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. Предлагается проанализировать свойства моделей в случае, когда выбрано информативное и не информативное априорное распределения на веса 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 картинке. Предполагается, что каждой картинке соответсвует свой эксперт. В качестве данных рассматриваются синтетически сгенерированные окружности с разным уровнем шума. Предлагается сравнить устойчивость к шуму мультимомделей с априорными знаниями и без них.

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

DOI: 00.00000/00000000000000

1 Введение

В настоящее время в разных задач анализа данных присутствует необходимость обработки поступающих данных локально.

Мультимодели показывают отличные результаты во многих задачах. Классическими методами являются беггинг и градиентный бустинг [1], случайный лес [2]. Более современный 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

*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

эксперта — отдельной модели. В качестве шлюзовой функции выступает: softmax–регрессия, процесс Дирихле [5], нейронная сеть [4] с softmax на последнем слое.

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

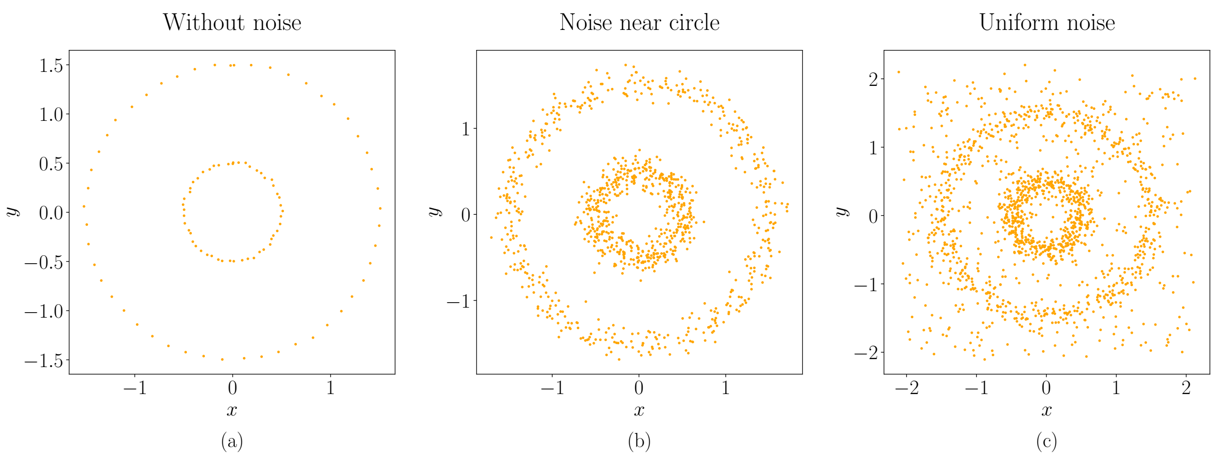

Рис. 1: Пример изображений с окружностями с разным уровнем шума: a) окружности без шума; b) окружности с зашумленным радиусом; c) окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению

В данной работе предлагается исследовать, как влияют априорные знания на сходимость мультимодели. Для данного исследования решается задача поиска окружностей на биноризованной картинке — где каждый пиксель закрашен в один из двух цветов. Пример картинок показан на рис. 1. Предлагается рассмотреть как ведет себя модель с априорными знанием и без них в случае картинок 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 Постановка задачи нахождения параметров окружностей

Задано бинарное изображение:

$$\mathbf{M} \in \{0, 1\}^{m_1 \times m_2}, \tag{2.1}$$

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

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

$$\mathbf{C} \in \mathbb{R}^{N \times 2}, \tag{2.2}$$

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

Пусть x_0, y_0 — центр окружности, которую нужно найти на бинарной картинке $\mathbf{M}$, а r ее радиус. Тогда элементы выборки $\mathbf{C}$ должны удовлетворять уравнению окружности:

$$(x_i - x_0)^2 + (y_i - y_0)^2 = r^2. \tag{2.3}$$

Раскрыв скобки получим следующие уравнение:

$$(2x_0) \cdot x_i + (2y_0) \cdot y_i + (r^2 - x_0^2 - y_0^2) \cdot 1 = x_i^2 + y_i^2. \tag{2.4}$$

Получаем следующую задачу линейной регрессии для нахождения параметров окружности:

$$\mathbf{Xw} \approx \mathbf{Y}, \quad \mathbf{X} = \mathbf{C} \times \mathbf{1}, \quad \mathbf{Y} = \{x^2 + y^2 | \forall x, y \in \mathbf{C}\}, \tag{2.5}$$

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

$$x_0 = \frac{w_1}{2}, \quad y_0 = \frac{w_2}{2}, \quad r = \sqrt{w_3 + x_0^2 + y_0^2}. \tag{2.6}$$

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.

3 Постановка задачи смеси экспертов

Задана выборка:

$$\mathbf{X} \in \mathbb{R}^{N \times n}, \tag{3.1}$$

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

Определение 3.1. *Смесь экспертов — мультимодель, определяющая правдоподобие каждой π_k каждой модели $\boldsymbol{f}_k$ на объекте $\boldsymbol{x}$ на основе его признакового опсиания.*

$$\hat{\boldsymbol{f}} = \sum_{k=1}^{K} \pi_k \boldsymbol{f}_k, \qquad \pi_k(\boldsymbol{x}, \boldsymbol{V}) : \mathbb{R}^{2 \times n} \to [0, 1], \qquad \sum_{k=1}^{K} \pi_k(\boldsymbol{x}, \boldsymbol{V}) = 1 \tag{3.2}$$

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

3.1 Общий случай

Правдоподобие модели:

$$p\left(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[\pi_k\left(\mathbf{x}_i, \mathbf{V}\right) p_k\left(y_i | \mathbf{w}_k, \mathbf{x}_i\right)\right]^{z_{ik}} \prod_{k=1}^{K} p^k\left(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k\right) \tag{3.3}$$

Логарифм правдоподобия модели:

$$\begin{aligned} \log p\left(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left[\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) + \log p_k\left(y_i | \mathbf{w}_k, \mathbf{x}_i\right)\right] + \\ + \sum_{k=1}^{K} \log p^k\left(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k\right) \end{aligned} \tag{3.4}$$

E-step

Найдем $q\left(\mathbf{Z}\right)$:

$$\begin{aligned} \log q\left(\mathbf{Z}\right) &= \mathsf{E}_{q/\mathbf{Z}} \log p\left(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) \\ p\left(z_{ik} = 1\right) &= \frac{\exp\left(\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) + \mathsf{E} \log p_k\left(y_i | \mathbf{w}_k, \mathbf{A}_k\right)\right)}{\sum_{k'=1}^{K} \exp\left(\log \pi_{k'}\left(\mathbf{x}_i, \mathbf{V}\right) + \mathsf{E} \log p_{k'}\left(y_i | \mathbf{w}'_k, \mathbf{A}'_k\right)\right)} \end{aligned} \tag{3.5}$$

Найдем $q\left(\mathbf{W}\right)$:

$$\begin{aligned} \log q\left(\mathbf{W}\right) &= \mathsf{E}_{q/\mathbf{W}} \log p\left(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathsf{E} z_{ik} \left[\log \pi_k\left(\mathbf{x}_{i,\mathbf{V}}\right) + \log p_k\left(y_i\left(y_i | \mathbf{w}_k, \mathbf{x}_i\right)\right)\right] + \\ &+ \sum_{k=1}^{K} \log p^k\left(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k\right) \end{aligned} \tag{3.6}$$

M-step

$$\begin{aligned} \mathsf{E}_q p\left(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) &= \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) \\ \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) &= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathsf{E} z_{ik} \left[\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) + \mathsf{E} \log p_k\left(y_i | \mathbf{w}_k, \mathbf{x}_i\right)\right] \\ &+ \sum_{k=1}^{K} \mathsf{E} \log p_k\left(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k\right) \end{aligned} \tag{3.7}$$

Найдем $\mathbf{A}$ из условия:

$$\frac{\partial \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}\right)}{\partial \mathbf{A}^{-1}} = 0 \tag{3.8}$$

Найдем $\mathbf{V}$ из условия:

$$\frac{\partial \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}\right)}{\partial \mathbf{V}} = 0 \tag{3.9}$$

Найдем $\mathbf{W}^0$ из условия:

$$\frac{\partial \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}\right)}{\partial \mathbf{W}^0} = 0 \tag{3.10}$$

3.2 Случай линейной регресионной модели

Рассмотрим следующий случай распределений:

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

Правдоподобие модели:

$$p\left(\mathbf{y}, \mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) = \prod_{i=1}^{N}\prod_{k=1}^{K}\left[\pi_k\left(\mathbf{x}_i, \mathbf{V}\right)\mathrm{N}\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right)\right]^{z_{ik}}\prod_{k=1}^{K}\mathrm{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right) \tag{3.11}$$

Логарифм правдоподобия модели:

$$\begin{aligned}\log p\left(\mathbf{y}, \mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\ + \sum_{k=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right]\end{aligned} \tag{3.12}$$

E-step

Найдем $q\left(\mathbf{Z}\right)$:

$$\begin{aligned}\log q\left(\mathbf{Z}\right) = \mathsf{E}_{q/\mathbf{Z}}\log p\left(\mathbf{y}, \mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) \propto \\ \propto \sum_{i+1}^{N}\sum_{k=1}^{K} z_{ik}\left[\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) - \frac{\beta}{2}\left(y_i^2 - \mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right) + \frac{1}{2}\log\frac{\beta}{2\pi}\right] \\ p\left(z_{ik} = 1\right) = \frac{\exp\left(\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) - \frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i - \mathbf{x}_i^{\mathsf{T}}\mathsf{T}\mathbf{w}_k\right)\right)}{\sum_{k'=1}^{K}\exp\left(\log \pi_{k'}\left(\mathbf{x}_i, \mathbf{V}\right) - \frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_{k'}\mathbf{w}_{k'}^{\mathsf{T}}\mathbf{x}_i - \mathbf{x}_i^{\mathsf{T}}\mathsf{T}\mathbf{w}_{k'}\right)\right)}\end{aligned} \tag{3.13}$$

Найдем $q(\mathbf{W})$:

$$\begin{aligned}\log q(\mathbf{W}) &= \mathsf{E}_{q/\mathbf{W}} \log p\left(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{E} z_{ik} \left[\log \pi_k\left(\mathbf{x}_{i,\mathbf{V}}\right) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\ &+ \sum_{k=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}} \mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] \\ &= \sum_{k=1}^{K}\left[\mathbf{w}_k^{\mathsf{T}}\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0 + \beta\sum_{i=1}^{N}\mathbf{x}_i y_i \mathsf{E} z_{ik}\right) - \frac{1}{2}\mathbf{w}_k^{\mathsf{T}}\left(\mathbf{A}_k^{-1} + \beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right)\mathbf{w}_k\right]\end{aligned} \tag{3.14}$$

Получаем распределение параметров:

$$q(\mathbf{w}_k) = \mathrm{N}\left(\mathbf{w}_k | \mathbf{m}_k, \mathbf{B}_k\right), \tag{3.15}$$

где введены обозначения

$$\mathbf{m}_k = \mathbf{B}_k\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0 + \beta\sum_{i=1}^{N}\mathbf{x}_i y_i \mathsf{E} z_{ik}\right) \qquad \mathbf{B}_k = \left(\mathbf{A}_k^{-1} + \beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right)^{-1} \tag{3.16}$$

M-step

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

$$\begin{aligned}\mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}\right) &= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{E} z_{ik}\left[\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) - \frac{\beta}{2}\mathsf{E}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\ &+ \sum_{k=1}^{K}\left[-\frac{1}{2}\mathsf{E}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right]\end{aligned} \tag{3.17}$$

Найдем $\mathbf{A}$ из условия:

$$\frac{\partial\mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}_k\right)}{\partial\mathbf{A}^{-1}} = \frac{1}{2}\mathbf{A}_k - \frac{1}{2}\mathsf{E}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^T = 0 \tag{3.18}$$

Получаем $\mathbf{A}$:

$$\mathbf{A}_k = \mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}} - 2\mathbf{w}_k^0\mathsf{E}\mathbf{w}_k^{\mathsf{T}} + \mathbf{w}_k^0\mathbf{w}_k^{0\mathsf{T}} \tag{3.19}$$

Найдем $\mathbf{V}$:

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

$$\mathbf{V}^{j+1} = \mathbf{V}^j + \alpha\frac{\partial\mathcal{F}\left(\mathbf{W}, \mathbf{V}^j, \beta\right)}{\partial\mathbf{V}} \tag{3.20}$$

Найдем $\mathbf{W}^0$ из условия:

$$\frac{\partial \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}\right)}{\partial \mathbf{w}_k^0} = -2\mathbf{A}_k^{-1}\mathsf{E}\mathbf{w}_k + 2\mathbf{A}_k^{-1}\mathbf{w}_k^0 = 0 \tag{3.21}$$

Получаем $\mathbf{W}^0$:

$$\mathbf{w}_k^0 = \mathsf{E}\mathbf{w}_k \tag{3.22}$$

4 Вычислительный эксперимент

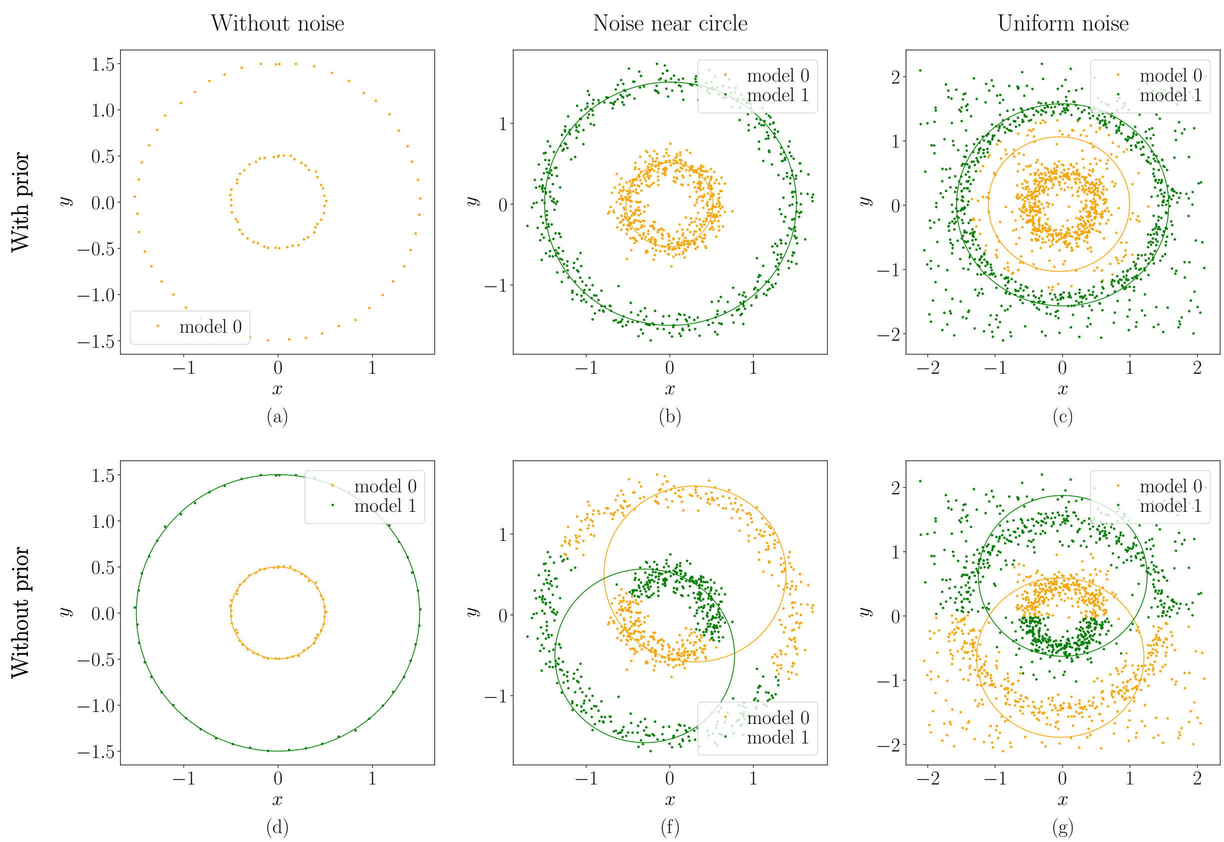

Рис. 2: Результат работы мультимодели в зависимости от априорных знаний и в зависимости от уровня шума: a) модель с априорными знаниями, окружности без шума; b) модель с априорными знаниями, окружности с зашумленным радиусом; c) модель с априорными знаниями, окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению; d) модель без априорных знаний, окружности без шума; f) модель без априорных знаний, окружности с зашумленным радиусом; g) модель без априорных знаний, окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению

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.

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

$$\mathrm{N}\left(\mathbf{w}_1|\mathbf{w}_1^0, \mathbf{I}\right), \quad \mathrm{N}\left(\mathbf{w}_2|\mathbf{w}_2^0, \mathbf{I}\right), \tag{4.1}$$

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.

Таблица 1: Результаты работы мультимоделей

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

На рис. 2 показан случайный результаты работы мультимоделей с априорными знаниями и без них. На всех картинках обе модели работала 30 итераций. Так-как сходимость мултимодели очень сильно зависит от начальной инициализации, также был проведен эксперперимент с множественным запуском мултимодели на одной и той же картиинке. Обе модели на каждой картинке запускались по 100 раз. В таб. 1 показано сколько мультимоделей правильно отыскали обе окружности на рисунке. Как видно мультимодель с использованием априорных знаний является более стабильной, чем мультимодель, которая не использует никаких априорных знаний.

5 Заключение

В данной работе проведено сравнение мультимоделей в случае, когда каждая модель имела априорные знания и в случае, когда априорных знаний 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 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.

В дальнейшем планируется улучшить мультимодель при помощи задания априорных знаний на шлюзовую функцию.

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.