \left[-\frac{1}{2} (\mathbf{w}_k - \mathbf{w}_k^0)^\top \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] \end{aligned} \quad (4.3)$$

Воспользуемся ЕМ-алгоритмом для решения оптимизационной задачи. В случае ограничений, которые предложены в данном разделе получим явный вид итерационных формул.

Е-step

Найдем $q(\mathbf{Z})$:

$$\begin{aligned} \log q(\mathbf{Z}) &= E_{q/\mathbf{Z}} \log p(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}) \propto \\ &\propto \sum_{i=1}^N \sum_{k=1}^K z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (y_i^2 - \mathbf{x}_i^\top E \mathbf{w}_k + \mathbf{x}_i^\top E \mathbf{w}_k \mathbf{w}_k^\top \mathbf{x}_i) + \frac{1}{2} \log \frac{\beta}{2\pi} \right] \\ p(z_{ik} = 1) &= \frac{\exp(\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (\mathbf{x}_i^\top E \mathbf{w}_k \mathbf{w}_k^\top \mathbf{x}_i - \mathbf{x}_i^\top E \mathbf{w}_k))}{\sum_{k'=1}^K \exp(\log \pi_{k'}(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (\mathbf{x}_i^\top E \mathbf{w}_{k'} \mathbf{w}_{k'}^\top \mathbf{x}_i - \mathbf{x}_i^\top E \mathbf{w}_{k'}))} \end{aligned} \quad (4.4)$$

Найдем $q(\mathbf{W})$:

$$\begin{aligned} \log q(\mathbf{W}) &= E_{q/\mathbf{W}} \log p(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}) \\ &= \sum_{i=1}^N \sum_{k=1}^K E z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (y_i - \mathbf{w}_k^\top \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\ &+ \sum_{k=1}^K \left[-\frac{1}{2} (\mathbf{w}_k - \mathbf{w}_k^0)^\top \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] \\ &= \sum_{k=1}^K \left[\mathbf{w}_k^\top \left(\mathbf{A}_k^{-1} \mathbf{w}_k^0 + \beta \sum_{i=1}^N \mathbf{x}_i y_i E z_{ik} \right) - \frac{1}{2} \mathbf{w}_k^\top \left(\mathbf{A}_k^{-1} + \beta \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{w}_k \right] \end{aligned} \quad (4.5)$$

Получаем распределение параметров:

$$q(\mathbf{w}_k) = N(\mathbf{w}_k | \mathbf{m}_k, \mathbf{B}_k), \quad (4.6)$$

где введены обозначения

$$\mathbf{m}_k = \mathbf{B}_k \left(\mathbf{A}_k^{-1} \mathbf{w}_k^0 + \beta \sum_{i=1}^N \mathbf{x}_i y_i \mathbb{E} z_{ik} \right) \quad \mathbf{B}_k = \left(\mathbf{A}_k^{-1} + \beta \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \quad (4.7)$$

M-step

$$\mathbb{E}_q \log p(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}) = \mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A})$$

$$\begin{aligned} \mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}) = & \sum_{i=1}^N \sum_{k=1}^K \mathbb{E} z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} \mathbb{E} (y_i - \mathbf{w}_k^\top \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\ & + \sum_{k=1}^K \left[-\frac{1}{2} \mathbb{E} (\mathbf{w}_k - \mathbf{w}_k^0)^\top \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] \end{aligned} \quad (4.8)$$

Найдем \mathbf{A} из условия:

$$\frac{\partial \mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}_k)}{\partial \mathbf{A}^{-1}} = \frac{1}{2} \mathbf{A}_k - \frac{1}{2} \mathbb{E} (\mathbf{w}_k - \mathbf{w}_k^0) (\mathbf{w}_k - \mathbf{w}_k^0)^\top = 0 \quad (4.9)$$

Получаем \mathbf{A} :

$$\mathbf{A}_k = \mathbb{E} \mathbf{w}_k \mathbf{w}_k^\top - 2 \mathbf{w}_k^0 \mathbb{E} \mathbf{w}_k^\top + \mathbf{w}_k^0 \mathbf{w}_k^{0\top} \quad (4.10)$$

Найдем \mathbf{V} :

Аналитически решение не ищется, поэтому воспользуемся градиентным спуском для максимизации правдоподобия модели:

$$\mathbf{V}^{j+1} = \mathbf{V}^j + \alpha \frac{\partial \mathcal{F}(\mathbf{W}, \mathbf{V}^j, \beta)}{\partial \mathbf{V}} \quad (4.11)$$

Найдем \mathbf{W}^0 из условия:

$$\frac{\partial \mathcal{F}(\mathbf{V}, \mathbf{W}^0, \mathbf{A})}{\partial \mathbf{w}_k^0} = -2 \mathbf{A}_k^{-1} \mathbb{E} \mathbf{w}_k + 2 \mathbf{A}_k^{-1} \mathbf{w}_k^0 = 0 \quad (4.12)$$

Получаем \mathbf{W}^0 :

$$\mathbf{w}_k^0 = \mathbb{E} \mathbf{w}_k \quad (4.13)$$

4.2 Регуляризация prior распределений на моделях

Рассмотрим случай, когда априорные распределения моделей зависимы. Рассмотрим следующую модель:

1. $p_k(y_i | \mathbf{w}_k, \mathbf{x}_i) = N(y_i | \mathbf{w}_k^\top \mathbf{x}_i, \beta^{-1})$
2. $p^k(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k) = N(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k)$
3. $\pi(\mathbf{x}_i, \mathbf{V}) = \text{softmax}(\mathbf{F}(\mathbf{x}_i, \mathbf{V}))$, где $F : \mathbb{R}^n \times \mathbb{R}^V \rightarrow \mathbb{R}^K$ — нейросеть, V — число параметров нейросети.
4. $p(\boldsymbol{\varepsilon}_{k,k'} | \boldsymbol{\alpha}) = N(\boldsymbol{\varepsilon}_{k,k'} | \mathbf{0}, \text{diag} \boldsymbol{\alpha})$, где $\boldsymbol{\varepsilon}_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0$, $\text{diag} \boldsymbol{\alpha}$ — диагональная матрица, диагональ которой равняется $\boldsymbol{\alpha}$, обозначим данную матрицу $\hat{\boldsymbol{\alpha}}$

С учетом введенных предположений о моделях логарифм правдоподобия имеет следующий вид:

$$\begin{aligned} \log p(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}) &= \sum_{i=1}^N \sum_{k=1}^K z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (y_i - \mathbf{w}_k^\top \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\ &+ \sum_{k=1}^K \left[-\frac{1}{2} (\mathbf{w}_k - \mathbf{w}_k^0)^\top \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] + \\ &+ \sum_{k=1}^K \sum_{k'=1}^K \left[-\frac{1}{2} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0)^\top \hat{\boldsymbol{\alpha}}^{-1} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0) + \frac{1}{2} \log \det \hat{\boldsymbol{\alpha}} - \frac{n}{2} \log 2\pi \right] \end{aligned} \quad (4.14)$$

В данном случае, изменится только выражения (4.12) следующим образом:

$$\mathbf{w}_k^0 = [\mathbf{A}_k^{-1} - K \hat{\boldsymbol{\alpha}}]^{-1} \left(\mathbf{A}_k^{-1} \mathbf{E} \mathbf{w}_k - \sum_{k'=1}^K \hat{\boldsymbol{\alpha}} \mathbf{w}_{k'}^0 \right). \quad (4.15)$$

5 Вычислительный эксперимент

Мультимодель с prior на моделях. Для сравнения мультимодели смеси экспертов с заданием априорного распределения на вектора параметров локальных моделей и мультимодели без задания априорного распределения на вектора параметров локальных моделей был проведен вычислительный эксперимент на синтетической выборке.

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

Предлагается сравнить две разные постановки задачи смеси экспертов: в случае когда используются априорные знания об картинке и в случае когда априорные знания отсутствуют.

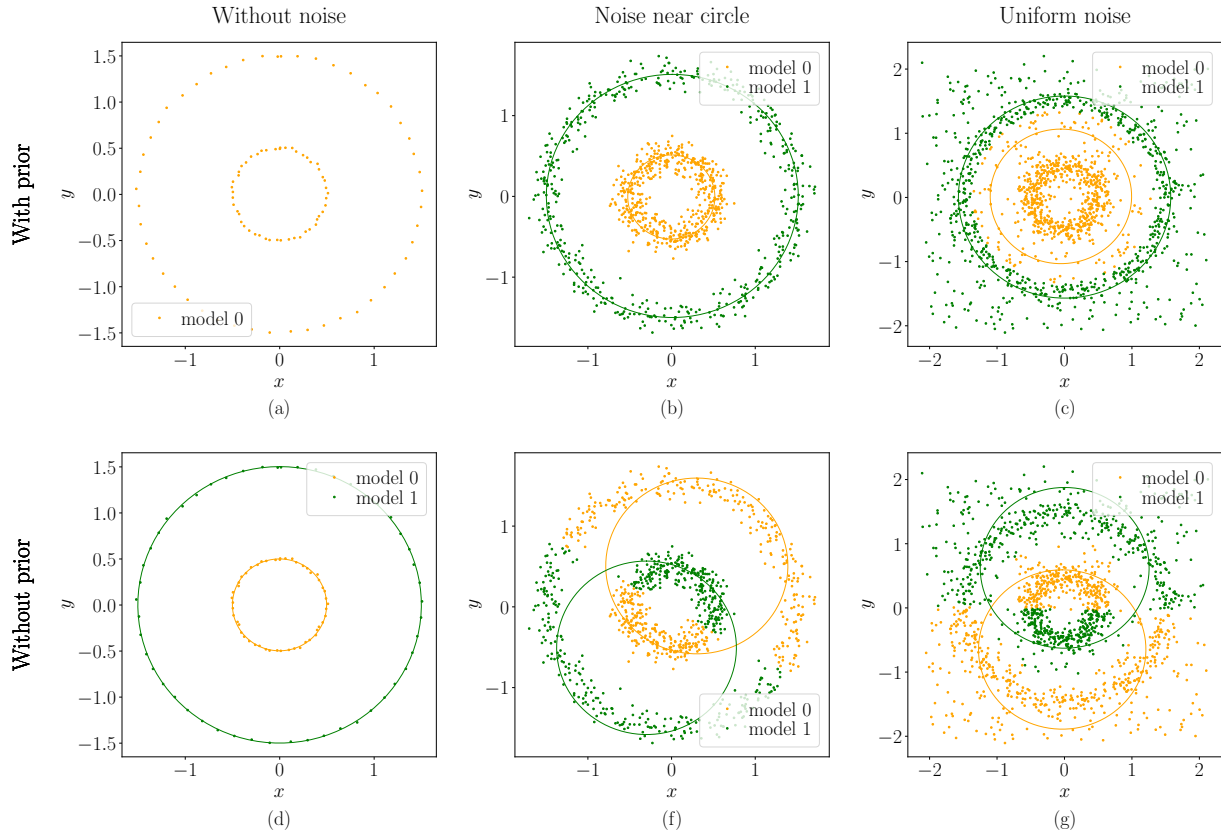


Рис. 3: Результат работы мультимодели в зависимости от априорных знаний и в зависимости от уровня шума: а) модель с априорными знаниями, окружности без шума; б) модель с априорными знаниями, окружности с зашумленным радиусом; в) модель с априорными знаниями, окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению; д) модель без априорных знаний, окружности без шума; е) модель без априорных знаний, окружности с зашумленным радиусом; ф) модель без априорных знаний, окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению

В качестве информативного априорного знания выступает, то, что предполагается что вектора параметров каждой модели имеют следующие распределения:

$$N(\mathbf{w}_1 | \mathbf{w}_1^0, \mathbf{I}), \quad N(\mathbf{w}_2 | \mathbf{w}_2^0, \mathbf{I}), \quad (5.1)$$

где $\mathbf{w}_1^0 = [0, 0, 0.1]$, $\mathbf{w}_2^0 = [0, 0, 5]$, что указывает, на то, что известно о концентричности окружностей, а также что у них радиусы различны.

На рис. 3 показан случайный результаты работы мультимодели с априорными знаниями и без них. На всех картинках обе модели работала 30 итераций. Так-как сходимость мультимодели очень сильно зависит от начальной инициализации, также был проведен эксперимент с множественным запуском мультимодели на одном и том же изображении. Обе модели на каждом изображении запускались по 100 раз.

Таблица 1: Результаты работы мультимodelей

Мультимodelь	Without Noise	Noise near circle	Uniform noise
With prior	99/100	97/100	95/100
Without prior	68/100	23/100	7/100

В таб. 1 показано сколько мультимodelей правильно отыскивали обе окружности на рисунке. Как видно мультимodelь с использованием априорных знаний является более стабильной, чем мультимodelь, которая не использует никаких априорных знаний.

Мультимodelь с регуляризацией. Проведен эксперимент на реальной выборке. В качестве данных рассматривались глаза, а точнее их предобработанное бинарное изображение с выделенными окружностями. Предлагается сравнить работу мультимodelей с регуляризацией и без.

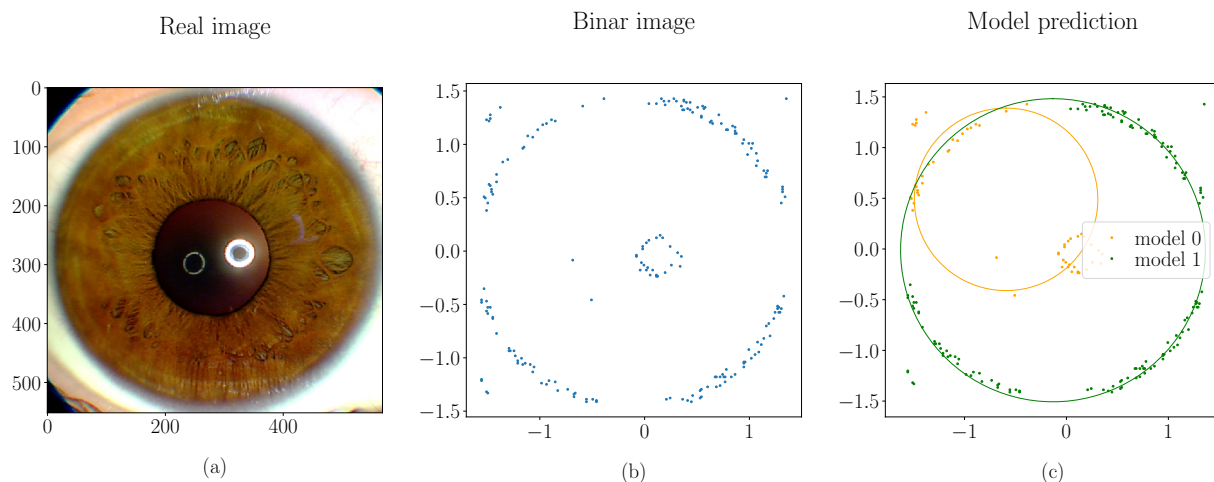


Рис. 4: Результат работы мультимodelи в случае, когда регуляризация не используется: а) исходное изображение; б) бинаризованное изображение; в) результат работы алгоритма

На рис. 4 показан результат работы мультимodelи в случае, когда информация о том, что центры окружностей должны совпадать не используется. В этом случае, мультимodelь не верно находит внутреннюю окружность.

На рис. 5 показан результат работы мультимodelи в случае, когда используется регуляризация априорных гиперпараметров w_0^k . На рис. 5, мультимodelь верно нашла обе окружности. Но окружность, которая соответствует модели 1 очень сильно сдвинута в сторону выброса из окружности.

На рис. 6 показано как меняются модели в процессе обучения мультимodelи в случае, когда используется регуляризация априорных распределений. Видно, что после 30 эпох мультимodelь остается неизменной.

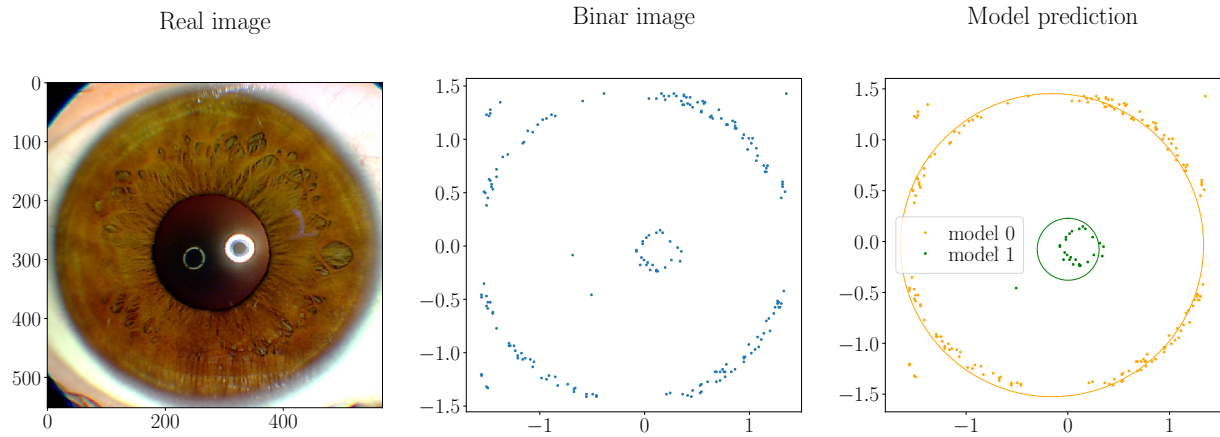


Рис. 5: Результат работы мультимодели в случае, когда регуляризация используется: а) исходное изображение; б) бинаризованное изображение; в) результат работы алгоритма

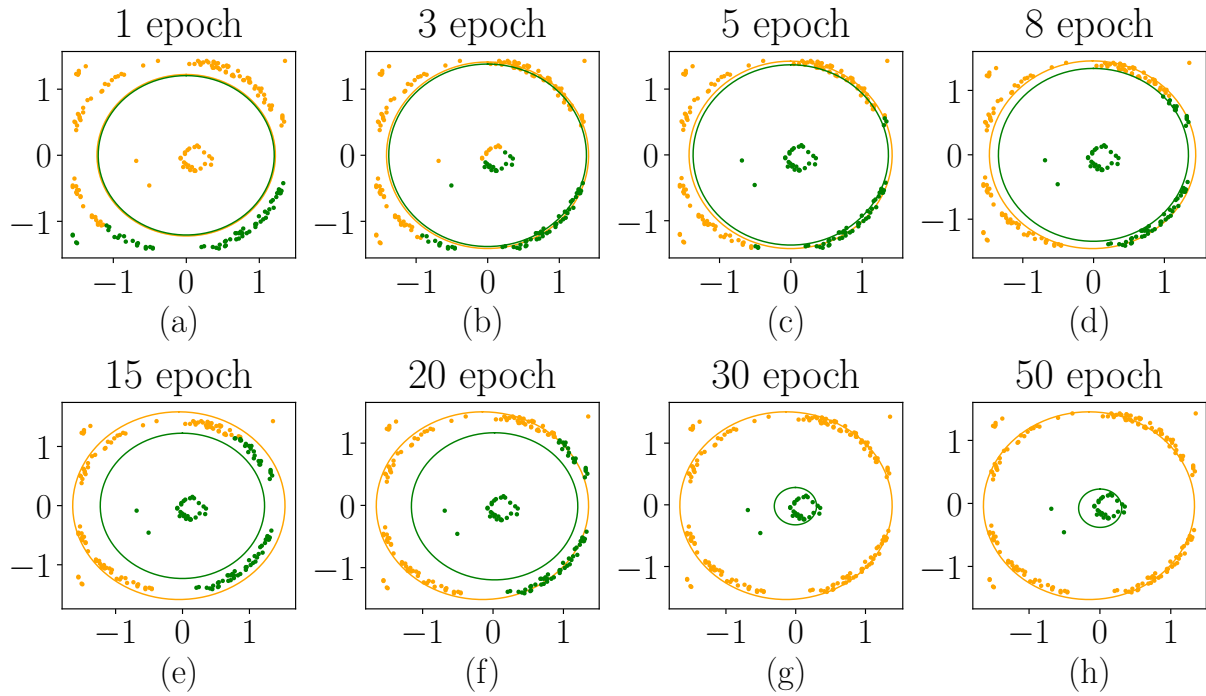


Рис. 6: Визуализация процесса обучения: а) результат работы после 1й эпохи; б) результат работы после 3й эпохи; в) результат работы после 5й эпохи; д) результат работы после 8й эпохи; е) результат работы после 20й эпохи; ф) результат работы после 30й эпохи; г) результат работы после 50й эпохи

6 Заключение

В данной работе проведено сравнение мультимodelей в случае, когда было задано априорное распределение параметров каждой модели внутри мультимодели и в слу-

чае, когда априорного распределения не было. В качестве данных использовались изображения концентрических окружностей с разным уровнем шума. Для поиска окружностей использовались линейные модели. В качестве плюсовой функции использовалась двухслойная нейросеть.

Как показано в эксперименте в случае, когда введены априорные знания на линейные модели, мультимодель является более устойчивой к шуму. Также в случае задания априорных знаний моделей, мультимодель менее зависит от начальной инициализации, что также позволяет сказать, что модель является более устойчивой к начальной инициализации.

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 плюсовую 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] *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.