

МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

На правах рукописи

Грабовой Андрей Валериевич

АПРИОРНОЕ РАСПРЕДЕЛЕНИЕ ПАРАМЕТРОВ
В ЗАДАЧАХ ВЫБОРА МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ

05.13.17 — Теоретические основы информатики

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель:
д.ф.-м.н. В. В. Стрижов

Москва — 2022

Оглавление

	Стр.
Введение	4
Глава 1. Априорное распределения параметров моделей	9
1.1. Привилегированное обучение Вапника и дистилляция Хинтона	11
1.2. Релевантность параметров моделей глубокого обучения	14
1.3. Смесь экспертов для аппроксимации мультимодальной выборки	17
Глава 2. Модели привилегированного обучения и дистилляции	18
2.1. Обобщенная вероятностная постановка задачи дистилляции	19
2.2. Подход дистилляции модели учителя в модель ученика	19
2.3. Анализ вероятностного подхода к дистилляции линейных моделей	24
Глава 3. Байесовская дистилляция моделей глубокого обучения	31
3.1. Постановка задачи дистилляции в терминах байесовского подхода	32
3.2. Выравнивание априорного распределения параметров ученика на основе параметров учителя	33
3.3. Последовательность выравнивающих преобразований	40
3.4. Анализ байесовской дистилляции полносвязных нейронных сетей	42
Глава 4. Априорные распределения параметров смеси экспертов	46
4.1. Локальные модели в задаче построения смеси экспертов	49
4.2. Вероятностное обоснование смеси экспертов	51
4.3. Априорное распределение для аппроксимации кривых второго порядка на изображении	54
4.4. Анализ качества аппроксимации смесью экспертов	57
Глава 5. Введение отношения порядка на множестве параметров аппрокси- мирующих моделей	71
5.1. Задача упорядочивания параметров аппроксимирующих моделей	72
5.2. Определение релевантности на основе метода Белсли	73
5.3. Анализ разных подходов к определению релевантности	75
5.4. Вычислительный эксперимент по упорядочиванию параметров	79
Глава 6. Анализ прикладных задач выбора моделей машинного обучения	86
6.1. Постановка задачи определения достаточного размера выборки	88
6.2. Байесовский подход к определению достаточного размера выборки	89
6.3. Анализ методов определения достаточного размера выборки	94
6.4. Кластеризация точек квазипериодических временных рядов	98
6.5. Анализ фазовых траекторий в задаче кластеризации	103
Заключение	109
Список основных обозначений	111

Список иллюстраций	112
Список таблиц	117
Список литературы	118

Введение

Актуальность темы. Построение и выбор оптимальной структуры нейронной сети является вычислительно сложной процедурой [1], которая значимо влияет на итоговое качество модели. При этом большинство параметров модели перестают значимо изменяться уже после небольшого числа итераций алгоритма оптимизации [2]. Своевременное определение начала сходимости параметров существенно снижает вычислительные затраты на обучение моделей с большим числом параметров. Примерами моделей, с большим числом параметров, являются AlexNet [3], VGGNet [4], ResNet [5], BERT [6, 7], mT5 [8], GPT3 [9]. Рост числа параметров моделей глубокого обучения влечет снижение интерпретируемости ответов этих моделей [10]. Проблема с неинтерпретируемыми моделями рассматривается в классе задач по состязательным атакам [11].

Проблемой моделей с большим числом параметров является увеличение вычислительной сложности. Использование избыточно сложных моделей с большим числом неинформативных параметров является препятствием для использования глубоких сетей на мобильных устройствах в режиме реального времени. *Сложность модели* определяется числом настраиваемых параметров модели. Для снижения числа параметров в литературе рассматривается метод дистилляции модели на основе предсказаний модели учителя [12, 13, 14]. Сложная модель с большим числом параметров называется *учителем*. Модель учителя дистиллируется в менее сложную модель с малым числом параметров, которая называется *ученик*. Методы дистилляции моделей глубокого обучения введены в работах Дж. Е. Хинтона и В. Н. Вапником [12, 13, 14]. Предлагается использовать предсказания модели учителя для повышения качества ученика. В [13] В. Н. Вапником вводит понятие привилегированной информации. Оно использует дополнительную информацию о данных в момент обучения модели. Работа [14] объединяет идеи дистилляции [12] с идеями привилегированного обучения [13], предложив метод дистилляции учителя в модель ученика в случае, когда признаковое описание объектов не совпадает.. В [14] решается двухэтапная задача. На первом этапе строится модель учителя с расширенным признаком описанием. На втором этапе при помощи дистилляции [12] обучается ученика в исходном признаковом описании. В работе Дж. Е. Хинтона [12] поставлены эксперименты по дистилляции моделей глубокого обучения для задачи классификации. Первый эксперимент анализирует выборку MNIST [15]. Он показывает, что предложенный метод дистилляции позволяет построить нейросетевую модель меньшей сложности на основе модели большей сложности. Второй эксперимент анализирует метод дистилляции ансамбля моделей в одну нейросетевую модель для решения задачи распознания речи. В работе [12] проводится сравнение дистилляции с моделью смеси экспертов. Дальнейшие работы по дистилляции моделей глубокого обучения исследуют методы, использующие информации о значениях параметров модели учителя, для оптимизации параметров модели ученика. В [16] предлагается метод передачи селективно-

сти [17] нейрона минимизирующий специальную функцию потерь. Эта функция основывается на максимизация среднего описания между выходами слоев модели учителя и модели ученика. В рамках вычислительного эксперимента сравнивалось качество базовой дистилляции с предложенным методом на примере выборок CIFAR [18] и ImageNet [19].

Дистилляция моделей глубокого обучения предполагает, что архитектура модели ученика уже известна. Для выбора архитектуры модели ученика предлагается использовать методы прореживания нейросетевых моделей. В работах [20, 21] предлагается использовать алгоритм градиентного спуска для оптимизации сети. В [22] используются байесовские методы [23] оптимизации параметров нейронных сетей. Существуют методы поиска оптимальной структуры используя удаления параметров сложной модели [24, 25, 26]. В работе [24] предлагается удалять наименее *релевантные* параметры на основе значений первой и второй производных функции ошибки. В [27] предложен метод определения релевантности параметров аппроксимирующих моделей при помощи метода Белсли. *Релевантность* параметров в работе [27] определяется на основе ковариационной матрицы параметров модели. Другим примером задания порядка на множестве параметров служит l_1 -регуляризация [28] и регуляризация ElasticNet [29] для линейных моделей. Порядок, заданный на множестве значений коэффициентов регуляризации, индуцирует порядок на множестве признаковых описаний и указывает на важность признаков. В случае нейросетей для регуляризации параметров используется метод исключения параметров [30, 22]. Он также задает порядок на множестве параметров модели.

Порядок на множестве параметров нейросети используется не только для удаления неимение релевантных параметров, а и для фиксации параметров в процессе оптимизации параметров. Работа [31] посвящена оптимизации структуры нейронной сети, а также выбору параметров, которые фиксируются после некоторой итерации градиентного метода.

Цели работы.

1. Предложить байесовский метод выбора моделей с использованием модели учителя с привилегированной и накопленной информацией.
2. Предложить метод назначения априорного распределения параметров модели ученика с использованием апостериорного распределения параметров модели учителя.
3. Предложить вероятностную интерпретацию дистилляции моделей глубокого обучения.
4. Предложить метод использования экспертной информации о решаемой задачи прогнозирования при построении априорного распределения параметров.
5. Предложить метод назначения релевантности параметров моделей глубокого обучения для выбора модели машинного обучения.

Методы исследования. Для достижения поставленных целей используются методы вариационного байесовского вывода [32, 33], вероятностные [34] методы к анализу моделей глубокого обучения, статистические методы [35, 33] анализа распределений параметров моделей глубокого обучения.

Основные положения, выносимые на защиту.

1. Предложен байесовский метод выбора моделей с использованием модели учителя с привилегированной и накопленной информацией.
2. Доказаны теоремы о свойствах дистилляции,
 - *теоремы об эквивалентности* для дистилляции моделей в случае задачи регрессии и классификации,
 - *теоремы о виде априорного распределения* параметров модели ученика в байесовской дистилляции.
3. Предложен метод выравнивания вероятностных пространств параметров. Предложен метод выбора априорного распределения параметров модели ученика с использованием апостериорного распределения параметров модели учителя для случаев
 - различных размерностей пространств параметров отдельных слоев,
 - различного числа слоев нескольких моделей.
4. Предложены методы задания порядка на множестве параметров моделей
 - на основе корреляции параметров,
 - на основе оценки скорости сходимости параметров.
5. Предложена вероятностная интерпретации дистилляции моделей глубокого обучения. Исследованы свойства дистилляции моделей глубокого обучения.

Научная новизна. Разработаны новые подходы к назначению априорного распределения параметров моделей. Предложен метод назначения априорного распределения используя экспертную информацию о задаче. Предложены методы задания порядка на множестве параметров нейросетевых моделей на основе анализа мультиколлинеарности параметров и скорости их сходимости. Предложено вероятностное обобщение дистилляции моделей. Предложено байесовское обобщение дистилляции моделей глубокого обобщения.

Теоретическая значимость. Диссертационная работа носит теоретический характер. В работе проводится теоретический анализ методов снижения размерности пространства параметров нейросетевых моделей. Доказаны *теоремы об эквивалентности* для дистилляции моделей в случае задачи регрессии и классификации. Доказаны *теоремы об априорном распределении* модели для байесовской дистилляции.

Практическая значимость. Предложенные в работе методы предназначены для построения моделей глубокого обучения в прикладных задачах регрессии и классификации; снижения пространства параметров моделей глубокого обучения; использования экспертной информации для построения моделей; дистилляции параметрических моделей на основе выравнивания архитектур.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных методов на реальных задачах выбора моделей глубокого обучения; публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК. Результаты работы докладывались и обсуждались на следующих научных конференциях.

1. Задача обучения с экспертом для построения интерпретируемых моделей машинного обучения, Международная конференция «Интеллектуализация обработки информации», 2020.
2. Привилегированная информация и дистилляция моделей, Всероссийская конференция «63-я научная конференция МФТИ», 2020.
3. Введение отношения порядка на множестве параметров нейронной сети, Всероссийская конференция «Математические методы распознавания образов ММРО», 2019.
4. Анализ априорных распределений в задаче смеси экспертов, Всероссийская конференция «62-я научная конференция МФТИ», 2019.
5. Поиск оптимальной модели при помощи алгоритмов прореживания, Всероссийская конференция «61-я научная конференция МФТИ», 2018.
6. Автоматическое определение релевантности параметров нейросети, Международная конференция «Интеллектуализация обработки информации», 2018.

Работа поддержана грантами Российского фонда фундаментальных исследований:

- 1) 19-07-00875, Развитие методов автоматического построения и выбора вероятностных моделей субоптимальной сложности в задачах глубокого обучения,
- 2) 19-07-01155, Развитие теории порождения моделей локальной аппроксимации для классификации сигналов носимых устройств,
- 3) 19-07-00885, Выбор моделей в задачах декодирования временных рядов высокой размерности.

Публикации по теме диссертации. Основные результаты по теме диссертации изложены в 6 печатных изданиях в журналах, рекомендованных ВАК.

1. Грабовой А.В., Стрижсов В.В. Байесовская дистилляция моделей глубокого обучения // Автоматика и телемеханика — 2021. — Т. 11. — С. 16–29.

2. Грабовой А.В., Стрижсов В.В. Анализ выбора априорного распределения для смеси экспертов // Журнал вычислительной математики и математической физики — 2021. — Т. 61, № 7. — С. 1149–1161.
3. Grabovoy A., Strijov V. Quasi-periodic time series clustering for human // Lobachevskii Journal of Mathematics — 2020. Vol. 41. — Pp. 333–339.
4. Грабовой А.В., Бахтеев О.Ю., Стрижсов В.В. Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения — 2020. — Т. 14, № 2. — С. 58–65.
5. Грабовой А.В., Бахтеев О.Ю., Стрижсов В.В. Определение релевантности параметров нейросети // Информатика и ее применения — 2019. — Т. 13, № 2. — С. 62–70.
6. Грабовой А.В., Стрижсов В.В. Вероятностная интерпретация задачи дистилляции // Автоматика и телемеханика — 2022. — Т. 1. — С. 150–168.

Личный вклад. Все приведенные результаты, кроме отдельно оговоренных случаев, получены диссертантом лично при научном руководстве д.ф.-м.н. В. В. Стрижова.

Структура и объем работы. Диссертация состоит из оглавления, введения, шести разделов, заключения, списка иллюстраций, списка таблиц, перечня основных обозначений и списка литературы из 99 наименований. Основной текст занимает 124 страницы.

Краткое содержание работы по главам. В главе 1 вводятся основные понятия, поставлены задачи выбора априорного распределения параметров моделей машинного обучения. Проанализированы методы дистилляции и привилегированного обучения предложенные Владимиром Наумовичем Вапником и Джеком Хинтоном. Анализируются существующие методы задания порядка на множестве параметров нейросетевых моделей.

В главе 2 предложены методы обобщения дистилляции и привилегированного обучения на основе вероятностного подхода.

В главе 3 предложен байесовский подход для дистилляции моделей глубокого обучения на основе вариационного вывода.

В главе 4 предложены методы задания априорного распределения параметров локальных моделей в задаче обучения смеси экспертов.

В главе 5 предложены методы введения отношения порядка на множестве параметров аппроксимирующих моделей.

В главе 6 проведен анализ прикладных задач, которые используют экспертную информацию.

Глава 1

Априорное распределения параметров моделей

Повышение точности аппроксимации в задачах машинного обучения влечет повышение сложности моделей и снижает их интерпретируемость. Примеры моделей с повышенной сложностью являются AlexNet [3], VGGNet [4], ResNet [5], BERT [6, 7], mT5 [8], GPT3 [9], а также ансамбли этих моделей. Табл. 1.1 описывает глубокие модели машинного обучения. Число параметров моделей машинного обучения с годами растет. Это влечет снижение интерпретируемости моделей. Данная проблема рассматривается в специальном классе задач по сопротивительным атакам (англ. adversarial attack) [11].

Таблица 1.1: Анализ роста числа параметров при развитии моделей глубокого обучения

Название	AlexNet	VGGNet	ResNet	BERT	mT5	GPT3
Год	2012	2014	2015	2018	2020	2020
Тип данных	изображение	изображение	изображение	текст	текст	текст
Число параметров, млрд	0,06	0,13	0,06	0,34	13	175

При построении модели машинного обучения оптимизируются два критерия: сложность модели и точность аппроксимации модели.

Определение 1. Сложность модели (*структурная сложность*) — число обучаемых параметров, используемых предсказательной моделью.

Модель, которая имеет меньшую сложность при фиксированной точности, является более предпочтительной [36]. Для снижения сложности рассматривается метод *дистилляции* моделей глубокого обучения. Он строит новые модели на основе ранее обученных моделей.

Определение 2. Дистилляция модели — снижение сложности модели путем выбора модели в множестве более простых моделей на основе анализа пространства параметров и предсказаний целевой переменной более сложной фиксированной модели.

Исследуется проблема снижения числа обучаемых параметров моделей машинного обучения. Большое число параметров требует больших вычислительных ресурсов. Из-за этого данные модели не могут быть использованы в мобильных устройствах. Идея дистилляции предложена в работах Дж. Е. Хинтона и В. Н. Вапником [12, 13, 14]. В этих работах предлагается использовать ответы учителя в качестве целевой переменной для обучения модели ученика. Для снижения числа параметров предложен метод дистилляции модели [12, 13, 14].

Дистилируемая модель с большим числом параметров называется *учителем*, а модель получаемая путем дистилляции называется *учеником*. При оптимизации параметров модели ученика используется модель учителя с фиксированными параметрами.

В работе [12] Дж. Е. Хинтоном предлагается метод дистилляции моделей машинного обучения для задачи классификации и проведены эксперименты дистилляции моделей. Проведен эксперимент на выборке MNIST [15], в котором нейросеть с избыточным числом параметров дистиллирована в нейросеть меньшей сложности. Эксперимент по распознаванию речи, в котором ансамбль моделей дистиллирован в одну модель. Проведен эксперимент по обучению экспертных моделей на основе одной большой модели.

Определение 3. *Привилегированная информация — множество признаков, доступных только при выборе модели, но не в при тестировании.*

В работе [13] В. Н. Вапником введено понятие *привилегированной информации*. В работе [14] метод дистилляции [12] используется вместе с привилегированным обучением [13]. На первом этапе обучается модель *учителя* в пространстве привилегированной информации. На втором этапе обучается модель *ученика* в исходном признаковом пространстве используя *дистилляцию* [12]. Для обучения строится функция ошибки специального вида, которая подробно анализируется во 2й главе. Эта функция состоит из нескольких слагаемых. Она включает ошибку учителя, ученика и регуляризирующие элементы. Первый вариант этой функции ошибки предложен А. Г. Ивахненко [10].

Определение 4. *Учитель — фиксированная модель, ответы которой используются при выборе модели ученика.*

Определение 5. *Ученик — модель, которая выбирается согласно заданного критерия качества использующего учителя.*

Поставлен ряд экспериментов, в которых проводилась дистилляция моделей для задачи классификации машинного обучения. Базовый эксперимент на выборке MNIST [15] показал применимость метода для дистилляции избыточно сложной модели в модель меньшей сложности. Эксперимент по дистилляции ансамбля моделей в одну модель для решения задачи распознания речи. Также в работе [12] проведен эксперимент по обучению экспертных моделей на основе одной модели с большим числом параметров при помощи предложенного метода дистилляции на ответах учителя.

В работе [16] предложен метод передачи селективности нейронов (англ. neuron selectivity transfer) основанный на минимизации специальной функции потерь основаной на максимальном среднем отклонении (англ. maximum mean discrepancy) между выходами всех слоев модели учителя и ученика. Вычислительный эксперимент показал эффективность данного метода для задачи классификации изображений на примере выборок CIFAR [18] и ImageNet [19].

Важным свойством дистиллированных является то, что избыточная сложность модели учителя заключается в большом числе не релевантных параметров.

Определение 6. *Релевантность параметров — численная характеристика описывающая влияние параметров на предсказания моделей.*

Предлагается удалять наименее релевантные параметры модели. Под *релевантностью* [24] подразумевается то, насколько параметр влияет на функцию ошибки. Малая релевантность указывает на то, что удаление этого параметра не влечет значимого изменения функции ошибки. Метод предлагает построение исходной избыточной сложности нейросети с большим количеством избыточных параметров.

В работах предлагается [24, 26] метод введения отношения порядка на множестве параметров сложных параметрических моделей, таких как нейросеть. Рассматривается порядок, заданный при помощи ковариационной матрицы градиентов функции ошибки по параметрам модели [37]. В работе [2] предложен итерационный алгоритм для поиска ковариационной матрицы градиентов. Данный итерационный алгоритм интегрируется в градиентный алгоритм оптимизации Adam [38].

1.1. Привилегированное обучение Вапника и дистилляция Хинтона

Задано множество объектов Ω и множество целевых переменных \mathbb{Y} . Множество $\mathbb{Y} = \{1, \dots, R\}$ для задачи классификации, где R число классов, множество $\mathbb{Y} = \mathbb{R}$ для задачи регрессии. Для каждого объекта из $\omega_i \in \Omega$ задана целевая переменная $\mathbf{y}_i = \mathbf{y}(\omega_i)$. Множество целевых переменных для всех объектов обозначим \mathbf{Y} . Для множества Ω задано отображение в некоторое признаковое пространство \mathbb{R}^n :

$$\varphi : \Omega \rightarrow \mathbb{R}^n, \quad |\Omega| = m,$$

где n размерность признакового пространства, а m количество объектов в множестве Ω . Отображение φ отображает объект $\omega_i \in \Omega$ в соответствующий ему вектор признаков $\mathbf{x}_i = \varphi(\omega_i)$. Пусть для объектов $\Omega^* \subset \Omega$ задана привилегированная информация:

$$\varphi^* : \Omega^* \rightarrow \mathbb{R}^{n^*}, \quad |\Omega^*| = m^*,$$

где $m^* \leq m$ — число объектов с привилегированной информацией, n^* — число признаков в пространстве привилегированной информации. Отображение φ^* отображает объект $\omega_i \in \Omega^*$ в соответствующий ему вектор признаков $\mathbf{x}_i^* = \varphi^*(\omega_i)$.

Множество индексов объектов с известной привилегированной информацией обозначим \mathcal{I} :

$$\mathcal{I} = \{1 \leq i \leq m \mid \text{для } i\text{-го объекта задана привилегированная информация}\},$$

а множество индексов объектов с не известной привилегированной информацией обозначим $\{1, \dots, m\} \setminus \mathcal{I} = \bar{\mathcal{I}}$.

Пусть на множестве привилегированных признаков задана функция учителя $\mathbf{f}(\mathbf{x}^*)$:

$$\mathbf{f} : \mathbb{R}^{n^*} \rightarrow \mathbb{Y}^*,$$

где $\mathbb{Y}^* = \mathbb{Y}$ для задачи регрессии и \mathbb{Y}^* является единичным симплексом \mathcal{S}_R в пространстве размерности R для задачи классификации. Модель учителя \mathbf{f} ставит объекты \mathbf{X}^* в соответствие объектам \mathbf{S} , то есть $\mathbf{f}(\mathbf{x}_i^*) = \mathbf{s}_i$.

Требуется выбрать модель ученика $\mathbf{g}(\mathbf{x})$ из множества:

$$\mathfrak{G} = \{\mathbf{g} | \mathbf{g} : \mathbb{R}^n \rightarrow \mathbb{Y}^*\}, \quad (1.1)$$

например для задачи классификации множество \mathfrak{G} может быть параметрическим семейством функций линейных моделей:

$$\mathfrak{G}_{\text{lin,cl}} = \{\mathbf{g}(\mathbf{W}, \mathbf{x}) | \mathbf{g}(\mathbf{W}, \mathbf{x}) = \text{softmax}(\mathbf{W}\mathbf{x}), \quad \mathbf{W} \in \mathbb{R}^{n \times R}\}.$$

Рассмотрим описание метода предложенного в работах [12, 14]. В рамках данных работ предполагается, что для всех данных доступна привилегированная информация $\mathcal{I} = \{1, 2, \dots, m\}$. В работе [12] решается задача классификации вида:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \quad \mathbf{x}_i \in \mathbb{R}^n, \quad y_i \in \mathbb{Y} = \{1, \dots, R\},$$

где y_i — это класс объекта, также обозначим \mathbf{y}_i вектором вероятности для класса y_i .

В постановке Хинтона рассматривается параметрическое семейство функций:

$$\mathfrak{G}_{\text{cl}} = \{\mathbf{g} | \mathbf{g} = \text{softmax}(\mathbf{z}(\mathbf{x})/T), \quad \mathbf{z} : \mathbb{R}^n \rightarrow \mathbb{R}^R\}, \quad (1.2)$$

где \mathbf{z} — это дифференцируемая параметрическая функция заданной структуры, T — параметр температуры. В качестве модели учителя \mathbf{f} рассматривается функция из множества \mathfrak{F}_{cl} :

$$\mathfrak{F}_{\text{cl}} = \{\mathbf{f} | \mathbf{f} = \text{softmax}(\mathbf{v}(\mathbf{x})/T), \quad \mathbf{v} : \mathbb{R}^n \rightarrow \mathbb{R}^R\}, \quad (1.3)$$

где \mathbf{v} — это дифференцируемая параметрическая функция заданной структуры, T — параметр температуры. Параметр температуры T имеет свойства:

1. при $T \rightarrow 0$ получаем вектор, в котором один из классов имеет единичную вероятность;
2. при $T \rightarrow \infty$ получаем равновероятные классы.

Функция потерь \mathcal{L} учитывает перенос информации от модели учителя \mathbf{f} к модели ученика \mathbf{g} имеет вид:

$$\begin{aligned} \mathcal{L}_{st}(\mathbf{g}) &= - \sum_{i=1}^m \underbrace{\sum_{r=1}^R y_i^r \log \mathbf{g}(\mathbf{x}_i)}_{\text{исходная функция потерь}} \Big|_{T=1} \\ &\quad - \sum_{i=1}^m \underbrace{\sum_{r=1}^R \mathbf{f}(\mathbf{x}_i) \Big|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)}_{\text{слагаемое дистилляция}} \Big|_{T=T_0}, \end{aligned} \quad (1.4)$$

где $\cdot|_{T=t}$ обозначает, что параметр температуры T в предыдущей функции равняется t .

Получаем оптимизационную задачу:

$$\hat{\mathbf{g}} = \arg \min_{\mathbf{g} \in \mathfrak{G}_{cl}} \mathcal{L}_{st}(\mathbf{g}). \quad (1.5)$$

Работа [14] обобщает метод предложенный в работе [12]. Решение задачи оптимизации (1.5) зависит только от вектора ответов модели учителя \mathbf{f} . Следовательно признаковые пространства учителя и ученика могут различаться. Получаем постановку задачи:

$$\mathfrak{D} = \{(\mathbf{x}_i, \mathbf{x}_i^*, y_i)\}_{i=1}^m, \quad \mathbf{x}_i \in \mathbb{R}^n, \quad \mathbf{x}_i^* \in \mathbb{R}^{n^*}, \quad y_i \in \{1, \dots, R\},$$

где \mathbf{x}_i это информация доступна на этапах обучения и контроля, а \mathbf{x}_i^* это информация доступна только на этапе обучения. Модель учителя принадлежит множеству моделей \mathfrak{F}_{cl}^* :

$$\mathfrak{F}_{cl}^* = \{\mathbf{f} | \mathbf{f} = \text{softmax}(\mathbf{v}^*(\mathbf{x}^*)/T), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \rightarrow \mathbb{R}^R\}, \quad (1.6)$$

где \mathbf{v}^* — это дифференцируемая параметрическая функция заданной структуры, T — параметр температуры. Множество моделей \mathfrak{F}_{cl}^* отличается от множества моделей \mathfrak{F}_{cl} из выражения (1.3). В множестве \mathfrak{F}_{cl} модели используют пространство исходных признаков, а в множестве \mathfrak{F}_{cl}^* модели используют пространство привилегированных признаков. Функция потерь (1.4) в случае модели учителя $\mathbf{f} \in \mathfrak{F}_{cl}^*$ принимает вид:

$$\mathcal{L}_{st}(\mathbf{g}) = - \sum_{i=1}^m \sum_{r=1}^R y_i^r \log \mathbf{g}(\mathbf{x}_i)|_{T=1} - \sum_{i=1}^m \sum_{r=1}^R \mathbf{f}(\mathbf{x}_i^*)|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)|_{T=T_0}, \quad (1.7)$$

где $\cdot|_{T=t}$ обозначает, что параметр температуры T в предыдущей функции равняется t .

Требуется построить модель, которая использует привилегированную информацию \mathbf{x}_i^* при обучении. Для этого рассмотрим двухэтапную модель обучения предложенную в работе [14]:

1. выбираем оптимальную модель учителя $\mathbf{f} \in \mathfrak{F}_{cl}^*$;
2. выбираем оптимальную модель ученика $\mathbf{g} \in \mathfrak{G}_{cl}$ используя дистилляцию [12].

Модель ученика — это функция минимизирующая (1.7). Модель учителя — это функция минимизирующая кросс-энтропийную функцию ошибки:

$$\mathcal{L}_{th}(\mathbf{f}) = - \sum_{i=1}^m \sum_{r=1}^R y_i^r \log \mathbf{f}(\mathbf{x}_i^*).$$

1.2. Релевантность параметров моделей глубокого обучения

Задана выборка

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}, i = 1, \dots, N,$$

где $\mathbf{x}_i \in \mathbb{R}^m$, $y_i \in \{1, \dots, Y\}$, Y — число классов. Рассмотрим модель $f(\mathbf{x}, \mathbf{w}) : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \{1, \dots, Y\}$, где $\mathbf{w} \in \mathbb{R}^n$ — пространство параметров модели,

$$f(\mathbf{x}, \mathbf{w}) = \text{softmax}(f_1(f_2(\dots(f_l(\mathbf{x}, \mathbf{w})))),$$

где $f_i(\mathbf{x}, \mathbf{w}) = \tanh(\mathbf{w}^\top \mathbf{x})$, l — число слоев нейронной сети, $i \in \{1 \dots l\}$. Параметр w_j модели f называется активным, если $w_j \neq 0$. Множество индексов активных параметров обозначим $\mathcal{A} \subset \mathcal{J} = \{1, \dots, n\}$. Задано пространство параметров модели:

$$\mathbb{W}_{\mathcal{A}} = \{\mathbf{w} \in \mathbb{R}^n | w_j \neq 0, j \in \mathcal{A}\},$$

Для модели f с множеством индексов активных параметров \mathcal{A} и соответствующего ей вектора параметров $\mathbf{w} \in \mathbb{W}_{\mathcal{A}}$ определим логарифмическую функцию правдоподобия выборки:

$$\mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) = \log p(\mathfrak{D} | \mathcal{A}, \mathbf{w}), \quad (1.8)$$

где $p(\mathfrak{D} | \mathcal{A}, \mathbf{w})$ — апостериорная вероятность выборки \mathfrak{D} при заданных \mathbf{w}, \mathcal{A} . Оптимальные значения \mathbf{w}, \mathcal{A} находятся из минимизации $-\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A})$ — логарифма правдоподобия модели:

$$\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}) = \log p(\mathfrak{D} | \mathcal{A}) = \log \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{A}}} p(\mathfrak{D} | \mathbf{w}) p(\mathbf{w} | \mathcal{A}) d\mathbf{w}, \quad (1.9)$$

где $p(\mathbf{w} | \mathcal{A})$ — априорная вероятность вектора параметров в пространстве $\mathbb{W}_{\mathcal{A}}$.

Так как вычисление интеграла (1.9) является вычислительно сложной задачей, рассмотрим вариационный подход [33] для решения этой задачи. Пусть задано распределение q :

$$q(\mathbf{w}) \sim \mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}),$$

где $\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}$ — вектор средних и матрица ковариации, аппроксимирующее неизвестное апостериорное распределение $p(\mathbf{w} | \mathfrak{D}, \mathcal{A})$:

$$p(\mathbf{w} | \mathcal{A}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}),$$

где $\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}$ — вектор средних и матрица ковариации.

Приблизим интеграл (1.9) методом из [33]:

$$\begin{aligned}
\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}) &= \log p(\mathfrak{D}|\mathcal{A}) = \\
&= \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} - \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathfrak{D}, \mathcal{A})}{q(\mathbf{w})} d\mathbf{w} \approx \\
&\approx \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} = \\
&= \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} + \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log p(\mathfrak{D}|\mathcal{A}, \mathbf{w}) d\mathbf{w} = \\
&= \mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) + \mathcal{L}_E(\mathfrak{D}, \mathcal{A}).
\end{aligned} \tag{1.10}$$

Первое слагаемое формулы (1.10) — это сложность модели. Оно определяется расстоянием Кульбака-Лейблера:

$$\mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) = -D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathcal{A})).$$

Второе слагаемое формулы (1.10) является матожиданием правдоподобия выборки $\mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \mathbf{w})$, рассматриваемое в качестве функции ошибки:

$$\mathcal{L}_E(\mathfrak{D}, \mathcal{A}) = \mathbb{E}_{\mathbf{w} \sim q} \mathcal{L}_{\mathfrak{D}}(\mathbf{y}, \mathfrak{D}, \mathcal{A}, \mathbf{w}).$$

Требуется найти параметры, доставляющие минимум суммарному функционалу потерь $\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}, \mathbf{w})$ из (1.10):

$$\begin{aligned}
\hat{\mathbf{w}} &= \arg \min_{\mathcal{A} \subset \mathcal{J}, \mathbf{w} \in \mathbb{W}_{\mathcal{A}}} -\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) = \\
&= \arg \min_{\mathcal{A} \subset \mathcal{J}, \mathbf{w} \in \mathbb{W}_{\mathcal{A}}} D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathcal{A})) - \mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}).
\end{aligned} \tag{1.11}$$

Случайное удаление. Метод случайного удаления заключается в том, что случайным образом удаляется некоторый параметр w_{ξ} из множества активных параметров сети. Индекс параметра ξ из равномерного распределения случайная величина, предположительно доставляющая оптимум в (1.11).

$$\xi \sim \mathcal{U}(\mathcal{A}).$$

Оптимальное прореживание. Метод оптимального прореживания [24] использует вторую производную целевой функции (1.8) по параметрам для определения нерелевантных параметров. Рассмотрим функцию потерь \mathcal{L} (1.8) разложенную в ряд Тейлора в некоторой окрестности вектора параметров \mathbf{w} :

$$\delta \mathcal{L} = \sum_{j \in \mathcal{A}} g_j \delta w_j + \frac{1}{2} \sum_{i,j \in \mathcal{A}} h_{ij} \delta w_i \delta w_j + O(||\delta \mathbf{w}||^3), \tag{1.12}$$

где δw_j — компоненты вектора $\delta \mathbf{w}$, g_j — компоненты вектора градиента $\nabla \mathcal{L}$, а h_{ij} — компоненты гессиана \mathbf{H} :

$$g_j = \frac{\partial \mathcal{L}}{\partial w_j}, \quad h_{ij} = \frac{\partial^2 \mathcal{L}}{\partial w_i \partial w_j}.$$

Задача является вычислительно сложной в силу размерности матрицы \mathbf{H} . Введем предположение [24], о том что удаление нескольких параметров приводит к такому же изменению функции потерь \mathcal{L} , как и суммарное изменение при индивидуальном удалении:

$$\delta\mathcal{L} = \sum_{j \in \mathcal{A}} \delta\mathcal{L}_j,$$

где \mathcal{A} — множество активных параметров, $\delta\mathcal{L}_j$ — изменение функции потерь, при удалении одного параметра \mathbf{w}_j .

В силу данного предположения рассматриваются только диагональные элементы матрицы \mathbf{H} . После введенного предположения, выражение (1.12) принимает вид

$$\delta\mathcal{L} = \frac{1}{2} \sum_{j \in \mathcal{A}} h_{jj} \delta w_j^2,$$

Получаем задачу оптимизации:

$$\xi = \arg \min_{j \in \mathcal{A}} h_{jj} \frac{w_j^2}{2},$$

где ξ — индекс наименее релевантного, удаляемого параметра, предположительно доставляющая оптимум в (1.11).

Удаление неинформативных параметров с помощью вариационного вывода. Для удаления параметров в работе [26] предлагается удалить параметры, которые имеют максимальное отношение плотности $p(\mathbf{w}|\mathcal{A})$ априорной вероятности в нуле к плотности вероятности априорной вероятности в математическом ожидании параметра.

Для гауссовского распределения с диагональной матрицей ковариации получаем:

$$p_j(\mathbf{w}|\mathcal{A})(x) = \frac{1}{\sqrt{2\sigma_j^2}} \exp\left(-\frac{(x - \mu_j)^2}{2\sigma_j^2}\right).$$

Разделив плотность вероятности в нуле к плотности в математическом ожидании

$$\frac{p_j(\mathbf{w}|\mathcal{A})(0)}{p_j(\mathbf{w}|\mathcal{A})(\mu_j)} = \exp\left(-\frac{\mu_j^2}{2\sigma_j^2}\right),$$

Получаем задачу оптимизации:

$$\xi = \arg \min_{j \in \mathcal{A}} \left| \frac{\mu_j}{\sigma_j} \right|,$$

где ξ — индекс наименее релевантного, удаляемого параметра.

1.3. Смесь экспертов для аппроксимации мультиодальной выборки

Исследуется проблема построения смеси экспертов. Смесь экспертов является мультиоделью, состоящей из набора локальных моделей и шлюзовой функции. Локальные модели называются *экспертами*. Смесь экспертов использует шлюзовую функцию для взвешивания прогнозов каждого эксперта. Весовые коэффициенты шлюзовой функции зависят от объекта, для которого производится прогноз. Примерами мультиоделей являются бэггинг, градиентный бустинг [39] и случайный лес [40]. В статье [41] предполагается, что вклад каждого эксперта в ответ зависит от объекта из набора данных.

Большое количество работ в области построения смеси экспертов посвящены выбору шлюзовой функции: используется softmax, процесс Дирихле [42], нейронная сеть [43] с функцией softmax на последнем слое. Ряд работ посвящены выбору моделей в качестве отдельных экспертов. В работах [44, 45] в качестве модели эксперта рассматривается линейная модель. Работы [46, 47] рассматривают модель SVM в качестве модели эксперта. В работе [41] представлен обзор методов и моделей в задачах смеси экспертов.

Смеси экспертов имеют приложения в прикладных задачах. Работы [48, 49, 50] посвящены применению смеси экспертов в задачах прогнозирования временных рядов. В работе [51] предложен метод распознавания рукописных цифр. Метод распознания текстов при помощи смеси экспертов исследуется в работах [52], распознание речи [53, 54, 55]. В работе [52] решается задача классификации текстов. В работах [49, 50, 47, 53, 56, 55, 48], используется смесь экспертов для прогнозирования временных рядов при распознавания человеческой речи, повседневной деятельности человека и прогнозирования стоимости ценных бумаг. В работе [51] смесь экспертов применяется для решения задачи распознавания рукописных чисел на изображениях. В работе [56] исследуется смесь экспертов для задачи распознавания трехмерных движений человека. В [57] описаны работы по исследованию обнаружения радужки глаза на изображении. В работах [58, 59] в частности описаны методы выделения границ радужки и зрачка.

Глава 2

Модели привилегированного обучения и дистилляции

Раздел посвящен методам понижения сложности аппроксимирующих моделей. Предлагается вероятностное обоснование методов дистилляции и привилегированного обучения. В данной главе рассматривается вероятностный подход к решению задачи дистилляции модели и задачи привилегированного обучения. Проанализирована задача обучения модели ученика с помощью модели учителя. Исследован метод дистилляции и привилегированного обучения. Предложено вероятностное обоснование дистилляции.

Приведены общие выводы для произвольной параметрической функции с наперед заданной структурой. Приводится теоретическое обоснование для частных случаев: линейной и логистической регрессии. Подход обобщается на случай, когда привилегированная информация доступна не для всех объектов из обучающей выборки. В рамках вероятностного подхода предлагается анализ и обобщение функции ошибки [12, 14]. Рассматриваются частные задачи классификации и регрессии [10].

В главе введены вероятностные предположения, описывающие дистилляцию моделей. В рамках данных вероятностных предположений анализируются модели для задачи классификации и регрессии. Результат анализа сформулирован в виде теорем 1 и 2. Теорема 2 показала, что обучение линейной регрессии с учителем эквивалентно замене обучающей выборки и вероятностных предположений о распределении истинных ответов. Для задачи классификации ответы учителя дают дополнительную информацию в виде распределения классов для каждого объекта из обучающей выборки. Данная информация не может быть представлена в виде классической задачи классификации. Требуется ввести распределение, которое представлено в теореме 1.

В вычислительном эксперименте анализируются методы как использующие, так и не использующие модель учителя при обучении модели ученика. Проведен анализ ответов модели ученика с использованием модели учителя и без нее.

Анализируются рассмотренные модели на синтетических выборках и реальных данных. В качестве реальных данных рассматриваются выборки FashionMNIST [60] и Twitter Sentiment Analysis [61]. Выборка FashionMNIST [60] является реальной выборкой для задачи классификации изображений, а выборка для Twitter Sentiment. Выборка FashionMNIST использовалась вместо выборки MNIST, так как последняя имеет приемлемое качество аппроксимации даже для линейного классификатора. Analysis [61] — задачи классификации текстов. Вычислительный эксперимент использует модели разной сложности: линейная модель, полно связная нейронная сеть, сверточная нейронная сеть [62], модель Bi-LSTM [63] и модель BERT [6].

Основным результатом данной главы является вероятностная интерпретация задачи дистилляции. Рассмотрен частный случай, когда признаковые описания модели учителя и ученика совпадают.

2.1. Обобщенная вероятностная постановка задачи дистилляции

Задано распределение целевой переменной $p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g})$. Для поиска $\hat{\mathbf{g}}$ воспользуемся методом максимального правдоподобия. В качестве $\hat{\mathbf{g}}$ выбирается функция, которая максимизирует правдоподобие модели:

$$\hat{\mathbf{g}} = \arg \max_{\mathbf{g} \in \mathfrak{G}} \prod_{i=1}^N p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}), \quad (2.1)$$

где множество \mathfrak{G} задается в (1.1).

2.2. Подход дистилляции модели учителя в модель ученика

Рассмотрим вероятностную постановку, в которой выполнены ограничения:

- 1) задано распределение целевой переменной $p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g})$;
- 2) задано совместное распределение целевой переменной и ответов модели учителя $p(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g})$;
- 3) для всех $\omega \in \Omega^*$ элементы $\mathbf{y}(\omega)$ и $\mathbf{s}(\omega)$ являются зависимыми величинами, так как ответы учителя должны коррелировать с истинными ответами;
- 4) если $|\Omega^*| = 0$, то решение должно соответствовать решению (2.1).

Рассмотрим совместное правдоподобие истинных меток и меток учителя:

$$p(\mathbf{Y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}) = \prod_{i \notin \mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) \prod_{i \in \mathcal{I}} p(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}). \quad (2.2)$$

Перепишем $p(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g})$ по формуле условной вероятности:

$$p(\mathbf{y}_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}) = p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g})p(\mathbf{s}_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}). \quad (2.3)$$

Подставляя выражения (2.3) в (2.2), получим

$$p(\mathbf{Y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}) = \prod_{i \notin \mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) \prod_{i \in \mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) \prod_{i \in \mathcal{I}} p(\mathbf{s}_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}).$$

Заметим, что \mathbf{y}_i и \mathbf{s}_i зависят только через переменную \mathbf{x}_i , тогда $p(\mathbf{s}_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}) = p(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g})$. Получаем совместное правдоподобие

$$p(\mathbf{Y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}) = \prod_{i \notin \mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) \prod_{i \in \mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) \prod_{i \in \mathcal{I}} p(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}). \quad (2.4)$$

Используя (2.4), получаем оптимизационную задачу для поиска $\hat{\mathbf{g}}$

$$\hat{\mathbf{g}} = \arg \max_{\mathbf{g} \in \mathfrak{G}} \prod_{i \notin \mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) \prod_{i \in \mathcal{I}} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{g}) \prod_{i \in \mathcal{I}} p(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}). \quad (2.5)$$

Для удобства минимизируется логарифм выражения. Тогда из (2.5) получаем, что

$$\hat{\mathbf{g}} = \arg \max_{\mathbf{g} \in \mathcal{G}} \sum_{i \notin \mathcal{I}} \log p(\mathbf{y}_i | \mathbf{x}_i, \mathbf{g}) + (1 - \lambda) \sum_{i \in \mathcal{I}} \log p(\mathbf{y}_i | \mathbf{x}_i, \mathbf{g}) + \lambda \sum_{i \in \mathcal{I}} \log p(\mathbf{s}_i | \mathbf{x}_i, \mathbf{g}), \quad (2.6)$$

где параметр $\lambda \in [0, 1]$ введен для взвешивания ошибок на истинных ответах и ошибок ответов учителя.

На рис. 2.1 показан вид вероятностной модели в графовой нотации для произвольной функции g .

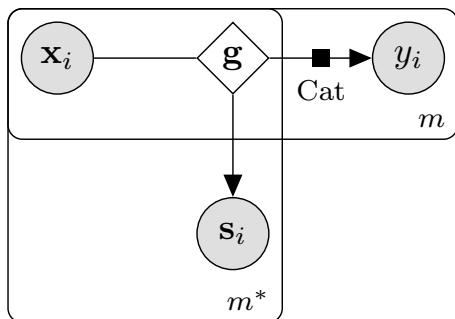


Рис. 2.1: Вероятностная модель в графовой нотации.

Для каждой реализации \mathbf{g} соответствующий блок требует уточнения. На рис. 2.3 показана более подробная реализация в случае, когда \mathbf{g} — линейная модель.

Классификация. Для задачи многоклассовой классификации рассматриваются вероятностные *предположения*:

- 1) рассматривается функция учителя $f \in \mathfrak{F}_{cl}^*$ (1.6);
 - 2) рассматривается функция ученика $g \in \mathfrak{G}_{cl}$ (1.2);
 - 3) для истинных меток рассматривается категориальное распределение $p(y|x, g) = \text{Cat}(g(x))$, где $g(x)$ задает вероятность каждого класса;
 - 4) для меток учителя введем плотность распределения

$$p(\mathbf{s}|\mathbf{x}, \mathbf{g}) = C \prod_{r=1}^R g_r(\mathbf{x})^{s_r}, \quad (2.7)$$

где g^r — вероятность класса r , которую предсказывает модель ученика, а s^r — вероятность класса r , которую предсказывает модель учителя.

Теорема 1. Пусть вероятность каждого класса отделима от нуля и единицы, т.е. для всех r выполняется условие

$$1 > 1 - \varepsilon > g_r(\mathbf{x}) > \varepsilon > 0.$$

Тогда при

$$C = (-1)^R \frac{R^{R/2}}{2^{R(R-1)/2}} \prod_{r=1}^R g_r(\mathbf{x}) \log g_r(\mathbf{x})$$

функция $p(\mathbf{s}|\mathbf{x}, \mathbf{g})$, определенная в (2.7), является плотностью распределения.

Доказательство. Во-первых, покажем, что для произвольного вектора ответов $\mathbf{s} \in \mathcal{S}_R$ выполняется $p(\mathbf{s}|\mathbf{x}, \mathbf{g}) \geq 0$. Заметим, что для всех r выполняется

$$\log g_r(\mathbf{x}) < 0,$$

тогда

$$C = \underbrace{\frac{R^{R/2}}{2^{R(R-1)/2}}}_{>0} \prod_{r=1}^R \underbrace{g_r(\mathbf{x})}_{>\varepsilon} \underbrace{(-\log g_r(\mathbf{x}))}_{>0} > 0.$$

Так как $g_r(\mathbf{x}) > 0$ и $C > 0$, получаем, что $p(\mathbf{s}|\mathbf{x}, \mathbf{g}) \geq 0$. Во-вторых, покажем, что интеграл по всему пространству ответов \mathcal{S}_R является конечным:

$$\begin{aligned} \int_{\mathcal{S}_R} p(\mathbf{s}|\mathbf{x}, \mathbf{g}) d\mathbf{s} &= \int_{\mathcal{S}_R} \prod_{r=1}^R g_r(\mathbf{x})^{s^r} d\mathbf{s} = \prod_{r=1}^R \int_{\mathcal{S}_R} g_r(\mathbf{x})^{s^r} d\mathbf{s} = \\ &= \prod_{r=1}^R \int_0^1 \frac{r^{R-1} \sqrt{R}}{(R-1)! \sqrt{2^{R-1}}} g_r(\mathbf{x})^r dr = \prod_{r=1}^R \underbrace{\frac{\sqrt{R}}{(R-1)! \sqrt{2^{R-1}}}}_D \int_0^1 r^{R-1} g_r(\mathbf{x})^r dr = \\ &= D^R \prod_{r=1}^R \int_0^1 r^{R-1} \exp(r \log g_r(\mathbf{x})) dr = \\ &= (-D)^R \prod_{r=1}^R \log g_r(\mathbf{x}) (\Gamma(R) - \Gamma(R, -\log g_r(\mathbf{x}))) = \\ &= (-D)^R (R-1)!^R \prod_{r=1}^R \log g_r(\mathbf{x}) (1 - g_r(\mathbf{x}) \exp_{R-1}(-\log g_r(\mathbf{x})) + g_r(\mathbf{x})) = \\ &= \frac{(-\sqrt{R})^R}{2^{R(R-1)/2}} \prod_{r=1}^R \log g_r(\mathbf{x}) (1 - g_r(\mathbf{x}) \exp_{R-1}(-\log g_r(\mathbf{x})) + g_r(\mathbf{x})) < \infty, \end{aligned} \tag{2.8}$$

где $\Gamma(R)$ является гамма-функцией, $\Gamma(R, -\log g_r(\mathbf{x}))$ является неполной гамма функцией, $\exp_r(x)$ является суммой Тейлора из первых r слагаемых. В рамках

приближенных расчетов считается, что $\exp_r(x) \approx \exp(x)$, тогда с учетом (2.8) получаем

$$C(\mathbf{g}, \mathbf{x}) = \int_{\mathcal{S}_R} p(\mathbf{s}|\mathbf{x}, \mathbf{g}) ds \approx (-1)^R \frac{R^{R/2}}{2^{R(R-1)/2}} \prod_{r=1}^R g_r(\mathbf{x}) \log g_r(\mathbf{x}). \quad (2.9)$$

Полученное выражение (2.9) заканчивает доказательство теоремы 1. \square

Из теоремы 1 следует, что плотность, введенная для меток учителя, является плотностью распределения. Поэтому можно воспользоваться выражением (2.6). Используя предположения 1–4 и подставляя в (2.6), получаем оптимизационную задачу:

$$\begin{aligned} \hat{\mathbf{g}} = \arg \max_{\mathbf{g} \in \mathcal{G}} & \sum_{i \notin \mathcal{I}} \sum_{r=1}^R y_i^r \log g_r(\mathbf{x}_i) \Big|_{T=1} + \\ & + (1 - \lambda) \sum_{i \in \mathcal{I}} \sum_{r=1}^R y_i^r \log g_r(\mathbf{x}_i) \Big|_{T=1} + \lambda \sum_{i \in \mathcal{I}} \sum_{r=1}^R s_{i,r} \log g_r(\mathbf{x}_i) \Big|_{T=T_0} + \\ & + \lambda \sum_{i \in \mathcal{I}} \sum_{r=1}^R \left(\log g_r(\mathbf{x}_i) \Big|_{T=T_0} + \log \log \frac{1}{g_r(\mathbf{x}_i)} \Big|_{T=T_0} \right). \end{aligned} \quad (2.10)$$

Проанализировав выражение (2.10), получаем, что первые три слагаемых совпадают со слагаемыми в выражении (1.4) при $\mathcal{I} = \{1, \dots, m\}$ и $\lambda = \frac{1}{2}$, а четвертое слагаемое является некоторым регуляризатором, который получен из вида распределения. Анализируя первые три слагаемых в выражении (2.10) при $T_0 = 1$, получаем сумму кросс-энтропий между двумя распределениями для каждого объекта:

- 1) первое распределение — это выпуклая комбинация с весами $1 - \lambda$ и λ распределения, задаваемого метками объектов $\text{Cat}(\mathbf{y})$, и распределения, задаваемого моделью учителя $\text{Cat}(\mathbf{s})$;
- 2) второе распределение — это распределение, задаваемое моделью ученика $\text{Cat}(\mathbf{g}(\mathbf{x}))$.

Следовательно, модель ученика восстанавливает плотность не исходных меток, а новую плотность, которая является выпуклой комбинацией плотности исходных меток и меток учителя.

Регрессия. Для задачи регрессии рассматриваются вероятностные *пределожения*:

- 1) рассматривается функция учителя $\mathbf{f} \in \mathfrak{F}_{rg}^*$,

$$\mathfrak{F}_{rg}^* = \{ \mathbf{f} | \mathbf{f} = \mathbf{v}^*(\mathbf{x}^*), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \rightarrow \mathbb{R} \},$$

где \mathbf{v}^* — дифференцируемая параметрическая функция;

2) рассматривается функция ученика $\mathbf{g} \in \mathfrak{G}_{rg}$,

$$\mathfrak{G}_{rg} = \{\mathbf{g} | \mathbf{g} = \mathbf{z}(\mathbf{x}), \quad \mathbf{z} : \mathbb{R}^n \rightarrow \mathbb{R}^R\},$$

где \mathbf{z} — дифференцируемая параметрическая функция;

3) истинные метки имеют нормальное распределение

$$p(y|\mathbf{x}, \mathbf{g}) = \mathcal{N}(y|\mathbf{g}(\mathbf{x}), \sigma);$$

4) метки учителя имеют распределение

$$p(s|\mathbf{x}, \mathbf{g}) = \mathcal{N}(s|\mathbf{g}(\mathbf{x}), \sigma_s).$$

Используя предположения 1–4 и подставляя в (2.6), получаем оптимизационную задачу:

$$\begin{aligned} \hat{g} = \arg \min_{g \in \mathcal{G}} & \sum_{i \notin \mathcal{I}} \sigma^2 (y_i - g(\mathbf{x}_i))^2 + \\ & + (1 - \lambda) \sum_{i \in \mathcal{I}} \sigma^2 (y_i - g(\mathbf{x}_i))^2 + \lambda \sum_{i \in \mathcal{I}} \sigma_s^2 (s_i - g(\mathbf{x}_i))^2. \end{aligned} \quad (2.11)$$

Выражение (2.11) записано с точностью до аддитивной константы относительно \mathbf{g} .

Теорема 2. Пусть множество \mathcal{G} описывает класс линейных функций вида $\mathbf{g}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. Тогда решение оптимизационной задачи (2.11) эквивалентно решению задачи линейной регрессии:

$$\mathbf{y}'' = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad (2.12)$$

где $\Sigma^{-1} = \text{diag}(\boldsymbol{\sigma}')$ и \mathbf{y}'' имеют вид:

$$\begin{aligned} \sigma'_i &= \begin{cases} \sigma^2, & \text{если } i \notin \mathcal{I}, \\ (1 - \lambda)\sigma^2 + \lambda\sigma_s^2, & \text{иначе,} \end{cases} \\ \mathbf{y}'' &= \Sigma \mathbf{y}', \\ y'_i &= \begin{cases} \sigma^2 y_i, & \text{если } i \notin \mathcal{I}, \\ (1 - \lambda)\sigma^2 y_i + \lambda\sigma_s^2 s_i, & \text{иначе.} \end{cases} \end{aligned} \quad (2.13)$$

Доказательство. Обозначим $\mathbf{a}_{\mathcal{J}} = [a_i | i \in \mathcal{J}]^\top$, где \mathbf{a} — произвольный вектор, а \mathcal{J} — произвольное непустое индексное множество. Подвектор вектора ответов \mathbf{y} , для элементов которого доступна привилегированная информация, обозначим $\mathbf{y}_{\mathcal{I}} = [y_i | i \in \mathcal{I}]^\top$. Аналогично обозначим матрицу $\mathbf{X}_{\mathcal{I}} = [\mathbf{x}_i | i \in \mathcal{I}]^\top$.

В случае линейной модели $\mathbf{g}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ выражение (2.11) принимает вид:

$$\begin{aligned}\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{W}} & \sigma^2 (\mathbf{y}_{\bar{\mathcal{I}}} - \mathbf{X}_{\bar{\mathcal{I}}} \mathbf{w})^\top (\mathbf{y}_{\bar{\mathcal{I}}} - \mathbf{X}_{\bar{\mathcal{I}}} \mathbf{w}) + \\ & + \sigma^2 (1 - \lambda) (\mathbf{y}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}} \mathbf{w})^\top (\mathbf{y}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}} \mathbf{w}) + \sigma_s^2 \lambda (\mathbf{s}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}} \mathbf{w})^\top (\mathbf{s}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}} \mathbf{w}).\end{aligned}$$

Раскроем скобки и сгруппируем:

$$\begin{aligned}\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{W}} & \sigma^2 (\mathbf{w}^\top \mathbf{X}_{\bar{\mathcal{I}}}^\top \mathbf{X}_{\bar{\mathcal{I}}} \mathbf{w} - 2 \mathbf{y}_{\bar{\mathcal{I}}}^\top \mathbf{X}_{\bar{\mathcal{I}}} \mathbf{w}) + \\ & + (1 - \lambda) \sigma^2 (\mathbf{w}^\top \mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}} \mathbf{w} - 2 \mathbf{y}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}} \mathbf{w}) + \lambda \sigma_s^2 (\mathbf{w}^\top \mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}} \mathbf{w} - 2 \mathbf{s}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}} \mathbf{w}).\end{aligned}$$

Продифференцируем выражение, приравняем к нулю и сгруппируем элементы:

$$\begin{aligned}(\sigma^2 \mathbf{X}_{\bar{\mathcal{I}}}^\top \mathbf{X}_{\bar{\mathcal{I}}} + (1 - \lambda) \sigma^2 \mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}} + \lambda \sigma_s^2 \mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}}) \mathbf{w} = & 2 \sigma^2 \mathbf{X}_{\bar{\mathcal{I}}}^\top \mathbf{y}_{\bar{\mathcal{I}}} + \\ & + 2(1 - \lambda) \sigma^2 \mathbf{X}_{\mathcal{I}}^\top \mathbf{y}_{\mathcal{I}} + 2 \lambda \sigma_s^2 \mathbf{X}_{\mathcal{I}}^\top \mathbf{s}_{\mathcal{I}}.\end{aligned}\tag{2.14}$$

Воспользуемся равенствами:

$$\begin{aligned}\sigma^2 \mathbf{X}_{\bar{\mathcal{I}}}^\top \mathbf{X}_{\bar{\mathcal{I}}} + (1 - \lambda) \sigma^2 \mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}} + \lambda \sigma_s^2 \mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}} &= \mathbf{X}^\top \Sigma^{-1} \mathbf{X}, \\ 2 \sigma^2 \mathbf{X}_{\bar{\mathcal{I}}}^\top \mathbf{y}_{\bar{\mathcal{I}}} + 2(1 - \lambda) \sigma^2 \mathbf{X}_{\mathcal{I}}^\top \mathbf{y}_{\mathcal{I}} + 2 \lambda \sigma_s^2 \mathbf{X}_{\mathcal{I}}^\top \mathbf{s}_{\mathcal{I}} &= 2 \mathbf{X} \mathbf{y}',\end{aligned}\tag{2.15}$$

где Σ и \mathbf{y}' из условия задачи (2.13).

Подставляя (2.15) в (2.14), получаем:

$$\mathbf{w} = 2 (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X} \Sigma^{-1} \mathbf{y}',$$

что соответствует решению задачи (2.12). Теорема 2 доказана. \square

Теорема 2 показывает, что обучение с учителем для задачи регрессии можно свести к задаче оптимизации в линейной регрессии.

2.3. Анализ вероятностного подхода к дистилляции линейных моделей

Проводится вычислительный эксперимент для анализа моделей, которые получены путем дистилляции модели учителя в модель ученика. Как показано в теореме 2, задачу регрессии с учителем можно свести к задаче регрессии без учителя, поэтому в эксперименте рассматривается только случай классификации. Во всех частях вычислительного эксперимента для поиска оптимальных параметров нейросетей использовался градиентный метод оптимизации Adam [38].

Выборка FashionMNIST. Эксперимент проводился для задачи классификации для выборки FashionMNIST [60]. В качестве модели учителя \mathbf{f} рассматривается нейросеть с двумя сверточными слоями и с тремя полно связанными слоями, в качестве функции активации рассматривается ReLu. Модель учителя содержит 30 тысяч обучаемых параметров. В качестве модели ученика рассматривается модель логистической регрессии для многоклассовой классификации. Модель ученика содержит 7850 обучаемых параметров.

На рис. 2.2 показан график зависимости кросс-энтропии между истинными метками объектов и вероятностями, которые предсказывает модель ученика. На графике сравнивается модель, которая обучалась без учителя (в задаче оптимизации (2.10) присутствует только первое слагаемое) с моделью, которая получена путем дистилляции модели нейросети в линейную модель. Из графика видно, что обе модели начинают переобучаться после 30-й итерации. Но модель, которая получена путем дистилляции, переобучается не так быстро: ошибка на тестовой выборке растет медленнее, а на обучающей выборке падает также медленнее.

В таблице показано, что для выборки FashionMnist итоговые модели ученика с учителем и без учителя сравнимы по точности и кросс-энтропийной ошибке, если учитывать дисперсию этих величин.

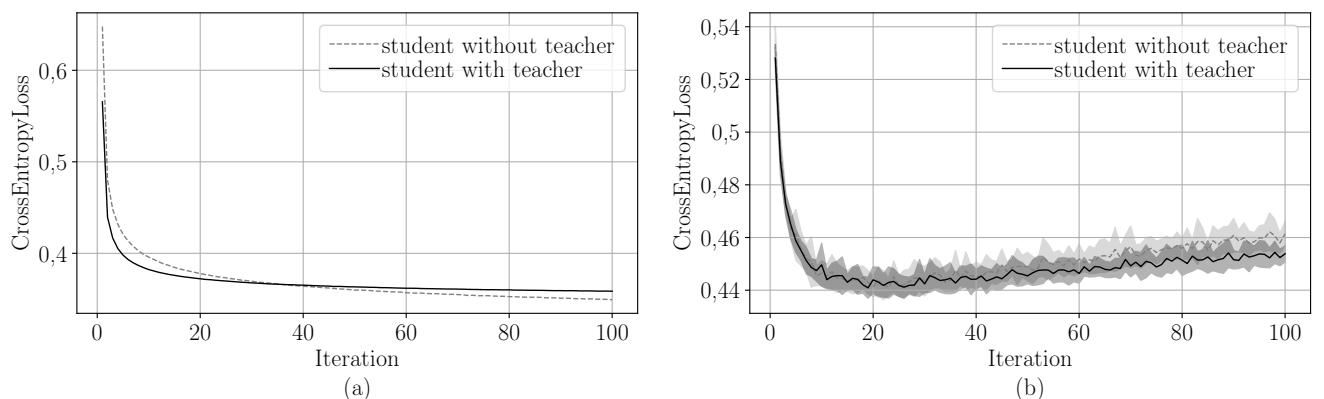


Рис. 2.2: Зависимость кросс-этропии между истинными метками и предсказанными учеников вероятностями классов: а) на обучающей выборке; б) на тестовой выборке

На рис. 2.2 показан график зависимости кросс-энтропии между истинными метками объектов и вероятностями предсказанными модель ученика. На графике сравнивается модель, которая обучалась без учителя (в задаче оптимизации (2.10) присутствует только первое слагаемое) с моделью полученной путем дистилляции модели нейросети в линейную модель. На графике видно, что обе модели начинают переобучаться после 30–й итерации, но модель, которая получена путем дистилляции переобучается не так быстро, что следует из того, что ошибка на тестовой выборке растет медленней, а на обучающей выборке падает также медленней.

Синтетический эксперимент. Проанализируем модель на синтетической выборке. Гипотеза порождения данных:

$$\mathbf{W} = [\mathcal{N}(w_{jr}|0, 1)]_{n \times R}, \quad \mathbf{X} = [\mathcal{N}(x_{ij}|0, 1)]_{m \times n}, \\ \mathbf{S} = \text{softmax}(\mathbf{X}\mathbf{W}), \quad \mathbf{y} = [\text{Cat}(y_i|\mathbf{s}_i)],$$

где функция softmax берется построчно. Строки матрицы \mathbf{S} рассматриваются как предсказание учителя, т.е. учитель знает истинные вероятности каждого класса. На рис. 2.3 показана вероятностная модель в графовой нотации. В эксперименте число признаков $n = 10$, число классов $R = 3$, для обучения сгенерировано $m_{\text{train}} = 1000$ и $m_{\text{test}} = 100$ объектов.

На рис. 2.4 показано распределение по классам для 20 объектов из обучающей выборки. Каждому столбцу на графике соответствует объект, а каждой строке соответствует вероятность класса. Видно, что для каждого рассмотренного объекта вероятности разных классов близки. Получается, что если в качестве истинных меток взять класс с максимальной вероятностью, то выборка будет сильно зашумленной и модель будет описывать эти данные некорректно.

Построим в качестве ученика линейную модель, которая минимизирует кросс-энтропийную (первое слагаемое в формуле (2.10)). Представление данной модели в виде графовой модели показано на рис. 2.3.

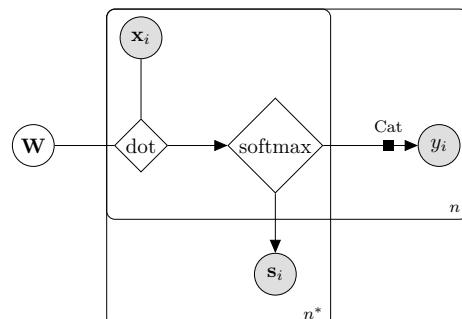


Рис. 2.3: Вероятностная модель используемая в синтетическом эксперименте

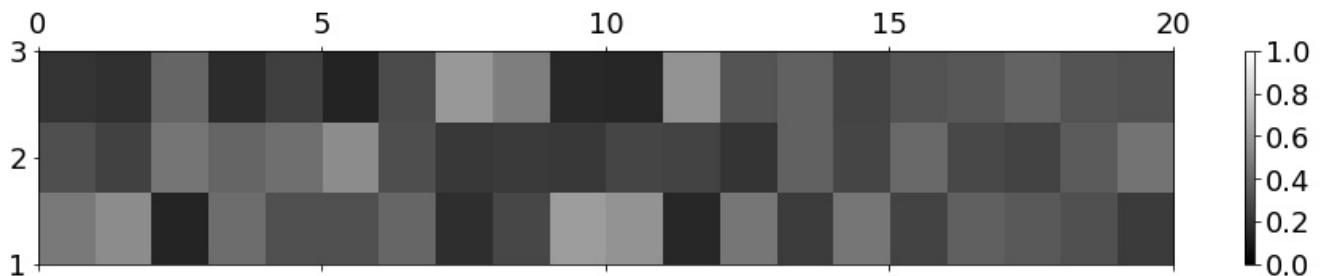


Рис. 2.4: Истинное распределение объектов по классам

На рис. 2.5 показано распределение вероятностей классов, которое предсказала модель. Видно, что полученное распределение не соответствует истинному, так как модель сосредотачивает всю вероятность в одном классе.

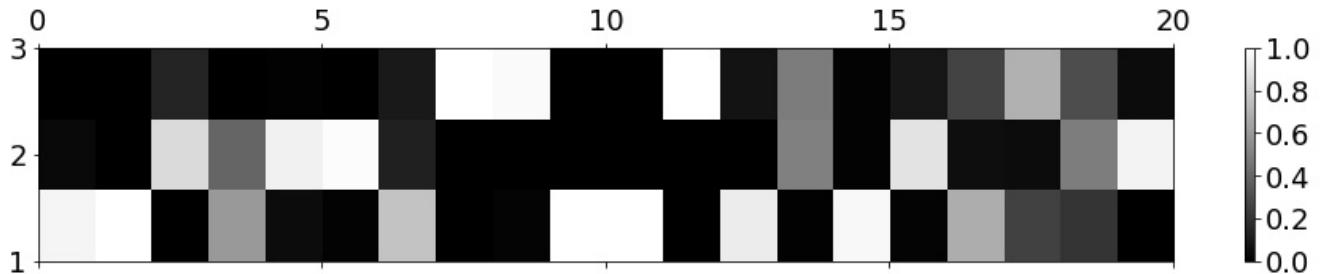


Рис. 2.5: Распределение предсказанное моделью без использования информации об истинном распределении на классах

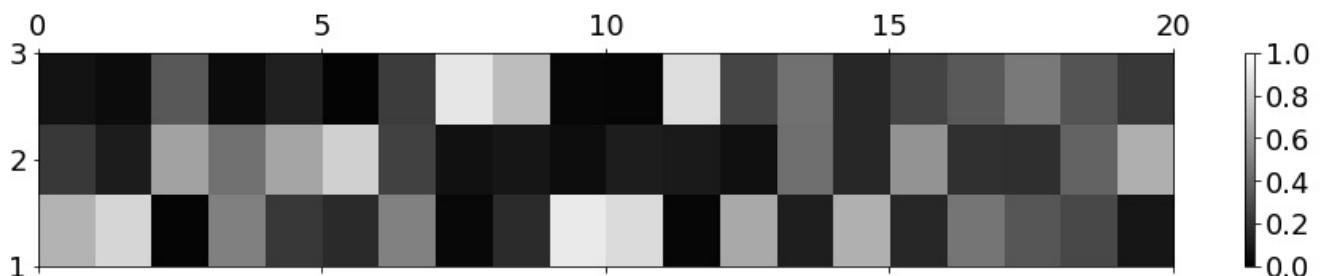


Рис. 2.6: Распределение предсказанное моделью с использованием информации об истинном распределении на классах

Рассмотрим модель, которая учитывает информацию об истинных распределениях на классах для каждого объекта. Для этого минимизируются первые три слагаемых в формуле (2.10) при $T_0 = 1$ и $\lambda = 0,75$. В качестве меток учителя $s_{i,r}$ использовались истинные вероятности для каждого класса данного объекта. На рис. 2.6 показано распределение, которое дала модель. В данном случае видно, что распределения являются сглаженными. Концентрации всей вероятности в одном классе не наблюдается.

Заметим, что в данном примере предполагается, что модель учителя учитывает не только метки классов, но и распределение на метках классов, в то время как в выборке $\{\mathbf{X}, \mathbf{y}\}$ имеются только точечные оценки в виде меток.

В данном примере используются истинные распределения в качестве предсказаний учителя, но их можно заменить предсказаниями модели учителя, которая предсказывает не только сами метки, но и их распределение для каждого объекта.

На рис. 2.7 показана зависимость вероятности верного класса от температуры T и параметра доверия λ для одного из объектов из тестовой выборки. На рис. 2.7 видно, что изменение температуры T влечет изменение концентрации вероятностной меры. При уменьшении параметра температуры и приближении его к нулю наблюдаем, что вероятность одного из классов приближается к единице, а остальных классов — к нулю. С другой стороны, при увеличении параметра температуры вероятности классов сглаживаются и распределение

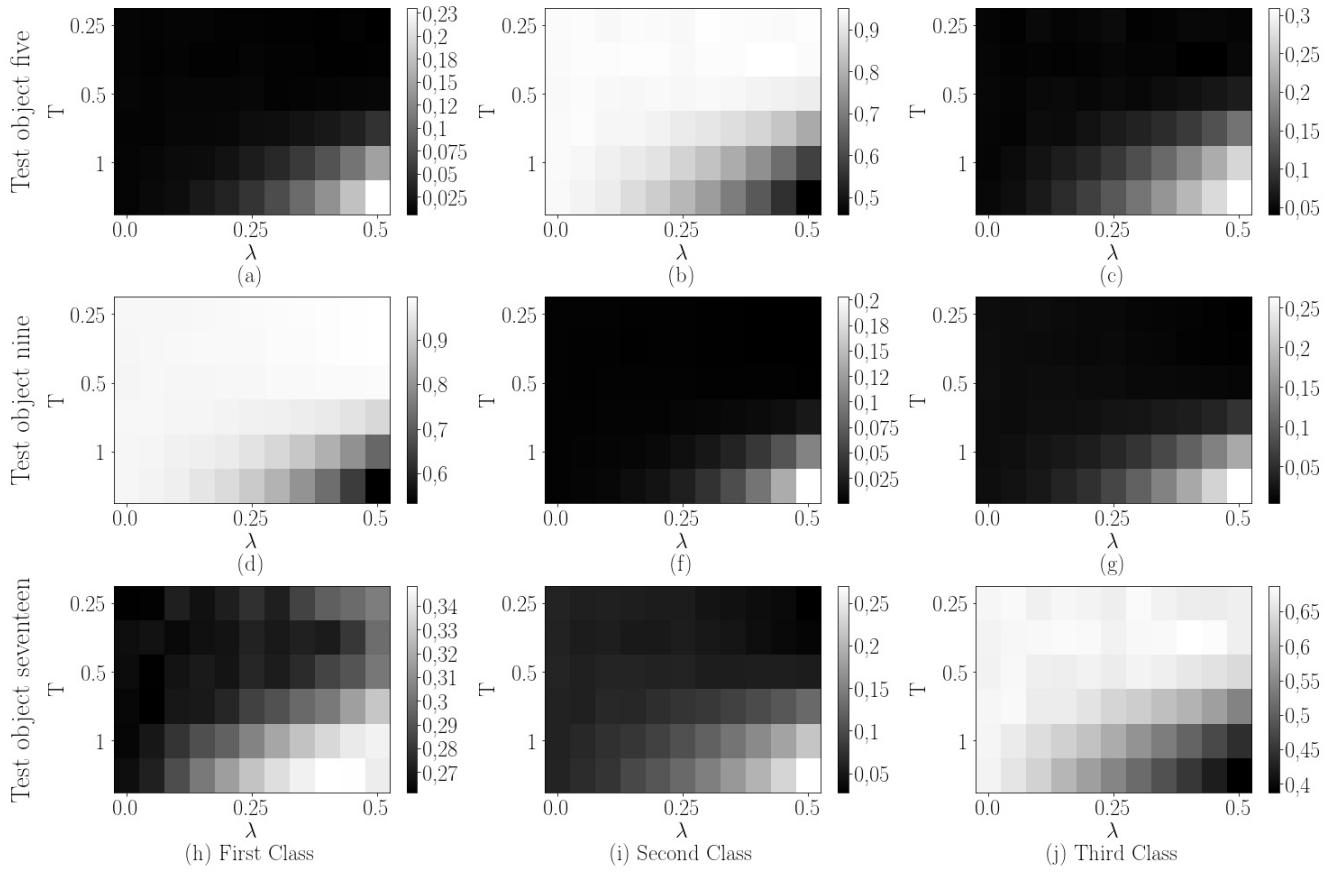


Рис. 2.7: Вероятности классов для разных объектов

классов для каждого объекта становится близким к равномерному.

В таблице 2.1 в колонке “Кросс-энтропийная ошибка с реальными вероятностями” показано сравнение кросс-энтропии в случае, если в качестве истинных вероятностей меток рассмотреть не onehot-кодированные вероятности классов, а истинные вероятности:

$$\mathcal{L}_{\text{real}}(\mathbf{g}) = - \sum_{i=1}^m \sum_{r=1}^R s_i^r \log g_r(\mathbf{x}_i),$$

где \mathbf{g} — модель ученика. Видно, что модель с учителем лучше аппроксимирует истинные вероятности классов. Также в таблице 2.1 представлено среднее значение разницы максимальной вероятности с минимальной вероятностью для каждого объекта:

$$\mathcal{L}_{\text{maxmin}}(\mathbf{g}) = \frac{1}{m} \sum_{i=1}^m \left(\max_r g_r(\mathbf{x}_i) - \min_r g_r(\mathbf{x}_i) \right).$$

Видно, что модель учителя имеет меньшую разницу между вероятностями классов, т.е. вероятности классов не концентрируются в одном классе.

Анализ твитов пользователей. Проводится эксперимент на выборке Twitter Sentiment Analysis. Данная выборка содержит короткие сообщения, для

Таблица 2.1: Сводная таблица результатов вычислительного эксперимента

Выборка	Модель	Кросс-Энтропийная ошибка	Кросс-Энтропийная ошибка с реальными вероятностями	Вероятностная разница	Точность	Число Параметров
FashionMnist	с учителем	$0,453 \pm 0,003$	-	$0,84 \pm 0,13$	$0,842 \pm 0,002$	7850
	без учителя	$0,461 \pm 0,005$	-	$0,86 \pm 0,18$	$0,841 \pm 0,002$	7850
Systetic	с учителем	$0,618 \pm 0,001$	$1,17 \pm 0,05$	$0,45 \pm 0,20$	$0,828 \pm 0,002$	33
	без учителя	$0,422 \pm 0,002$	$2,64 \pm 0,02$	$0,75 \pm 0,22$	$0,831 \pm 0,001$	33
Twiter	с учителем	$0,489 \pm 0,003$	-	$0,79 \pm 0,17$	$0,764 \pm 0,005$	1538
	без учителя	$0,501 \pm 0,006$	-	$0,83 \pm 0,22$	$0,747 \pm 0,004$	1538

которых требуется предсказать эмоциональный окрас: содержит твит позитивный окрас или негативный. Выборка разделена на 1,18 млн твитов для обучения и 0,35 млн твитов для тестирования. Выполнена предобработка твитов: все твиты переведены в нижний регистр, все никнеймы вида “@andrey” заменены на токен “name”, все цифры заменены на токен “number”.

Результаты данной части эксперимента показаны в таблице. В качестве модели учителя использовалась модель Bi-LSTM с линейным слоем на выходе. В качестве векторного представления токенов обучалась матрица параметров. В ней каждая строка соответствует токену из обучающей выборки. Суммарное число обучаемых параметров модели учителя составляет более 30 млн. Обученная модель учителя имеет точность предсказания 0,835. В качестве модели ученика рассматривается линейная модель с 1538 параметрами, где в качестве векторного представления предложения рассматривается выход предобученной модели BERT с размерностью векторного пространства 768. Признаковое описание модели учителя и модели ученика различаются. Модель учителя в качестве признакового описания рассматривает исходные слова в предложении. Модель ученика в качестве признакового описания использует готовое векторное представление предложения, которое получено при помощи модели BERT.

В таблице 2.1 показано качество модели ученика с использованием предсказания модели учителя и без него. В рамках данных результатов качество модели ученика с дистилляцией выше, чем модели ученика без дистилляции, но разница находится в пределах погрешности, что не позволяет говорить о значительных улучшениях качества.

В таблице 2.1 показаны результаты вычислительного эксперимента для разных выборок. Из результатов эксперимента видно, что модель ученика наследует распределение вероятностей по классам от модели учителя. Когда модель учителя адекватно описывает данные, то описание данных моделью ученика также улучшается, что показано в вычислительном эксперименте на синтетических данных. Показано, что точность аппроксимации выборки учеником повышается при использовании модели учителя. Задача регрессии не приведена в вычислительном эксперименте, так как в теореме 2 показана ее эквивалентность задаче линейной регрессии. Для задачи классификации проведен вычислительный эксперимент. Из вычислительного эксперимента видно, что дистилляция

влияет на распределение классов в рамках одного объекта. Вероятности классов для каждого объекта являются более разреженными, а не концентрируются в одном классе. Данное свойство хорошо видно в синтетической выборке, так как она генерировалась с максимальной дисперсией в вероятностях классов.

Глава 3

Байесовская дистилляция моделей глубокого обучения

Исследуется проблема понижения сложности аппроксимирующих моделей. Рассматриваются методы, основанные на дистилляции моделей глубокого обучения. Вводятся понятия учителя и ученика. Предполагается, что модель ученика имеет меньшее число параметров, чем модель учителя. Предлагается байесовский подход к выбору модели ученика. Предложен метод назначения априорного распределения параметров ученика на основе апостериорного распределения параметров модели учителя. Так как пространства параметров учителя и ученика не совпадают, предлагается механизм выравнивания пространства параметров модели учителя и пространства параметров модели ученика путем изменения структуры модели учителя. В данном разделе приводится теоретический анализ предложенного механизма выравнивания.

В данной главе представлены методы основанный на байесовском выводе. В качестве априорного распределения параметров модели ученика предлагается использовать апостериорное распределение параметров модели учителя. Решается задача выравнивание пространства параметров модели учителя и модели ученика. Предлагается подход, основанный на последовательном выравнивании пространств параметров модели ученика и учителя.

Определение 7. *Структура модели – упорядоченный набор структурных параметров модели, которые задают вид суперпозиции.*

Определение 8. *Выравнивание параметрических моделей – изменение структуры модели (одной или нескольких моделей) в результате которого векторы параметров различных моделей лежат в одном пространстве.*

В следствие этого выравнивания в качестве априорного распределения параметров модели ученика выбирается апостериорное распределение параметров модели учителя. В данной работе в качестве параметрической моделей рассматривается полносвязная нейронная сеть и рекурентная нейронная сеть. В качестве структурных параметров модели выбраны число слоев, а также размер каждого скрытого слоя.

В рамках предложенного метода выравнивания параметрических моделей не оговорен выбор порядка на множестве параметров модели учителя. Для этого предлагается упорядочивать параметры модели учителя на основе их релевантности. Первый нейрон является наиболее релевантным, а последний нейрон наименее релевантным. Порядок задается на основе отношения плотности распределения упорядочиваемого параметра к плотности распределения параметра в нуле [26] или на основе метода Белсли [27].

В рамках вычислительного эксперимента проводится теоретический анализ. Предложенный метод дистилляции анализируется на примере синтетической выборки, а также на реальной выборке FashionMnist [60].

3.1. Постановка задачи дистилляции в терминах байесовского подхода

Задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \quad \mathbf{x}_i \in \mathbb{R}^n, \quad y_i \in \mathbb{Y},$$

где \mathbf{x}_i, y_i — признаковое описание и целевая переменная i -го объекта, число объектов в обучающей выборке обозначается m . Размер признакового описания объектов обозначается n . Множество $\mathbb{Y} = \{1, \dots, R\}$ для задачи классификации, где R число классов, множество $\mathbb{Y} = \mathbb{R}$ для задачи регрессии.

Задана модель учителя в виде суперпозиций линейных и нелинейных преобразований:

$$f = \boldsymbol{\sigma} \circ \mathbf{U}_T \boldsymbol{\sigma} \circ \mathbf{U}_{T-1} \circ \dots \mathbf{U}_2 \boldsymbol{\sigma} \circ \mathbf{U}_1,$$

где T — число слоев модели учителя, $\boldsymbol{\sigma}$ — функция активации, а \mathbf{U}_t обозначает матрицу линейного преобразования. Матрицы \mathbf{U} соединяются в вектор параметров \mathbf{u} модели учителя f :

$$\mathbf{u} = \text{vec}([\mathbf{U}_T, \mathbf{U}_{T-1}, \dots, \mathbf{U}_1]), \quad (3.1)$$

где vec операция векторизации соединенных матриц. Каждая матрица \mathbf{U}_t имеет размер $n_t \times n_{t-1}$, где $n_0 = n$, а $n_T = 1$ для задачи регрессии и $n_T = R$ для задачи классификации на R классов. Число параметров N_{tr} учителя f

$$N_{\text{tr}} = \sum_{t=1}^T n_t n_{t-1}. \quad (3.2)$$

Для построения вектора параметров \mathbf{u} задается полный порядок на элементов матриц \mathbf{U}_t . Для полносвязной нейронной сети вводится естественный порядок, индуцированный номером слоя t , номером нейрона, и номером элемента вектора параметров нейрона: выбирается матрица \mathbf{U}_t , строка этой матрицы и элемент строки.

Например, для модели учителя в задаче регрессии:

$$f(\mathbf{x}) = \boldsymbol{\sigma} \circ \mathbf{U}_3 \boldsymbol{\sigma} \circ \mathbf{U}_2 \boldsymbol{\sigma} \circ \mathbf{U}_1 \mathbf{x}, \quad (3.3)$$

вектор параметров \mathbf{u} принимает вид

$$\mathbf{u} = [u_1^{1,1}, \dots, u_1^{1,n}, \dots, u_1^{n_1,1}, \dots, u_1^{n_1,n}, u_2^{1,1}, \dots, u_2^{1,n_1}, \dots, u_2^{n_2,1}, \dots, u_2^{n_2,n_1}, u_3^{1,1}, \dots, u_3^{1,n_2}].$$

Пусть для вектора параметров учителя f известно апостериорное распределение параметров $p(\mathbf{u}|\mathfrak{D})$. На основе выборки \mathfrak{D} и апостериорного распределения

параметров учителя f требуется выбрать модель ученика из параметрического семейства функций:

$$g = \boldsymbol{\sigma} \circ \mathbf{W}_L \boldsymbol{\sigma} \circ \dots \circ \mathbf{W}_1, \quad \mathbf{W}_l \in \mathbb{R}^{n_l \times n_{l-1}},$$

где L число слоев модели ученика. Число параметров N_{st} модели ученика g вычисляется аналогично выражению (3.2). Вектор параметров модели ученика \mathbf{w} строится аналогичным образом (3.1). Модель g задается своим вектором параметров \mathbf{w} . Следовательно, задача выбора модели g эквивалентна задаче оптимизации вектора параметров $\mathbf{w} \in \mathbb{R}^{N_{\text{st}}}$.

Параметры $\hat{\mathbf{w}} \in \mathbb{R}^{N_{\text{st}}}$ оптимизируются при помощи вариационного вывода на основе совместного правдоподобия модели и данных:

$$\mathcal{L}(\mathfrak{D}, \mathbf{A}) = \log p(\mathfrak{D}|\mathbf{A}) = \log \int_{\mathbf{w} \in \mathbb{R}^{N_{\text{st}}}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{A})d\mathbf{w}, \quad (3.4)$$

где $p(\mathbf{w}|\mathbf{A})$ — априорное распределение вектора параметров модели ученика. Так как взятие интеграла (3.4) является вычислительно сложной задачей, используется вариационный вывод [26, 27]. Для этого задается вариационное распределение параметров модели ученика $q(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, которое аппроксимирует неизвестное апостериорное распределение $p(\mathbf{w}|\mathfrak{D})$

$$q(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \approx p(\mathbf{w}|\mathfrak{D}).$$

Оптимизация параметров \mathbf{w} сводится к решению задачи:

$$\hat{\mathbf{w}} = \arg \min_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{w}} D_{\text{KL}}(q(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) || p(\mathbf{w}|\mathbf{A})) - \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}). \quad (3.5)$$

Выражение (3.5) не учитывает параметры учителя $f(\mathbf{x}, \mathbf{u})$. Для использования параметров учителя при решении оптимизационной задачи (3.5) предлагается рассмотреть зависимость параметров априорного распределения $p(\mathbf{w}|\mathbf{A})$ от параметров апостериорного распределения учителя $p(\mathbf{u}|\mathfrak{D})$.

3.2. Выравнивание априорного распределения параметров ученика на основе параметров учителя

Апостериорное распределение параметров модели учителя предполагается нормальным:

$$p(\mathbf{u}|\mathfrak{D}) = \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma}),$$

где \mathbf{m} и $\boldsymbol{\Sigma}$ параметры этого распределения. На основе параметров \mathbf{m} и $\boldsymbol{\Sigma}$ требуется задать параметры \mathbf{A} априорного распределения $p(\mathbf{w}|\mathbf{A})$. Структуры моделей учителя и ученика задаются числом слоев и размером этих слоев, то возможны варианты: 1) число слоев и размер каждого слоя совпадает; 2) число слоев совпадает, а размеры различаются; 3) не совпадает число слоев.

Учитель и ученик принадлежат одному семейству. Рассмотрим условия:

- 1) число слоев модели учителя равняется числу слоев модели ученика $L = T$;
- 2) размеры соответствующих слоев совпадают, другими словами, для всех t, l таких, что $t = l$ выполняется $n_l = n_t$, где n_t обозначает размер t -го слоя учителя, а n_l размер l -го слоя ученика.

При выполнении этих условий, априорное распределение параметров модели ученика приравнивается к апостериорному распределению параметров учителя $p(\mathbf{w}|\mathbf{A}) = p(\mathbf{u}|\mathfrak{D})$.

Удаление нейрона в слое учителя. Проведем выравнивание модели учителя и модели ученика, согласно определению 8 при помощи последовательных преобразований параметров \mathbf{u} . Рассмотрим преобразование

$$\phi(t, \mathbf{u}) : \mathbb{R}^{N_{\text{tr}}} \rightarrow \mathbb{R}^{N_{\text{tr}} - 2n_t}$$

вектора \mathbf{u} , которое описывает удаление одного нейрона из t -го слоя учителя. Обозначим новый вектор параметров $\mathbf{v} = \phi(t, \mathbf{u})$, а элементы вектора, которые удалены как $\bar{\mathbf{v}}$. Заметим, что векторы \mathbf{v} и $\bar{\mathbf{v}}$ являются случайными величинами.

Теорема 3. *Пусть задано распределение вектора параметров $p(\mathbf{u})$. Тогда распределение вектора параметров $p(\mathbf{v})$ представимо в виде:*

$$p(\mathbf{v}) = \int_{\boldsymbol{\nu}_2 \in \mathbb{R}^{n_{t-1}}} p(\bar{\boldsymbol{\nu}}_1 | \mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}) d\boldsymbol{\nu}_2.$$

Доказательство. Пусть $\phi(t, \mathbf{u})$ удаляет j -й нейрон в t -м слое, что является удалением j -й строки матрицы \mathbf{U}_t . Заметим, что удаление j -й строки матрицы \mathbf{U}_t влечет удаление j -й компоненты вектора z_{t+1} , где

$$\mathbf{z}_t = \boldsymbol{\sigma} \circ \mathbf{U}_{t-1} \boldsymbol{\sigma} \circ \dots \mathbf{U}_2 \boldsymbol{\sigma} \circ \mathbf{U}_1 \mathbf{x}.$$

Удаление j -й компоненты вектора \mathbf{z}_{t+1} эквивалентно занулению j -го столбца матрицы \mathbf{U}_{t+1} . Заметим, что тогда предсказание модели не зависит от параметров j -й строки матрицы \mathbf{U}_t , а следовательно данными параметрами также можно пренебречь.

Найдем распределение вектора \mathbf{v} . Для поиска распределения вектора параметров после зануления j -го столбца матрицы \mathbf{U}_{t+1} воспользуемся формулой условной вероятности $p(\bar{\boldsymbol{\nu}}_1 | \mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0})$, а для удаления j -й строки матрицы \mathbf{U}_t воспользуемся маргинализацией распределения $p(\bar{\boldsymbol{\nu}}_1 | \mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0})$. Обозначим зануляемые параметры модели как $\boldsymbol{\nu}_1$, а удаляемые параметры как $\boldsymbol{\nu}_2$. Также обозначим все параметры, которые не занулены как $\bar{\boldsymbol{\nu}}_1 = [\mathbf{v}^\top, \boldsymbol{\nu}_2^\top]$. Итоговое распределение параметров принимает вид:

$$p(\mathbf{v} | \mathfrak{D}) = \int_{\boldsymbol{\nu}_2} p(\bar{\boldsymbol{\nu}}_1 | \mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}) d\boldsymbol{\nu}_2.$$

□

Теорема 4. Пусть выполняются условия:

- 1) апостериорное распределение параметров $p(\mathbf{u}|\mathfrak{D}) = \mathcal{N}(\mathbf{m}, \Sigma)$,
- 2) число слоев модели учителя равняется числу слоев модели ученика $T = L$,
- 3) размеры соответствующих слоев не совпадают, другими словами, для всех t, l , таких что $t = l$, выполняется $n_t \geq n_l$.

Тогда распределение параметров $p(\mathbf{v}|\mathfrak{D})$ также является нормальным.

Доказательство. Из свойств нормального распределения следует, что распределение

$$p(\bar{\boldsymbol{\nu}}_1|\mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}) \quad (3.6)$$

также является нормальным распределением с параметрами $\boldsymbol{\mu}, \boldsymbol{\Xi}$:

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{m}_{\bar{\boldsymbol{\nu}}_1} + \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1, \boldsymbol{\nu}_1} \boldsymbol{\Sigma}_{\boldsymbol{\nu}_1, \boldsymbol{\nu}_1}^{-1} (\mathbf{0} - \mathbf{m}_{\boldsymbol{\nu}_1}), \\ \boldsymbol{\Xi} &= \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1, \bar{\boldsymbol{\nu}}_1} - \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1, \boldsymbol{\nu}_1} \boldsymbol{\Sigma}_{\boldsymbol{\nu}_1, \boldsymbol{\nu}_1}^{-1} \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1, \boldsymbol{\nu}_1}, \end{aligned}$$

где векторы $\mathbf{m}_{\bar{\boldsymbol{\nu}}_1}, \mathbf{m}_{\boldsymbol{\nu}_1}$ являются подвекторами вектора \mathbf{m} , который относится к параметрам $\bar{\boldsymbol{\nu}}_1$ и $\boldsymbol{\nu}_1$ соответственно. Ковариационная матрица $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1, \boldsymbol{\nu}_1}$ обозначает подматрицу матрицы $\boldsymbol{\Sigma}$, которая соответствует ковариационной матрице параметров $\bar{\boldsymbol{\nu}}_1$ и $\boldsymbol{\nu}_1$.

Распределение $p(\mathbf{v}|\mathfrak{D})$ найдем при помощи маргинализации распределения (3.6) по параметрам $\boldsymbol{\nu}_2$. Используя свойства нормального распределения получаем распределение

$$p(\mathbf{v}|\mathfrak{D}) = \mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\Xi}_{v,v}), \quad (3.7)$$

где $\boldsymbol{\mu}_v$ обозначает подвектор вектора $\boldsymbol{\mu}$, который относится к параметру \mathbf{v} , а матрица $\boldsymbol{\Xi}_{v,v}$ является подматрицей матрицы $\boldsymbol{\Xi}$, которая относится к вектору параметров \mathbf{v} . \square

Теорема 4 задает апостериорное распределение параметров (3.7) после замуления нейронов в модели нейросети — учителя. Заметим, что аналогичным образом можно удалить подмножество нейронов в одном слое. Если число нейронов отличается в нескольких слоях модели нейросети учителя, то выполняется последовательно применения отображения $\phi(t, \mathbf{u})$ для каждого t -го слоя.

Приведем пояснение доказательства теоремы. Введем обозначение:

$$\hat{\mathbf{w}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} = \arg \min_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{w}} D_{KL}(q(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) || p(\mathbf{w}|\mathbf{A})) - \sum_{i=1}^m \log p(y_i|\mathbf{x}_i, \mathbf{w}).$$

Параметры \mathbf{u} модели \mathbf{f} делятся на **удаляемые $\boldsymbol{\nu}_2$, зануляемые $\boldsymbol{\nu}_1$** , оставшиеся \mathbf{v} . Суперпозиция слоев модели учителя \mathbf{f} в окрестности t -го слоя:

$$\mathbf{f}(\mathbf{x}) = \dots \circ \underbrace{\begin{pmatrix} u_{1,1} & \dots & u_{1,j} & \dots & u_{1,n_t} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{n_{t+1},1} & \dots & u_{n_{t+1},j} & \dots & u_{n_{t+1},n_t} \end{pmatrix}}_{\mathbf{U}_{t+1}} \boldsymbol{\sigma} \circ \underbrace{\begin{pmatrix} u_{1,1} & \dots & u_{1,n_{t-1}} \\ \vdots & \ddots & \vdots \\ u_{j,1} & \dots & u_{j,n_{t-1}} \\ \vdots & \ddots & \vdots \\ u_{n_t,1} & \dots & u_{n_t,n_{t-1}} \end{pmatrix}}_{\mathbf{U}_t} \boldsymbol{\sigma} \circ \dots \circ \mathbf{U}_1 \mathbf{x}$$

Апостериорное распределение параметров \boldsymbol{v} модели \mathbf{f} :

$$p(\boldsymbol{v}|\mathfrak{D}) = \int_{\nu_2 \in \mathbb{R}^{n_t-1}} p(\bar{\nu}_1|\mathfrak{D}, \nu_1 = \mathbf{0}) d\nu_2.$$

Из свойства распределения

$$p(\bar{\nu}_1|\mathfrak{D}, \nu_1 = \mathbf{0}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Xi}),$$

с параметрами $\boldsymbol{\mu}, \boldsymbol{\Xi}$:

$$\begin{aligned}\boldsymbol{\mu} &= \mathbf{m}_{\bar{\nu}_1} + \boldsymbol{\Sigma}_{\bar{\nu}_1, \nu_1} \boldsymbol{\Sigma}_{\nu_1, \nu_1}^{-1} (\mathbf{0} - \mathbf{m}_{\nu_1}), \\ \boldsymbol{\Xi} &= \boldsymbol{\Sigma}_{\bar{\nu}_1, \bar{\nu}_1} - \boldsymbol{\Sigma}_{\bar{\nu}_1, \nu_1} \boldsymbol{\Sigma}_{\nu_1, \nu_1}^{-1} \boldsymbol{\Sigma}_{\nu_1, \bar{\nu}_1}.\end{aligned}$$

Маргинализация нормального распределения

$$p(\boldsymbol{v}|\mathfrak{D}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{v}}, \boldsymbol{\Xi}_{\boldsymbol{v}, \boldsymbol{v}}),$$

что соответствует полученному в теореме 4.

Удаление слоя учителя. Проведем выравнивание модели учителя и модели ученика, согласно определению 8 при помощи последовательных преобразований вектора параметров \mathbf{u} . Рассмотрим преобразование:

$$\psi(t, \mathbf{u}) : \mathbb{R}^{N_{\text{tr}}} \rightarrow \mathbb{R}^{N_{\text{tr}} - n_t n_{t-1}}$$

вектора \mathbf{u} которое описывает удаление одного t -го слоя. Обозначим новый вектор параметров $\boldsymbol{v} = \psi(t, \mathbf{u})$, а элементы вектора, которые удалены как $\bar{\boldsymbol{v}}$.

Теорема 5. Пусть выполняются условия:

- 1) апостериорное распределение параметров $p(\mathbf{u}|\mathfrak{D}) = \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$,
- 2) соответствующие размеры слоев совпадают, $n_t = n_{t-1}$, т.е. матрица \mathbf{U}_t является квадратной,
- 3) функция активации удовлетворяет свойству идемпотентности $\boldsymbol{\sigma} \circ \boldsymbol{\sigma} = \boldsymbol{\sigma}$.

Тогда апостериорное распределение также описывается нормальным распределением с плотностью распределения

$$p(\boldsymbol{v}|\mathfrak{D}) = \mathcal{N}(\mathbf{m}_{\boldsymbol{v}} + \boldsymbol{\Sigma}_{\boldsymbol{v}, \bar{\boldsymbol{v}}} \boldsymbol{\Sigma}_{\bar{\boldsymbol{v}}, \bar{\boldsymbol{v}}}^{-1} (\mathbf{i} - \bar{\boldsymbol{v}}), \boldsymbol{\Sigma}_{\boldsymbol{v}, \boldsymbol{v}} - \boldsymbol{\Sigma}_{\boldsymbol{v}, \bar{\boldsymbol{v}}} \boldsymbol{\Sigma}_{\bar{\boldsymbol{v}}, \bar{\boldsymbol{v}}}^{-1} \boldsymbol{\Sigma}_{\bar{\boldsymbol{v}}, \boldsymbol{v}}), \quad (3.8)$$

где

$$\mathbf{i} = [\underbrace{1, 0, \dots, 0}_{n_t}, \underbrace{0, 1, \dots, 0}_{n_t}, \underbrace{0, 0, 1, \dots, 0}_{n_t}, \underbrace{0, 0, \dots, 1}_{n_t}]^T.$$

Доказательство. Рассмотрим две структуры нейронной сети с T слоями и $T+1$ слоем соответственно. Рассмотрим t -й слой, для которого выполняются условия этой теоремы. Заметим, что если t -й слой нейронной сети с $T+1$ слоем приравнять к единичной матрице, то он будет эквивалентным архитектуре с T слоями:

$$\begin{aligned} f &= \sigma \circ \mathbf{U}_{T+1} \sigma \circ \mathbf{U}_T \dots \sigma \circ \mathbf{U}_t \sigma \circ \dots \mathbf{U}_2 \sigma \circ \mathbf{U}_1 = \\ &= \sigma \circ \mathbf{U}_{T+1} \sigma \circ \mathbf{U}_T \dots \sigma \circ \mathbf{I} \sigma \circ \dots \mathbf{U}_2 \sigma \circ \mathbf{U}_1 = \\ &= \sigma \circ \mathbf{U}_{T+1} \sigma \circ \mathbf{U}_T \dots \sigma \circ \sigma \circ \dots \mathbf{U}_2 \sigma \circ \mathbf{U}_1. \end{aligned}$$

Используя свойство идемпотентности функции σ получаем:

$$f = \sigma \circ \mathbf{U}_{T+1} \sigma \circ \mathbf{U}_T \dots \sigma \circ \dots \mathbf{U}_2 \sigma \circ \mathbf{U}_1.$$

Получаем, что удаление t -го слоя нейросети эквивалентно приравниванию матрицы параметров t -го слоя к единичной матрице. Распределение параметров после приравнивания к единичной матрице вычисляется при помощи условного распределения. В силу общих свойств нормального распределения, условное распределение также является нормальным распределением с параметрами μ, Σ :

$$\begin{aligned} \mu &= \mathbf{m}_v + \Sigma_{v,\bar{v}} \Sigma_{\bar{v},\bar{v}}^{-1} (\mathbf{i} - \bar{v}), \\ \Sigma &= \Sigma_{v,v} - \Sigma_{v,\bar{v}} \Sigma_{\bar{v},\bar{v}}^{-1} \Sigma_{v,\bar{v}}, \end{aligned}$$

где вектор \mathbf{m}_v является подвектором вектора \mathbf{m} соответствующий параметрам v , а матрица $\Sigma_{v,\bar{v}}$ является подматрицей ковариационной матрицы Σ соответствующий векторам параметров v и \bar{v} . \square

Теорема 5 задает апостериорное распределение (3.8) параметров после удаления слоя нейросети. Полученное распределение $p(v|\mathcal{D})$ является оценкой апостериорного распределения модели без одного слоя.

Выравнивание рекурентных сетей. Рассматривается задача аппроксимации последовательности:

$$[\mathbf{x}_1, \dots, \mathbf{x}_q, \dots, \mathbf{x}_Q],$$

где \mathbf{x}_q — вектор признакового описания q -го элемента последовательности. Структура модели RNN задается размерностью скрытых слоев n_1, n_2, \dots, n_T . Скрытое представление \mathbf{h}_t^q элемента последовательности \mathbf{x}_q для t -го слоя задается выражением:

$$\mathbf{h}_t^q = \sigma(\mathbf{U}_t^1 \mathbf{h}_{t-1}^q + \mathbf{U}_t^2 \mathbf{h}_t^{q-1}),$$

где матрицы $\{\mathbf{U}_t^1, \mathbf{U}_t^2\}_{t=1}^T$ задают линейные отображения в модели RNN.

Отображение

$$\phi_{\text{RNN}}(t, \mathbf{u}) : \mathbb{R}^{p_{\text{tr}}} \rightarrow \mathbb{R}^{p_{\text{tr}} - 3n_t}$$

задает снижение размерности пространства после t -го слоя на единицу. Полученный вектор параметров после снижения размерности обозначается

$$\mathbf{v} = \phi(t, \mathbf{u}).$$

Исходный вектор \mathbf{u} состоит из подвекторов $\boldsymbol{\nu}_2, \boldsymbol{\nu}_1, \mathbf{v}$, которые описывают удаляемые, зануляемые и оставшиеся параметры соответственно.

Теорема 6. Пусть задано распределение вектора параметров $p(\mathbf{u})$. Тогда распределение вектора параметров $p(\mathbf{v})$ представимо в виде:

$$p(\mathbf{v}) = \int_{\boldsymbol{\nu}_2 \in \mathbb{R}^{n_t-1}} p(\bar{\boldsymbol{\nu}}_1 | \mathcal{D}, \boldsymbol{\nu}_1 = \mathbf{0}) d\boldsymbol{\nu}_2.$$

Доказательство. Пусть $\phi(t, \mathbf{u})$ удаляет j -ю компоненту вектора скрытого представления \mathbf{h}_t^q . Удаление j -й компоненты вектора \mathbf{h}_t^q эквивалентно занулению j -го столбца матрицы \mathbf{U}_{t+1}^1 . Тогда предсказание модели не зависит от параметров в j -й строке матрицы \mathbf{U}_t^1 и \mathbf{U}_t^2 . Следовательно этими параметрами можно пренебречь.

Найдем распределение вектора \mathbf{v} . Для поиска распределения вектора параметров $\bar{\boldsymbol{\nu}}_1$ после зануления j -го столбца матрицы \mathbf{U}_{t+1}^1 воспользуемся формулой условной вероятности $p(\bar{\boldsymbol{\nu}}_1 | \mathcal{D}, \boldsymbol{\nu}_1 = \mathbf{0})$, а для удаления j -й строки матриц \mathbf{U}_t^1 и \mathbf{U}_t^2 воспользуемся маргинализацией распределения $p(\bar{\boldsymbol{\nu}}_1 | \mathcal{D}, \boldsymbol{\nu}_1 = \mathbf{0})$. Обозначим зануляемые параметры модели как $\boldsymbol{\nu}_1$, а удаляемые параметры как $\boldsymbol{\nu}_2$. Также обозначим все параметры, которые не занулены как $\bar{\boldsymbol{\nu}}_1 = [\mathbf{v}^\top, \boldsymbol{\nu}_2^\top]$. Итоговое распределение параметров принимает вид:

$$p(\mathbf{v} | \mathcal{D}) = \int_{\boldsymbol{\nu}_2} p(\bar{\boldsymbol{\nu}}_1 | \mathcal{D}, \boldsymbol{\nu}_1 = \mathbf{0}) d\boldsymbol{\nu}_2.$$

□

Теорема 7. Пусть выполняются условия:

- 1) апостериорное распределение параметров $p(\mathbf{u} | \mathcal{D}) = \mathcal{N}(\mathbf{m}, \Sigma)$,
 - 2) число слоев модели учителя равняется числу слоев модели ученика $T = L$,
 - 3) размеры соответствующих пространств не совпадают $n_t \geq n_t$.
- Тогда распределение параметров $p(\mathbf{v} | \mathcal{D})$ также является нормальным.

Доказательство. Из свойств нормального распределения следует, что распределение

$$p(\bar{\boldsymbol{\nu}}_1 | \mathcal{D}, \boldsymbol{\nu}_1 = \mathbf{0})$$

также является нормальным распределением с параметрами $\boldsymbol{\mu}, \boldsymbol{\Xi}$:

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{m}_{\bar{\boldsymbol{\nu}}_1} + \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1, \boldsymbol{\nu}_1} \boldsymbol{\Sigma}_{\boldsymbol{\nu}_1, \boldsymbol{\nu}_1}^{-1} (\mathbf{0} - \mathbf{m}_{\boldsymbol{\nu}_1}), \\ \boldsymbol{\Xi} &= \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1, \bar{\boldsymbol{\nu}}_1} - \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1, \boldsymbol{\nu}_1} \boldsymbol{\Sigma}_{\boldsymbol{\nu}_1, \boldsymbol{\nu}_1}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\nu}_1, \bar{\boldsymbol{\nu}}_1}, \end{aligned}$$

где векторы $\mathbf{m}_{\bar{\boldsymbol{\nu}}_1}, \mathbf{m}_{\boldsymbol{\nu}_1}$ являются подвекторами вектора \mathbf{m} , который относится к параметрам $\bar{\boldsymbol{\nu}}_1$ и $\boldsymbol{\nu}_1$ соответственно. Ковариационная матрица $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1, \boldsymbol{\nu}_1}$ обозначает подматрицу матрицы $\boldsymbol{\Sigma}$, которая соответствует ковариационной матрице параметров $\bar{\boldsymbol{\nu}}_1$ и $\boldsymbol{\nu}_1$.

Распределение $p(\mathbf{v}|\mathfrak{D})$ найдем при помощи маргинализации распределения (3.6) по параметрам $\boldsymbol{\nu}_2$. Используя свойства нормального распределения получаем распределение

$$p(\mathbf{v}|\mathfrak{D}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\Xi}_{\mathbf{v},\mathbf{v}}),$$

где $\boldsymbol{\mu}_{\mathbf{v}}$ обозначает подвектор вектора $\boldsymbol{\mu}$, который относится к параметру \mathbf{v} , а матрица $\boldsymbol{\Xi}_{\mathbf{v},\mathbf{v}}$ является подматрицей матрицы $\boldsymbol{\Xi}$, которая относится к вектору параметров \mathbf{v} . \square

Преобразование t -го слоя учителя:

$$\psi_{\text{RNN}}(t) : \mathbb{R}^{p_{\text{tr}}} \rightarrow \mathbb{R}^{p_{\text{tr}} - n_t n_{t-1}}$$

описывает удаление t -го слоя. Новый вектор параметров $\mathbf{v} = \psi(t, \mathbf{u})$, а удаляемые элементы вектора — $\bar{\mathbf{v}}$.

Теорема 8. Пусть выполняются условия:

- 1) апостериорное распределение параметров $p(\mathbf{u}|\mathfrak{D}) = \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$,
- 2) размеры соответствующих пространств совпадают: $n_t = n_{t+1}$.
- 3) функция активации удовлетворяет свойству идемпотентности $\sigma \circ \sigma = \sigma$.

Тогда апостериорное распределение также описывается нормальным распределением с плотностью распределения:

$$p(\mathbf{v}|\mathfrak{D}) = \mathcal{N}(\mathbf{m}_{\mathbf{v}} + \boldsymbol{\Sigma}_{\mathbf{v},\bar{\mathbf{v}}} \boldsymbol{\Sigma}_{\bar{\mathbf{v}},\bar{\mathbf{v}}}^{-1} (\mathbf{i} - \bar{\mathbf{v}}), \boldsymbol{\Sigma}_{\mathbf{v},\mathbf{v}} - \boldsymbol{\Sigma}_{\mathbf{v},\bar{\mathbf{v}}} \boldsymbol{\Sigma}_{\bar{\mathbf{v}},\bar{\mathbf{v}}}^{-1} \boldsymbol{\Sigma}_{\mathbf{v},\bar{\mathbf{v}}}),$$

где

$$\mathbf{i} = [\underbrace{1, 0, \dots, 0}_{n_t}, \underbrace{0, 1, \dots, 0}_{n_t}, \underbrace{0, 0, 1, \dots, 0}_{n_t}, \underbrace{0, \dots, 1}_{n_t}, \underbrace{0, \dots, 0}_{n_t \times n_t}]^T.$$

Доказательство. Рассмотрим две структуры нейронной сети с T слоями и $T+1$ слоем соответственно. Рассмотрим t -й слой, для которого выполняется условие теоремы. Заметим, что если в t -м слое матрицу \mathbf{U}_t^1 приравнять к единичной, а матрицу \mathbf{U}_t^2 к нулевой, то полученная нейросеть с $T+1$ слоем будет эквивалентной нейросети с T слоями:

$$\begin{aligned} \mathbf{h}_{t+1}^q &= \sigma(\mathbf{U}_{t+1}^1 \sigma(\mathbf{U}_t^1 \sigma(\mathbf{U}_{t-1}^1 \mathbf{h}_{t-2}^q + \mathbf{U}_{t-1}^2 \mathbf{h}_{t-1}^{q-1}) + \mathbf{U}_t^2 \mathbf{h}_t^{q-1}) + \mathbf{U}_t^2 \mathbf{h}_{t+1}^{q-1}) = \\ &= \sigma(\mathbf{U}_{t+1}^1 \sigma(\sigma(\mathbf{U}_{t-1}^1 \mathbf{h}_{t-2}^q + \mathbf{U}_{t-1}^2 \mathbf{h}_{t-1}^{q-1})) + \mathbf{U}_t^2 \mathbf{h}_{t+1}^{q-1}), \end{aligned}$$

где используя свойство идемпотентности получаем:

$$\begin{aligned} \mathbf{h}_{t+1}^q &= \sigma(\mathbf{U}_{t+1}^1 \sigma(\mathbf{U}_{t-1}^1 \mathbf{h}_{t-2}^q + \mathbf{U}_{t-1}^2 \mathbf{h}_{t-1}^{q-1}) + \mathbf{U}_t^2 \mathbf{h}_{t+1}^{q-1}) = \\ &= \sigma(\mathbf{U}_{t+1}^1 \mathbf{h}_{t-1}^q + \mathbf{U}_t^2 \mathbf{h}_{t+1}^{q-1}). \end{aligned}$$

Получаем, что удаление t -го слоя нейросети эквивалентно приравниванию матриц параметров t -го слоя к единичной и нулевой матрице соответственно. Распределение параметров вычисляется при помощи условного распределения. В

силу общих свойств нормального распределения, условное распределения также является нормальным распределением с параметрами $\boldsymbol{\mu}, \boldsymbol{\Xi}$:

$$\begin{aligned}\boldsymbol{\mu} &= \mathbf{m}_{\mathbf{v}} + \boldsymbol{\Sigma}_{\mathbf{v}, \bar{\mathbf{v}}} \boldsymbol{\Sigma}_{\bar{\mathbf{v}}, \bar{\mathbf{v}}}^{-1} (\mathbf{i} - \bar{\mathbf{v}}), \\ \boldsymbol{\Xi} &= \boldsymbol{\Sigma}_{\mathbf{v}, \mathbf{v}} - \boldsymbol{\Sigma}_{\mathbf{v}, \bar{\mathbf{v}}} \boldsymbol{\Sigma}_{\bar{\mathbf{v}}, \bar{\mathbf{v}}}^{-1} \boldsymbol{\Sigma}_{\bar{\mathbf{v}}, \mathbf{v}},\end{aligned}$$

где вектор $\mathbf{m}_{\mathbf{v}}$ является подвектором вектора \mathbf{m} соответствующий параметрам \mathbf{v} , а матрица $\boldsymbol{\Sigma}_{\mathbf{v}, \bar{\mathbf{v}}}$ является подматрицей ковариационной матрицы $\boldsymbol{\Sigma}$ соответствующий векторам параметров \mathbf{v} и $\bar{\mathbf{v}}$. \square

Теорема 8 задает апостериорное распределение параметров после удаления одного слоя модели RNN. Полученное распределение $p(\mathbf{v}|\mathfrak{D})$ является оценкой апостериорного распределения модели без одного слоя.

3.3. Последовательность выравнивающих преобразований

Множество всех структур задается последовательностью натуральных чисел:

$$\mathfrak{H} = \{(n_1, n_2, \dots, n_T), \quad n_i \in \mathbb{N}, T \in \mathbb{N}\}.$$

Множество структур, порождаемое структурой $(n'_1, n'_2, \dots, n'_L)$ учителя \mathbf{f} :

$$\mathfrak{H}_f = \left\{ \mathbf{n} = (n_1, \dots, n_T) \mid n_i \in \mathbb{N}; \right. \\ \left. n_i \leq n'_i, \quad 3 \leq T \leq L, \quad n_T = n'_L; \quad n_{T-1} \leq n'_i, \quad i > T \right\},$$

описывает конечное подмножество структур в континуальном множестве.

Теорема 9. Для произвольной структуры $\mathbf{n} \in \mathfrak{H}_f$ существует последовательность локальных выравнивающих преобразований $\tau = (\dots, \boldsymbol{\phi}, \dots, \boldsymbol{\psi}, \dots)$, сохраняющее распределение параметров модели \mathbf{f} .

Доказательство. Преобразования $\boldsymbol{\phi}$ и $\boldsymbol{\psi}$ либо уменьшают n_i на единицу либо удаляют одну из компонент последовательности \mathbf{n} .

Рассмотрим структуру модели учителя \mathbf{f} вида $\mathbf{n}_f = (n_1, n_2, \dots, n_L)$, которая преобразовывается в структуру $\mathbf{n} = (n_1, n'_2, \dots, n'_{L-1}, n_L)$. Последовательность преобразований

$$\tau = \left(\underbrace{\boldsymbol{\phi}(2, \mathbf{u}), \dots, \boldsymbol{\phi}(2, \mathbf{u})}_{n_2 - n'_2}, \dots, \underbrace{\boldsymbol{\phi}(L-1, \mathbf{u}), \dots, \boldsymbol{\phi}(L-1, \mathbf{u})}_{n_{L-1} - n'_{L-1}} \right) \quad (3.9)$$

преобразует исходную структуру модели учителя \mathbf{n}_f в структуру \mathbf{n} .

Рассмотрим структуру модели учителя \mathbf{f} вида $\mathbf{n}_f = (n_1, n_2, \dots, n_L)$, которая преобразовывается в структуру

$$\mathbf{n} = (n_1, n'_2, \dots, n'_{T-1}, n'_T).$$

Выравнивания состоит из трех этапов.

1. Уменьшить размеры всех слоев n_T, \dots, n_{L-1} . Сделать матрицы всех преобразований квадратными с размером n_{T-1} . Получиться последовательность преобразований

$$\tau_1 = \left(\underbrace{\phi(T, \mathbf{u}), \dots, \phi(T, \mathbf{u})}_{n_T - n_{T-1}}, \dots, \underbrace{\phi(L-1, \mathbf{u}), \dots, \phi(L-1, \mathbf{u})}_{n_{L-1} - n_{T-1}} \right)$$

преобразующий исходную структуру модели учителя \mathbf{n}_f в структуру

$$\mathbf{n}_1 = (n_1, n_2, \dots, n_{T-2}, n_{T-1}, \dots, n_{T-1}, n_L).$$

2. При помощи последовательности преобразований

$$\tau_2 = (\psi(T, \mathbf{u}), \psi(T+1, \mathbf{u}), \dots, \psi(L-1, \mathbf{u}))$$

выполняется преобразования структуры s_1 в структуру

$$\mathbf{n}_2 = (n_1, n_2, \dots, n_{T-1}, n_L),$$

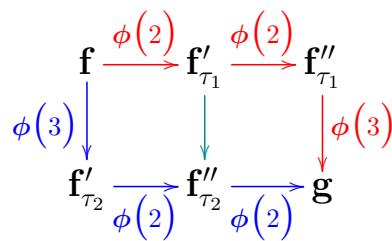
которая имеет число слоев T , что соответствует числу слоев в структуре \mathbf{n} .

3. Используя последовательность преобразований τ_3 аналогичную последовательности τ (3.9) для структур с одинаковым числом слоев получим итоговую структуру \mathbf{n} .

Суперпозиция преобразований $\tau_3 \circ \tau_2 \circ \tau_1$ задает преобразования из условия теоремы, что и требовалось доказать.

□

Рассмотрим теорему 9 на примере. Учитель \mathbf{f} имеет структуру $(10, 6, 7, 10)$, модель ученика \mathbf{g} имеет структуру $(10, 4, 6, 10)$.



Построенные последовательности выравнивающих преобразований модели учителя \mathbf{f} в модель ученика \mathbf{g} :

$$\tau_1 = (\phi(2), \phi(2), \phi(3)), \tau_2 = (\phi(3), \phi(2), \phi(2)), \tau_3 = (\phi(2), \phi(3), \phi(2)).$$

Цветом выделены разные последовательности преобразований. Из примера видно, что последовательность выравнивающих преобразований существует, но не единственна. Преобразования ϕ, ψ выравнивают пространства параметров учителя \mathbf{f} и ученика \mathbf{g} . После выравнивания параметрических моделей получаем, что параметры модели учителя и модели ученика принадлежат одному пространству параметров.

3.4. Анализ байесовской дистилляции полносвязных нейронных сетей

Проводится вычислительный эксперимент для анализа предложенного метода дистилляции на основе апостериорного распределения параметров модели учителя.

Проанализируем модель на синтетической выборке. Выборка построенная следующим образом:

$$\mathbf{w} = [w_j : w_j \sim \mathcal{N}(0, 1)]_{n \times 1}, \quad \mathbf{X} = [x_{ij} : x_{ij} \sim \mathcal{N}(0, 1)]_{m \times n}, \\ \mathbf{y} = [y_i : y_i \sim \mathcal{N}(\mathbf{x}_i^T \mathbf{w}, \beta)]_{m \times 1},$$

где $\beta = 0,1$ — уровень шума в данных. В эксперименте число признаков $n = 10$. Для обучения и тестирования сгенерировано $m_{\text{train}} = 900$ и $m_{\text{test}} = 124$ объекта.

В качестве модели учителя рассматривался многослойный перцептрон с двумя скрытыми слоями (3.3). Матрицы линейных преобразований имеют размер:

$$\mathbf{U}_1 \in \mathbb{R}^{100 \times 10}, \quad \mathbf{U}_2 \in \mathbb{R}^{50 \times 100}, \quad \mathbf{U}_3 \in \mathbb{R}^{1 \times 50}.$$

В качестве функции активации выбрана ReLu. Модель учителя предварительно обучена на основе вариационного вывода (3.5), где в качестве априорного распределения параметров выбрано стандартное нормальное распределение.

В качестве модели ученика выбраны две конфигурации. Первая конфигурация получается путем удаления нейронов в модели учителя:

$$g = \sigma \circ \mathbf{W}_3 \sigma \circ \mathbf{W}_2 \sigma \circ \mathbf{W}_1, \tag{3.10}$$

где σ является нелинейной функцией активации, а матрицы линейных преобразований имеют размер:

$$\mathbf{W}_1 \in \mathbb{R}^{10 \times 10}, \quad \mathbf{W}_2 \in \mathbb{R}^{10 \times 10}, \quad \mathbf{W}_3 \in \mathbb{R}^{1 \times 10}.$$

В качестве функции активации выбрана ReLu.

Рис. 3.1 сравнивает модели ученика, со структурой (3.10). Представлено сравнение разных моделей: модель без дистилляции, где в качестве априорного распределения выбирается стандартное нормальное распределение (на легенде обозначается student); модель с частичной дистилляцией, где в качестве среднего значения параметров выбираются параметры согласно выражения (3.7), а ковариационная матрица приравнивается к единичной матрице (на легенде обозначается distil-student); модель с полной дистилляцией согласно выражения (3.7) (на легенде обозначается distil-student-all). Видно, что обе модели ученика, где в качестве априорного распределения выбраны распределения, основанные на апостериорном распределении учителя, имеют большее правдоподобие, чем модель где в качестве априорного распределения выбрано стандартное нормальное $\mathcal{N}(0, 1)$. Также заметим, что использование параметра среднего из

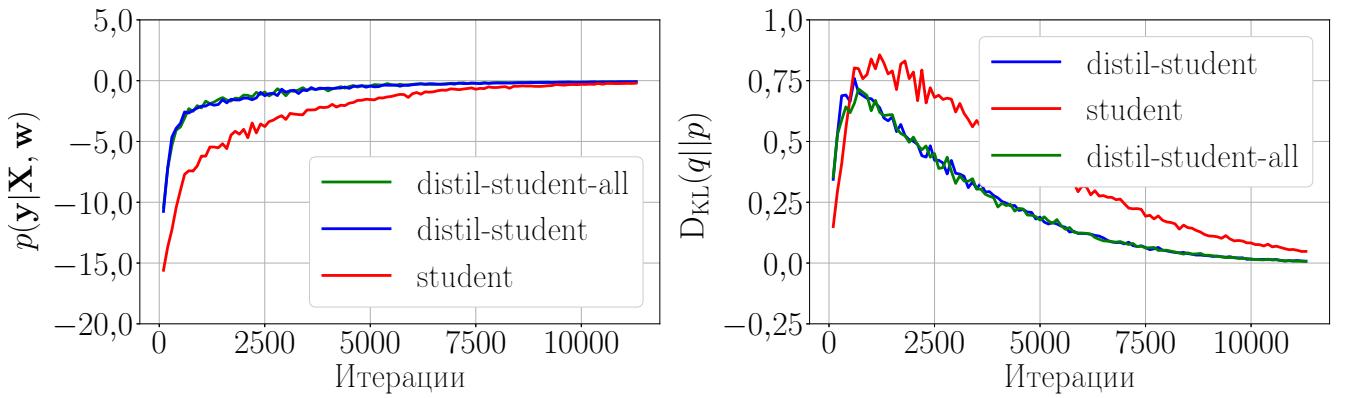


Рис. 3.1: Структура (3.10) модели ученика g . Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: KL–дивергенция между вариационным и априорным распределениями параметров модели

апостериорного распределения дает основной вклад при дистилляции, так как качество моделей distil-student и distil-student-all совпадает.

Вторая конфигурация получается путем удаления слоя модели учителя:

$$g = \sigma \circ \mathbf{W}_2 \sigma \circ \mathbf{W}_1, \quad (3.11)$$

где σ является нелинейной функцией активации, а матрицы линейных преобразований имеют размер

$$\mathbf{W}_1 \in \mathbb{R}^{1 \times 50}, \quad \mathbf{W}_2 \in \mathbb{R}^{50 \times 10}.$$

В качестве функции активации выбрана ReLu.

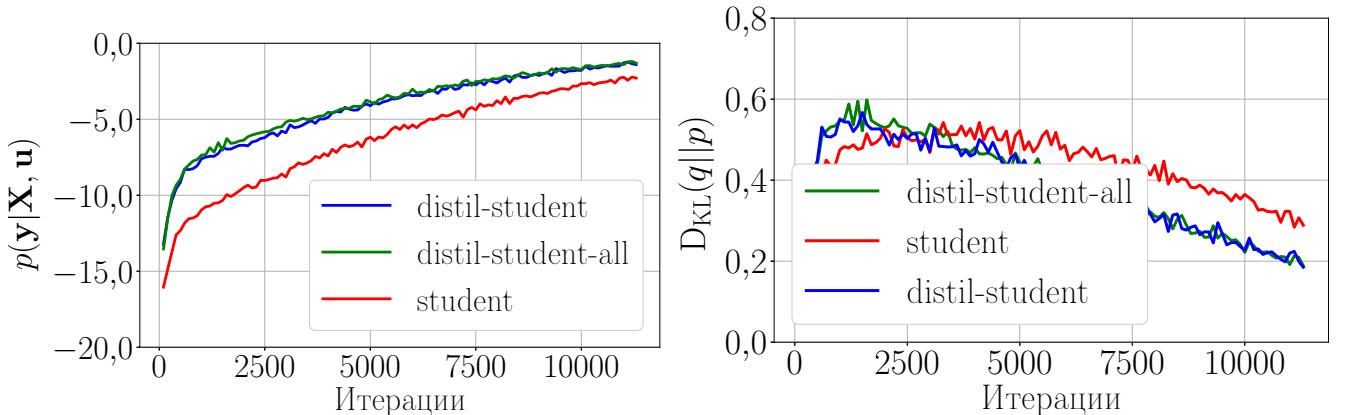


Рис. 3.2: Структура (3.11) модели ученика g . Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: KL–дивергенция между вариационным и априорным распределениями параметров модели

Рис. 3.2 сравнивает модели ученика со структурой (3.11). Аналогично рис. 3.1, на рис. 3.2 представлено сравнение модели без дистилляции (student), модели с дистилляцией параметра среднего значение (distil-student) и модели с

полной дистилляцией (distil-student-all). В рамках данного эксперимента, по дистилляции модели учителя в модель ученика с меньшим числом параметров получены результаты, которые подтверждают, что задание априорного распределения параметров ученика позволяет снизить число итераций при выборе оптимальных параметров модели ученика.

В этой части эксперимента проводился анализ байесовского подхода к дистилляции на реальных данных. В качестве реальных данных выбрана выборка FashionMnist [60] описывающая задачу классификации изображений на 10 классов.

В качестве модели учителя рассматривался многослойный перцептрон с двумя скрытыми слоями (3.3). Матрицы линейных преобразований имеют размер:

$$\mathbf{U}_1 \in \mathbb{R}^{800 \times 784}, \quad \mathbf{U}_2 \in \mathbb{R}^{50 \times 800}, \quad \mathbf{U}_3 \in \mathbb{R}^{10 \times 50},$$

В качестве функции активации выбрана ReLu. Модель учителя предварительно обучена на основе вариационного вывода (3.5), где в качестве априорного распределения параметров выбрано стандартное нормальное распределение.

В качестве модели ученика выбрана конфигурация с одним скрытым слоем (3.11), где матрицы линейных преобразований имеют размер:

$$\mathbf{W}_1 \in \mathbb{R}^{50 \times 784}, \quad \mathbf{W}_2 \in \mathbb{R}^{50 \times 10}.$$

В качестве функции активации выбрана ReLu.

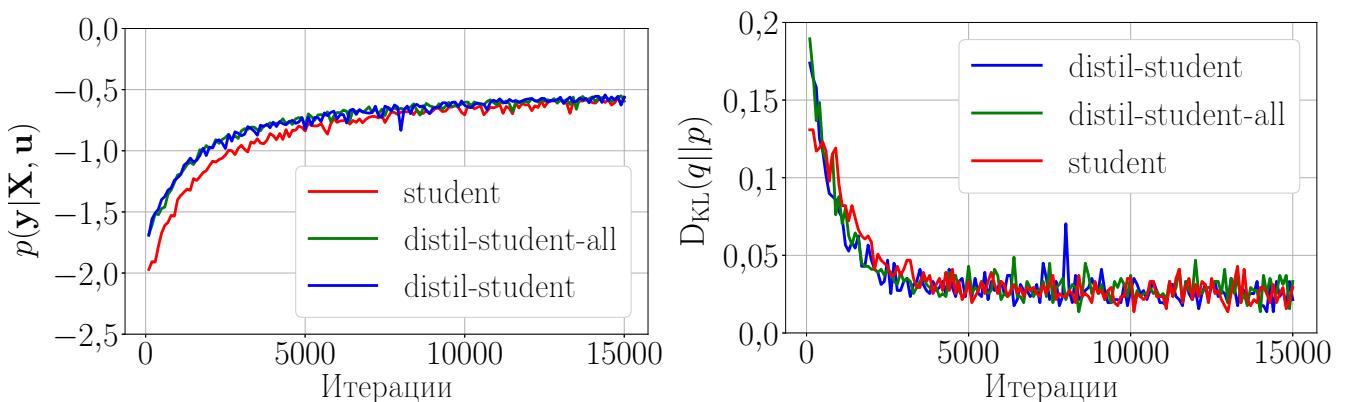


Рис. 3.3: Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: KL-дивергенция между вариационным и априорным распределениями параметров модели

Рис. 3.3 сравнивает модели ученика с разными априорными распределениями параметров. Аналогично синтетическому эксперименту, модель, где в качестве априорного распределения использовалось стандартное нормальное распределение, сравнивалась с моделью, где параметры распределения определялись на основе формулы (3.8). Видно, что у моделей с заданием априорного распределения на основе апостериорного распределения параметров учителя

Таблица 3.1: Сводная таблица результатов анализа байесовской дистилляции

	teacher	student	distil-student	distil-student-all
Эксперимент на синтетической выборке (удаление нейрона)				
Структура	[10, 100, 50, 1]	[10, 10, 10, 1]	[10, 10, 10, 1]	[10, 10, 10, 1]
Число параметров	6050	210	210	210
Разность площадей	-	0	16559	16864
Эксперимент на синтетической выборке (удаление слоя)				
Структура	[10, 100, 50, 1]	[10, 50, 1]	[10, 50, 1]	[10, 50, 1]
Число параметров	6050	550	550	550
Разность площадей S	-	0	23310	25506
Эксперимент на выборке FashionMnist				
Структура	[784, 800, 50, 10]	[784, 50, 10]	[784, 50, 10]	[784, 50, 10]
Число параметров	667700	39700	39700	39700
Разность площадей S	-	0	1165	1145

правдоподобие выборки выше, чем у модели, где в качестве априорного распределения выбрано стандартное нормальное распределение.

В табл. 3.1 представлен результат вычислительного эксперимента. Для численного сравнения качества моделей выбрана разность площадей графика правдоподобия $p(\mathbf{y}|\mathbf{X}, \mathbf{u})$ между моделью student и моделями distil-student и distil-student-all соответственно:

$$S = \sum_s p(\mathbf{y}|\mathbf{X}, \mathbf{u}_s^s) - p(\mathbf{y}|\mathbf{X}, \mathbf{u}_{ds}^s),$$

где $\mathbf{u}_s^s, \mathbf{u}_{ds}^s$ обозначает параметры модели студента и модели дистиллированного студента после s -й итерации оптимизационного процесса. Заметим, что площадь S имеет знак: чем больше значение положительного числа, тем дистиллированная модель точнее, чем модель построенная без учителя. Если площадь S принимает отрицательное значение, то это значит, что модель без дистилляции является точнее чем модель с дистилляцией. Из вычислительного эксперимента видно, что площадь S под графиками принимает положительные значения, то есть обе модели ученика полученные при помощи дистилляции, являются точнее чем модель ученика без дистилляции.

Глава 4

Априорные распределения параметров смеси экспертов

Данный раздел посвящен анализу свойств смеси экспертов. Рассматриваются различные способы выбора априорного распределения. Анализируется случай, когда выбраны как информативное так и неинформативное априорные распределения параметров каждого эксперта. Экспертами назначаются линейные модели. Смесь экспертов это комбинация экспертов при помощи шлюзовой функции. Рассматривается задача поиска окружностей на изображении. Каждой окружности на изображении соответствует свой эксперт. Рассматривается два случая с зависимыми и независимыми априорными распределениями параметрами локальных моделей — экспертов. Требуется найти на изображении синтетически генерированные окружности с разным уровнем шума. Сравнивается устойчивость к шуму смеси с заданными априорными распределениями на вектора параметров экспертов и без задания априорного распределения.

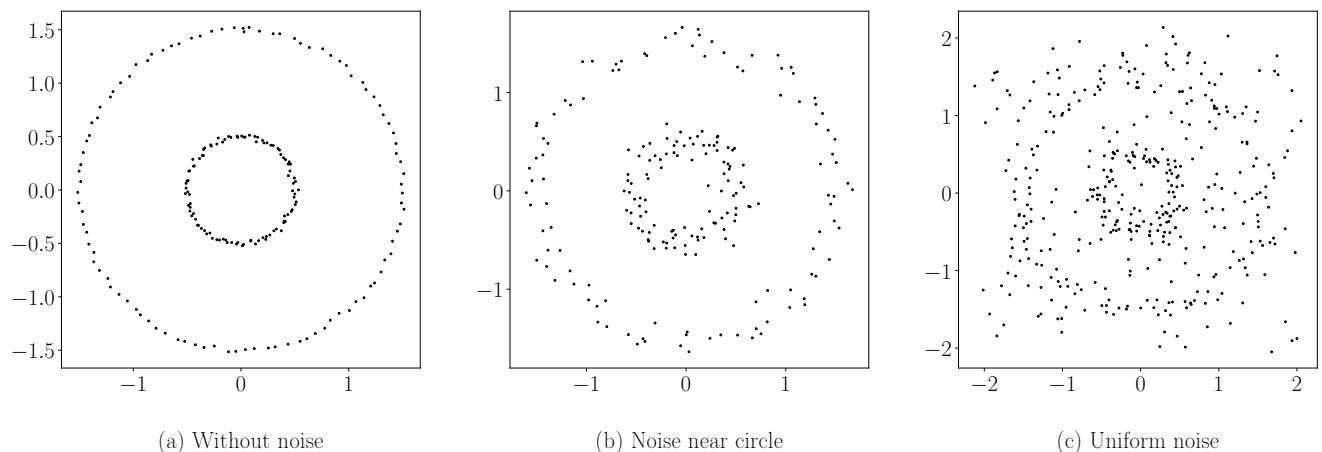


Рис. 4.1: Пример окружностей с разным уровнем шума: (а) окружности без шума; (б) окружности с зашумленным радиусом; (с) окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению

В рамках данной главы решается задача поиска окружностей на бинаризованном изображении. Предполагается, что радиусы окружностей различаются значимо, а также, что центры почти совпадают. Пример изображений показан на рис. 4.1. В качестве экспертов рассматриваются линейные модели — каждая модель аппроксимирует одну окружность. В качестве шлюзовой функции рассматривается двухслойная нейронная сеть.

Предлагается метод *обучения с экспертом*, который предполагает использование предметных знаний экспертов для улучшения качества аппроксимации, а также для получения интерпретируемых моделей машинного обучения. Предметные знания экспертов об образце будут называться *экспертная информация*. Предполагается, что использование экспертной информации позволяет аппроксимировать выборку простыми интерпретируемыми моделями, такими

как линейные модели. Методы машинного обучения, которые учитывают экспертные знания при построении моделей, называются *экспертным обучением*.

Решается задача аппроксимации кривых второго порядка на контурном изображении. Кривые второго порядка выбираются для анализа, так как они легко описываются линейными моделями. В этом случае эти фигуры необходимо восстанавливать в таких прикладных задачах, как задача распознавания радужной оболочки глаза [58, 59, 57], задача описания трека частицы в адронный коллайдер [64]. Экспертная информация о кривой второго порядка позволяет отображать точки на плоскости в новое описание объекта, где каждая кривая аппроксимируется одной линейной моделью. Модель аппроксимирующая одну кривую, называется *локальной моделью*. Для аппроксимации всего контурного изображения необходимо аппроксимировать несколько кривых второго порядка с помощью нескольких локальных моделей. Вводятся ограничения на изображения: а) изображение состоит только из кривых второго порядка; б) изображение аппроксимируется небольшим количеством кривых второго порядка; в) количество и тип кривых на изображении известны.

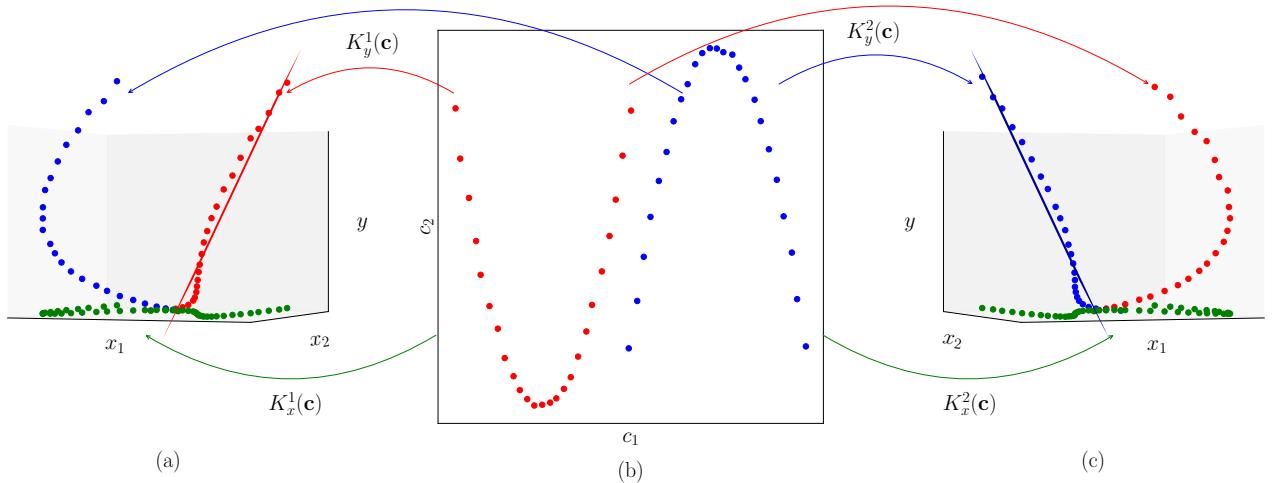


Рис. 4.2: Пример: а) экспертная информация первого эксперта; б) исходные данные; в) экспертная информация второго эксперта

На рисунке 4.2 показан пример кривых второго порядка, а также экспертная информация о кривых. На рисунке 4.2.a показана экспертная информация первого эксперта. Используя эту информацию, первая кривая аппроксимируется линейной моделью, а вторая кривая — шумом. На рисунке 4.2.b показана экспертная информация второго эксперта. Используя эту информацию, вторая кривая аппроксимируется линейной моделью, а первая кривая представляет собой шум.

Основной проблемой построения мультимоделей является то, что мультимодель зависит от начальной инициализации параметров локальных моделей. Для повышения устойчивости мультимодели предлагается использовать вероятностную постановку задачи для поиска оптимальных параметров шлюзовой

функции и параметров локальной модели. Предлагается задать априорное распределение на параметры локальных моделей, также, для повышения предлагаются учесть зависимость априорных распределений для разных моделей.

При аппроксимации нескольких кривых на одном контурном изображении строится мульти модель. Примером нескольких моделей является случайный лес [40], бустинг деревьев [2], смесь экспертов [41]. В данной работе смесь экспертов рассматривается как мульти модель. Смесь экспертов — это мульти модель, которая линейно взвешивает локальные модели аппроксимирующие часть выборки. Значения весовых коэффициентов зависят от объекта, для которого делается прогноз. Для решения проблемы смеси экспертов используется вариационный ЕМ-алгоритм [65, 33, 54].

В качестве примера рассматривается задача аппроксимации изображения радужной оболочки глаза. На рисунке 4.3а показан пример изображения для аппроксимации. Рассматриваем обработанное изображение, которое дано в виде схемы, пример такого изображения показан на рисунке 4.3б. На рисунке 4.3б показаны две модели окружностей, которые аппроксимируют радужную оболочку глаза. Окружности — простой пример кривой второго порядка.

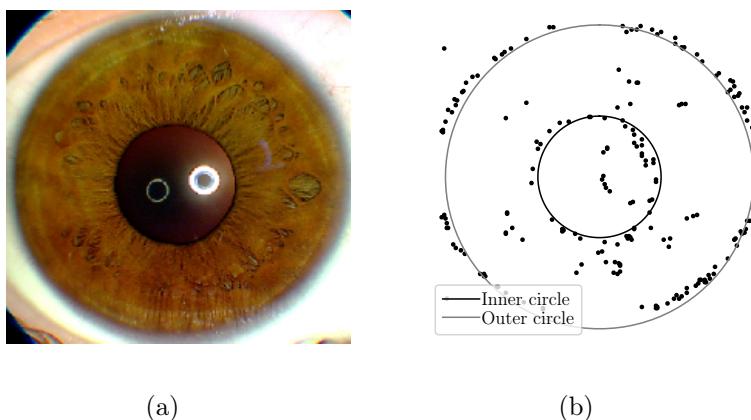


Рис. 4.3: Пример изображения радужной оболочки глаза и ее контурное изображение: а) изображение радужной оболочки глаза; б) контурное изображение радужной оболочки и аппроксимация заданного изображения окружностями

Для задачи аппроксимации радужной оболочки глаза используется экспертная информация: радужная оболочка глаза аппроксимируется двумя концентрическими окружностями. Экспертная информация используется для построения описания характеристик точек на плоскости, а также для построения функции оптимизации. Часть функции ошибок для оптимизации, использующая экспертную информацию, называется регуляризатором. Таким образом, информация о том, что изображение окружностей задается описанием признака, и информация о том, что концентрические окружности задаются с помощью специального регуляризатора.

В вычислительном эксперименте качество аппроксимации контурного изоб-

ражения анализируется в зависимости от заданной экспертной информации и от уровня шума в синтетически сгенерированных данных. Анализ качества аппроксимации диафрагмы проводится в зависимости от количества экспертной информации, которая использовалась при построении модели. Важно, что каждое изображение представляет собой отдельный набор точек, которые необходимо аппроксимировать.

4.1. Локальные модели в задаче построения смеси экспертов

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. 6.3а. 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:

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

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

Раскрыв скобки получаем

$$(2x_0) \cdot x_i + (2y_0) \cdot y_i + (r^2 - x_0^2 - y_0^2) \cdot 1 = x_i^2 + y_i^2 + \varepsilon_i. \quad (4.1)$$

Выражение (4.1) является задачей линейной регрессии с параметрами:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|, \quad \mathbf{X} = [\mathbf{C}, \mathbf{1}], \quad \mathbf{y} = [x_1^2 + y_1^2, x_2^2 + y_2^2, \dots, x_N^2 + y_N^2]^T. \quad (4.2)$$

Используя вектор параметров $\hat{\mathbf{w}} = [w_1, w_2, w_3]^T$ 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 (4.2) 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.

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 $\mathbf{C}'_k \subset \mathbf{C}$ соответствует 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 9. Модель \mathbf{g} называется локальной моделью для выборки \mathbf{U} , если \mathbf{g} аппроксимирует некоторое не пустое подмножество $\mathbf{U}' \subset \mathbf{U}$.

Определение 10. Мультимодель \mathbf{f} называется смесью экспертов, если

$$\mathbf{f} = \sum_{k=1}^K \pi_k \mathbf{g}_k(\mathbf{w}_k), \quad \pi_k(\mathbf{x}, \mathbf{V}) : \mathbb{R}^{n \times |\mathbf{V}|} \rightarrow [0, 1], \quad \sum_{k=1}^K \pi_k(\mathbf{x}, \mathbf{V}) = 1, \quad (4.3)$$

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

$$\mathbf{g}_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x}, \quad \boldsymbol{\pi}(\mathbf{x}, \mathbf{V}) = \text{softmax}(\mathbf{V}_1^\top \boldsymbol{\sigma}(\mathbf{V}_2^\top \mathbf{x})), \quad (4.4)$$

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

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

$$p(\mathbf{y}|\mathbf{X}, \mathbf{f}) = \prod_{i=1}^N \left(\sum_{k=1}^K \pi_k p_k(y_i | \mathbf{g}_k(\mathbf{x}_i)) \right), \quad (4.5)$$

где \mathbf{f} — это смесь экспертов, а $\mathbf{g}_k, \boldsymbol{\pi}$ определяются выражением (4.4).

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

$$p(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}) = \prod_{k=1}^K p^k(\mathbf{w}_k) \prod_{i=1}^N \left(\sum_{k=1}^K \pi_k p_k(y_i | \mathbf{w}_k, \mathbf{x}_i) \right), \quad (4.6)$$

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

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

4.2. Вероятностное обоснование смеси экспертов

Для построения смеси экспертов (4.3, 4.6), введем вероятностные предположения о данных (4.2):

- 1) правдоподобие $p_k(y_i|\mathbf{w}_k, \mathbf{x}_i) = \mathcal{N}(y_i|\mathbf{w}_k^\top \mathbf{x}_i, \beta^{-1})$, где параметр β является уровнем шума,
- 2) априорное распределение параметров $p^k(\mathbf{w}_k) = \mathcal{N}(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k)$, где \mathbf{w}_k^0 — вектор размерности $n \times 1$, а \mathbf{A}_k — ковариационная матрица размерности $n \times n$,
- 3) регуляризация априорного распределения $p(\boldsymbol{\varepsilon}_{k,k'}|\boldsymbol{\Xi}) = \mathcal{N}(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi})$, где $\boldsymbol{\Xi}$ — ковариационная матрица, а $\boldsymbol{\varepsilon}_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0$.

Предположение 2) задает априорное распределение вектора параметров локальных модели \mathbf{w}_k . Оно 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.

Предположение 3) задает регуляризацию априорных распределений. Данная регуляризация учитывает связь между априорными ограничениями разных локальных моделей. Например, если $\text{diag}(\boldsymbol{\Xi}) = [0.001, 0.001, 1]$, то центры разных окружностей совпадают.

Используя предположения 1), 2), 3) и выражение (4.6) получаем полное правдоподобие:

$$p(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) = \prod_{i=1}^N \left(\sum_{k=1}^K \pi_k \mathcal{N}(y_i|\mathbf{w}_k^\top \mathbf{x}_i, \beta^{-1}) \right) \cdot \prod_{k=1}^K \mathcal{N}(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k) \cdot \prod_{k,k'=1}^K \mathcal{N}(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}), \quad (4.7)$$

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

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 (4.7), а также взяв логарифм получаем:

$$\begin{aligned} \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) &= \\ &= \sum_{i=1}^N \sum_{k=1}^K z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (y_i - \mathbf{w}_k^\top \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\ &+ \sum_{k=1}^K \left[-\frac{1}{2} (\mathbf{w}_k - \mathbf{w}_k^0)^\top \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] + \\ &+ \sum_{k=1}^K \sum_{k'=1}^K \left[-\frac{1}{2} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0)^\top \boldsymbol{\Xi}^{-1} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0) + \frac{1}{2} \log \det \boldsymbol{\Xi} - \frac{n}{2} \log 2\pi \right]. \end{aligned} \quad (4.8)$$

Получаем новую задачу оптимизации обоснованности. Функция обоснованности получается при интегрировании выражения (4.8) по параметрам \mathbf{W}, \mathbf{Z} :

$$\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta = \arg \max_{\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta} \int_{\mathbf{W}, \mathbf{Z}} \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) d\mathbf{W} d\mathbf{Z}. \quad (4.9)$$

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

$$\begin{aligned} \log p(\mathbf{y} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) &= \int_{\mathbf{W}, \mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \log p(\mathbf{y} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) d\mathbf{W} d\mathbf{Z} = \\ &= \int_{\mathbf{W}, \mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \log \frac{p(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta)}{p(\mathbf{W}, \mathbf{Z} | \mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta)} d\mathbf{W} d\mathbf{Z} = \\ &= \int_{\mathbf{W}, \mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \log \frac{p(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) q(\mathbf{W}, \mathbf{Z})}{p(\mathbf{W}, \mathbf{Z} | \mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) q(\mathbf{W}, \mathbf{Z})} d\mathbf{W} d\mathbf{Z} = \\ &= \int_{\mathbf{W}, \mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \frac{p(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta)}{q(\mathbf{W}, \mathbf{Z})} d\mathbf{W} d\mathbf{Z} + \\ &\quad + \int_{\mathbf{W}, \mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \frac{q(\mathbf{W}, \mathbf{Z})}{p(\mathbf{W}, \mathbf{Z} | \mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta)} d\mathbf{W} d\mathbf{Z} = \\ &= \mathcal{L}(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta) + D_{KL}(q(\mathbf{W}, \mathbf{Z}) || p(\mathbf{W}, \mathbf{Z} | \mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta)) \end{aligned} \quad (4.10)$$

Используя (4.10) получаем нижнюю оценку обоснованности:

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

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

Используем ЕМ-алгоритм [65, 33] для решения оптимизационной задачи (4.9). Заметим, что ЕМ-алгоритм вместо оптимизации $\log p(\mathbf{y} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta)$ оптимизирует нижнюю оценку $\mathcal{L}(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)$.

Е-шаг. Е-шаг решает оптимизационную задачу

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

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

Пусть совместное распределение $q(\mathbf{Z}, \mathbf{W})$ удовлетворяет условию независимости $q(\mathbf{Z}, \mathbf{W}) = q(\mathbf{Z})q(\mathbf{W})$ [33]. Далее символом \propto обозначим то, что обе стороны выражения равны с точностью до аддитивной константы. Сначала

найдем распределение $q(\mathbf{Z})$:

$$\begin{aligned} \log q(\mathbf{Z}) &= \mathbb{E}_{q/\mathbf{Z}} \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) \propto \\ &\propto \sum_{i=1}^N \sum_{k=1}^K z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (y_i^2 - \mathbf{x}_i^\top \boldsymbol{\Xi} \mathbf{w}_k + \mathbf{x}_i^\top \boldsymbol{\Xi} \mathbf{w}_k \mathbf{w}_k^\top \mathbf{x}_i) + \frac{1}{2} \log \frac{\beta}{2\pi} \right] \\ p(z_{ik} = 1) &= \frac{\exp \left(\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (\mathbf{x}_i^\top \boldsymbol{\Xi} \mathbf{w}_k \mathbf{w}_k^\top \mathbf{x}_i - \mathbf{x}_i^\top \boldsymbol{\Xi} \mathbf{w}_k) \right)}{\sum_{k'=1}^K \exp \left(\log \pi_{k'}(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (\mathbf{x}_i^\top \boldsymbol{\Xi} \mathbf{w}_{k'} \mathbf{w}_{k'}^\top \mathbf{x}_i - \mathbf{x}_i^\top \boldsymbol{\Xi} \mathbf{w}_{k'}) \right)}. \end{aligned} \quad (4.11)$$

Используя выражения (4.11) получаем, что распределение $q(z_{ik})$ является бернулевским распределением с параметром z_{ik} , которое задается выражением (4.11). Далее найдем распределение $q(\mathbf{W})$:

$$\begin{aligned} \log q(\mathbf{W}) &= \mathbb{E}_{q/\mathbf{W}} \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) \propto \\ &\propto \sum_{i=1}^N \sum_{k=1}^K \mathbb{E} z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (y_i - \mathbf{w}_k^\top \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\ &+ \sum_{k=1}^K \left[-\frac{1}{2} (\mathbf{w}_k - \mathbf{w}_k^0)^\top \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] \\ &\propto \sum_{k=1}^K \left[\mathbf{w}_k^\top \left(\mathbf{A}_k^{-1} \mathbf{w}_k^0 + \beta \sum_{i=1}^N \mathbf{x}_i y_i \mathbb{E} z_{ik} \right) - \frac{1}{2} \mathbf{w}_k^\top \left(\mathbf{A}_k^{-1} + \beta \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{w}_k \right]. \end{aligned} \quad (4.12)$$

Используя выражение (4.12) получаем, что распределение $q(\mathbf{w}_k)$ является нормальным распределением со средним \mathbf{m}_k и ковариационной матрицей \mathbf{B}_k :

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

M-шаг. M-шаг решает оптимизационную задачу

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

где $q(\mathbf{W}, \mathbf{Z})$ является известной плотностью распределения. Распределение $q(\mathbf{Z}, \mathbf{W})$ является фиксированным, в то время как вариационная нижняя

оценка $\mathcal{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)$ максимизируется по параметрам $\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta$:

$$\begin{aligned}\mathcal{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta) &= \mathbb{E}_q \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) = \\ &= \sum_{i=1}^N \sum_{k=1}^K \mathbb{E} z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} \mathbb{E} (y_i - \mathbf{w}_k^\top \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\ &+ \sum_{k=1}^K \left[-\frac{1}{2} \mathbb{E} (\mathbf{w}_k - \mathbf{w}_k^0)^\top \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] + \quad (4.13) \\ &+ \sum_{k=1}^K \sum_{k'=1}^K \left[-\frac{1}{2} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0)^\top \boldsymbol{\Xi}^{-1} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0) + \frac{1}{2} \log \det \boldsymbol{\Xi} - \frac{n}{2} \log 2\pi \right].\end{aligned}$$

Во-первых, для нахождения оптимального параметра \mathbf{V} используется градиентный метод оптимизации, который сходится к некоторому локальному экстремуму. Во вторых, используя выражения (4.13) получаем оптимальное значения параметра \mathbf{A}_k

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

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

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \beta} &= \sum_{k=1}^K \sum_{i=1}^N \left(\frac{1}{\beta} \mathbb{E} z_{ik} - \frac{1}{2} \mathbb{E} z_{ik} [y_i^2 - 2y_i \mathbf{x}_i^\top \mathbb{E} \mathbf{w}_k + \mathbf{x}_i^\top \mathbf{w}_k \mathbf{w}_k^\top \mathbf{x}_i] \right) = 0, \\ \frac{1}{\beta} &= \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K [y_i^2 - 2y_i \mathbf{x}_i^\top \mathbb{E} \mathbf{w}_k + \mathbf{x}_i^\top \mathbb{E} \mathbf{w}_k \mathbf{w}_k^\top \mathbf{x}_i] \mathbb{E} z_{ik}. \\ \frac{\partial \mathcal{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \mathbf{w}_k^0} &= \mathbf{A}_k^{-1} (\mathbb{E} \mathbf{w}_k - \mathbf{w}_k^0) + \boldsymbol{\Xi} \sum_{k'=1}^K [\mathbf{w}_{k'}^0 - \mathbf{w}_k^0] = 0, \quad (4.14) \\ \mathbf{w}_k^0 &= [\mathbf{A}_k^{-1} + (K-1) \boldsymbol{\Xi}]^{-1} \left(\mathbf{A}_k^{-1} \mathbb{E} \mathbf{w}_k + \boldsymbol{\Xi} \sum_{k'=1, k' \neq k}^K \mathbf{w}_{k'}^0 \right).\end{aligned}$$

Выражения (4.11–4.14) задают итеративную процедуру, которая сходится к некоторому локальному максимуму оптимизационной задачи (4.9).

4.3. Априорное распределение для аппроксимации кривых второго порядка на изображении

Эксперт предполагает, что изображение состоит из кривой второго порядка Ω . Пусть для набора точек $\mathbf{C} \in \mathbb{R}^{N \times 2}$, образующих кривую Ω , дана экспертная

информация о фигуре $E(\Omega)$. Множество $E(\Omega)$ состоит из формы Ω , ожидаемой экспертом, и множества ее допустимых преобразований.

На основе экспертного описания введем отображения в новую задачу аппроксимации:

$$K_x(E(\Omega)) : \mathbb{R}^2 \rightarrow \mathbb{R}^n, \quad K_y(E(\Omega)) : \mathbb{R}^2 \rightarrow \mathbb{R}, \quad (4.15)$$

где K_x отображение объектов с признаковым описанием объектов, n число признаков, а K_y отображение в пространство целевых переменных для объекта. Применение отображений K_x, K_y для всех элементов выборки \mathbf{C} :

$$K_x(E(\Omega), \mathbf{c}) = \mathbf{x}, \quad K_y(E(\Omega), \mathbf{c}) = y, \quad (4.16)$$

где $\mathbf{c} = (x_i, y_i)$ точки из выборки \mathbf{C} .

Применив отображения (4.16) к точкам \mathbf{C} получаем выборку

$$\mathfrak{D} = \{(\mathbf{x}, y) \mid \forall \mathbf{c} \in \mathbf{C} \mathbf{x} = K_x(\mathbf{c}), y = K_y(\mathbf{c})\}. \quad (4.17)$$

Получаем, что исходная задача аппроксимации кривой Ω сводится к аппроксимации выборки \mathfrak{D} . Предполагается, что выборка \mathfrak{D} аппроксимируется линейной моделью:

$$g(\mathbf{x}, \mathbf{w}) = \mathbf{x}^\top \mathbf{w},$$

где \mathbf{w} вектор параметров, который необходимо найти.

Для поиска оптимального вектора параметров $\hat{\mathbf{w}}$, требуется решить оптимизационную задачу:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^n} \sum_{(\mathbf{x}, y) \in \mathfrak{D}} \|g(\mathbf{x}, \mathbf{w}) - y\|_2^2.$$

Таким образом, задача аппроксимации исходной кривой Ω сводится к решению задачи линейной регрессии, т.е. нахождению компонентов вектора $\hat{\mathbf{w}}$.

В случае, когда на изображении K кривые второго порядка $\Omega_1, \dots, \Omega_K$, для каждой из которых есть экспертная информация $E_k = E(\Omega_k)$, $k \in \{1, \dots, K\}$ ставится задача построения мульти модели, называемой смесью K экспертов.

Определение 11. *Мульти модель f называется смесью K экспертов*

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

где g_k является локальной моделью — экспертом, а \mathbf{x} является признаковым описанием объекта, π_k — шлюзовая функция, вектор \mathbf{w}_k является параметрами локальной модели, а матрица \mathbf{V} является параметрами шлюзовой функции.

Для каждой кривой второго порядка даны отображения (4.15). Обозначим $K_x^k(\mathbf{c}) = K_x(\Omega_k, \mathbf{c})$ и $K_y^k(\mathbf{c}) = K_y(\Omega_k, \mathbf{c})$. Затем с помощью локальных линейных моделей строится универсальная мультимодель, описывающая кривые $\Omega_1, \dots, \Omega_K$ на изображении \mathbf{M} :

$$f = \sum_{\mathbf{c} \in \mathbf{C}} \sum_{k=1}^K \pi_k(\mathbf{c}, \mathbf{V}) g_k(K_x^k(\mathbf{c}), \mathbf{w}_k), \quad (4.18)$$

где π_k задает шлюзовую функцию. Рассматривается случай, когда $\mathbf{x} = K_x^1(\mathbf{c}) = \dots = K_x^K(\mathbf{c})$, то есть выражение (4.18) принимает вид:

$$f = \sum_{\mathbf{c} \in \mathbf{C}} \sum_{k=1}^K \pi_k(\mathbf{x}, \mathbf{V}) g_k(\mathbf{x}, \mathbf{w}_k),$$

где шлюзовая функция π_k имеет вид:

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

где \mathbf{V} — параметры шлюзовой функции, а g_k — локальная модель.

Рассматривается вид функций распределения:

$$\boldsymbol{\pi}(\mathbf{x}, \mathbf{V}) = \text{softmax}(\mathbf{V}_1^\top \boldsymbol{\sigma}(\mathbf{V}_2^\top \mathbf{x})),$$

где $\mathbf{V} = \{\mathbf{V}_1, \mathbf{V}_2\}$ параметры шлюзовой функции, $\mathbf{V}_1 \in \mathbb{R}^{p \times k}$, $\mathbf{V}_2 \in \mathbb{R}^{n \times p}$.

Чтобы найти оптимальные параметры мультимодели, необходимо решить оптимизационную задачу:

$$\mathcal{L} = \sum_{(\mathbf{x}, y) \in \mathfrak{D}} \sum_{k=1}^K \pi_k(\mathbf{x}, \mathbf{V})(y - \mathbf{w}_k^\top \mathbf{x})^2 + R(\mathbf{V}, \mathbf{W}, E(\Omega)) \rightarrow \min_{\mathbf{V}, \mathbf{W}}, \quad (4.19)$$

где $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_k]$ — параметры локальных моделей, $R(\mathbf{V}, \mathbf{W}, E(\Omega))$ регуляризационный параметр на основе экспертной информации.

Единое пространство для кривых второго порядка. Произвольная кривая второго порядка, главная ось которой не параллельна оси ординат, задается выражением:

$$x^2 = B'xy + C'y^2 + D'x + E'y + F',$$

где на коэффициенты B', C' действуют ограничения, зависящие от типа кривой. Выражение (4.16) принимает вид:

$$K_x(\mathbf{c}_i) = [x_i y_i, y_i^2, x_i, y_i, 1], \quad K_y(\mathbf{c}_i) = x_i^2,$$

откуда получается задача линейной регрессии для восстановления параметров B', C', D', E', F' по выбранной выборке.

Окружность. Как частный случай кривой второго порядка, рассматривается окружность. Пусть (x_0, y_0) — это центр окружности на бинарном изображении \mathbf{M} , а r — его радиус. Элементы $(x_i, y_i) \in \mathbf{C}$ представляют собой геометрическое место точек. Это место требуется аппроксимировать уравнением окружности:

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

Раскладывая скобки, получаем:

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

Тогда выражение (4.16) принимает вид:

$$K_x(\mathbf{c}_i) = [x_i, y_i, 1] = \mathbf{x}, K_y(\mathbf{c}_i) = x_i^2 + y_i^2 = y.$$

Получаем задачу линейной регрессии (4.17). Компоненты вектора $\mathbf{w} = [w_0, w_1, w_2]^T$ связывают признаковое описание \mathbf{x} и целевую переменную y . Параметры окружности находятся из значений параметров линейной модели:

$$x_0 = \frac{w_0}{2}, \quad y_0 = \frac{w_1}{2}, \quad r = \sqrt{w_3 + x_0^2 + y_0^2}.$$

Композиция кривых второго порядка на изображении. Для построения композиции из фигур воспользуемся выражением (4.19):

$$\mathcal{L} = \sum_{\mathbf{c} \in \mathbf{C}} \sum_{k=1}^K \pi_k(\mathbf{c}, \mathbf{V}) (K_y^k(\mathbf{c}) - \mathbf{w}_k^T K_x^k(\mathbf{c}))^2 + R(\mathbf{V}, \mathbf{W}, E(\Omega)) \rightarrow \min_{\mathbf{V}, \mathbf{W}},$$

где K_x^k, K_y^k экспертное представление k -го эксперта. Предполагая, что все кривые на изображении описываются одним признаковым описанием $\mathbf{x} = K_x^1(\mathbf{c}) = \dots = K_x^K(\mathbf{c}), x = K_y^1(\mathbf{c}) = \dots = K_y^K(\mathbf{c})$, получаем задачу оптимизации:

$$\mathcal{L} = \sum_{(\mathbf{x}, y) \in \mathfrak{D}} \sum_{k=1}^K \pi_k(\mathbf{x}, \mathbf{V}) (y - \mathbf{w}_k^T \mathbf{x})^2 + R(\mathbf{V}, \mathbf{W}, E(\Omega)) \rightarrow \min_{\mathbf{V}, \mathbf{W}}, \quad (4.20)$$

В качестве регуляризатора R рассматриваются дополнительные ограничения на векторы параметров модели. Для решения задачи оптимизации (4.20) предлагается использовать ЕМ-алгоритм.

4.4. Анализ качества аппроксимации смесью экспертов

Для анализа качества различных мультимоделей для аппроксимации окружности проводится вычислительный эксперимент. В эксперимент рассматриваются

мультимодели: мультимодель \mathbf{f}_1 без использования априорных распределений, мультимодель \mathbf{f}_2 , которая использует априорные распределения (4.22) для параметров и мультимодель \mathbf{f}_3 , которая использует регуляризацию априорных распределений. Точность аппроксимации мультимодели \mathbf{f}_i задается:

$$\mathcal{S}_{\mathbf{f}_i} = \sum_{k=1}^K (x_0^k - x_{\text{pr}}^k)^2 + (y_0^k - y_{\text{pr}}^k)^2 + (r^k - r_{\text{pr}}^k)^2, \quad (4.21)$$

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 (4.5). В вычислительном эксперименте используется априорное распределение:

$$p^1(\mathbf{w}_1) \sim \mathcal{N}(\mathbf{w}_1^0, \mathbf{I}), \quad p^2(\mathbf{w}_2) \sim \mathcal{N}(\mathbf{w}_2^0, \mathbf{I}), \quad (4.22)$$

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 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. 4.4 показаны результаты для мультиформелей $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$. Все модели оптимизировались при помощи 50 итераций ЕМ-алгоритма. Мультиформели $\mathbf{f}_2, \mathbf{f}_3$ аппроксимируют окружности лучше чем мультиформель \mathbf{f}_1 . В табл. 4.1 показано качество аппроксимации (4.21) для всех мультиформелей.

Таблица 4.1: Качество аппроксимации мультиформели в зависимости от априорных распределений

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}

Анализ сходимости на синтетической выборке.

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

На рис. 4.5 показана зависимость предсказано центра и радуса в зависимости от номера итерации ЕМ-алгоритма. Мультиформель \mathbf{f}_2 , использующая априорное

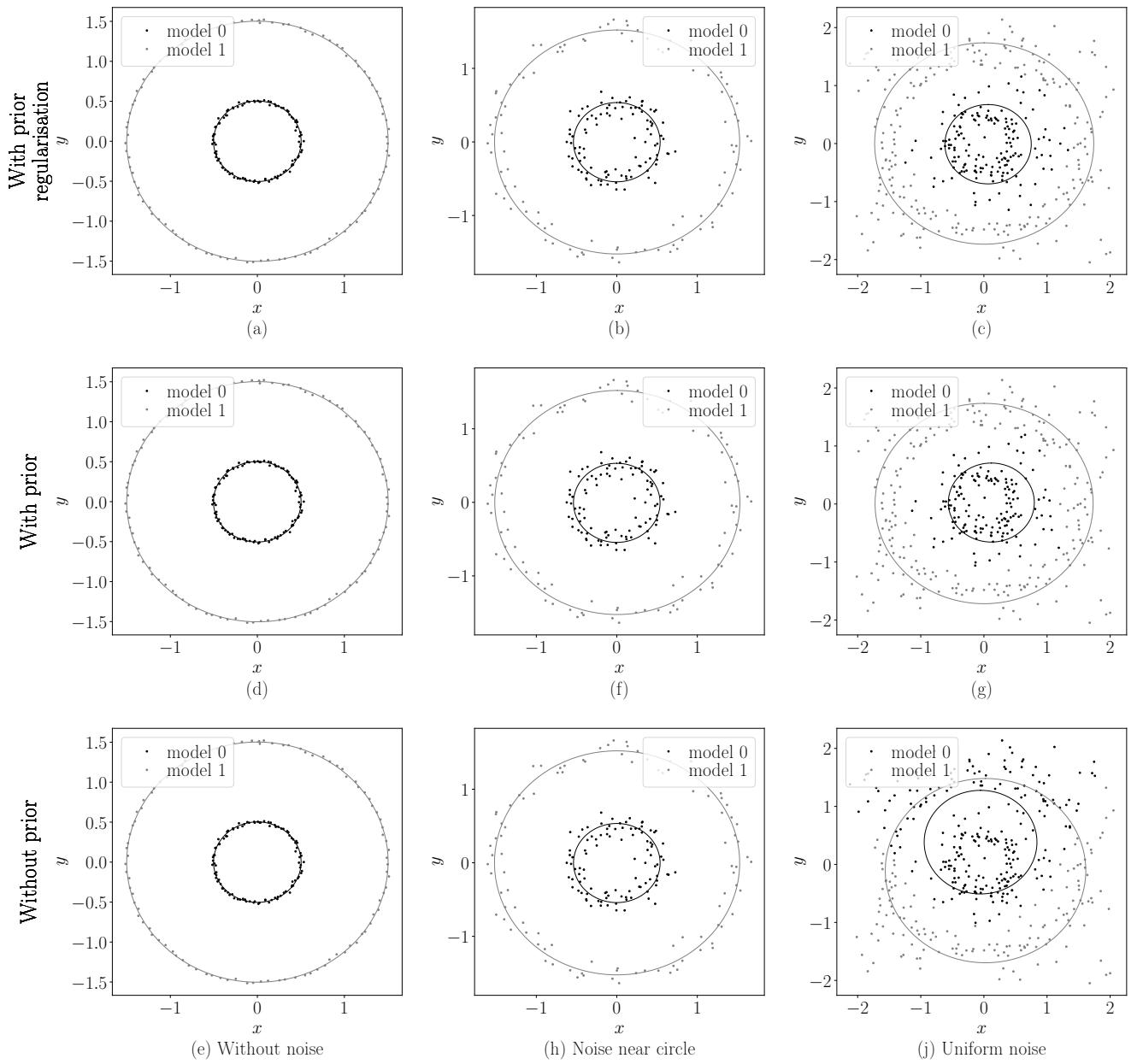


Рис. 4.4: Мульти модель в зависимости от разных априорных предположений и в зависимости от разного уровня шума: (а)–(с) модель с регуляризацией априорных распределений; (д)–(г) модель с заданными априорными распределениями на параметрах локальных моделей; (е)–(ж) модель без заданных априорных предположений

распределение, аппроксимирует окружность лучше мульти модели \mathbf{f}_1 , которая не использует никакого априорного распределения. Мульти модель \mathbf{f}_3 , использующая регуляризатор априорных распределений, является более стабильной, чем мульти модель \mathbf{f}_2 .

На рис. 4.6 показана зависимость логарифма правдоподобия (4.5) от номера итерации ЕМ-алгоритма. Логарифм правдоподобия мульти модели $\mathbf{f}_2, \mathbf{f}_3$ растет быстрее чем логарифм правдоподобия мульти модели \mathbf{f}_1 . После 20-й итерации все мульти модели имеют одинаковое правдоподобие.

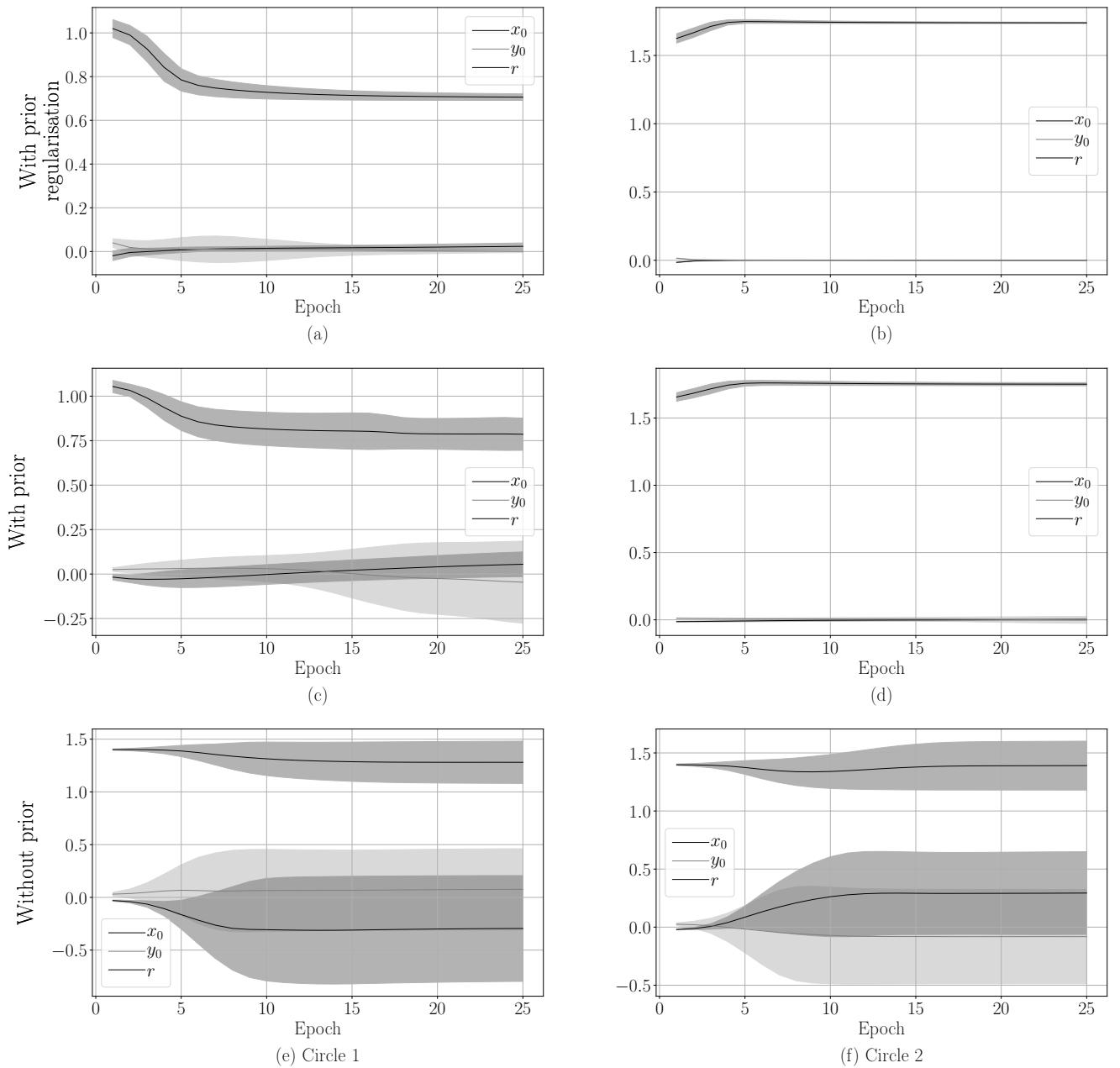


Рис. 4.5: График зависимости центра и радиуса окружностей от номера итерации: (a)–(b) модель с регуляризацией априорных распределений; (c)–(d) модель с заданными априорными распределениями на параметры моделей; (e)–(f) модель без задания априорных распределений

На рис. 4.7-4.9 показан процесс сходимости для разных мультимоделей $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$. На рис. 4.9 показана мультимодель \mathbf{f}_1 , которая аппроксимирует окружности не верно. На рис. 4.7-4.8 показаны мультимодели $\mathbf{f}_2, \mathbf{f}_3$, которые аппроксимируют окружности верно.

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

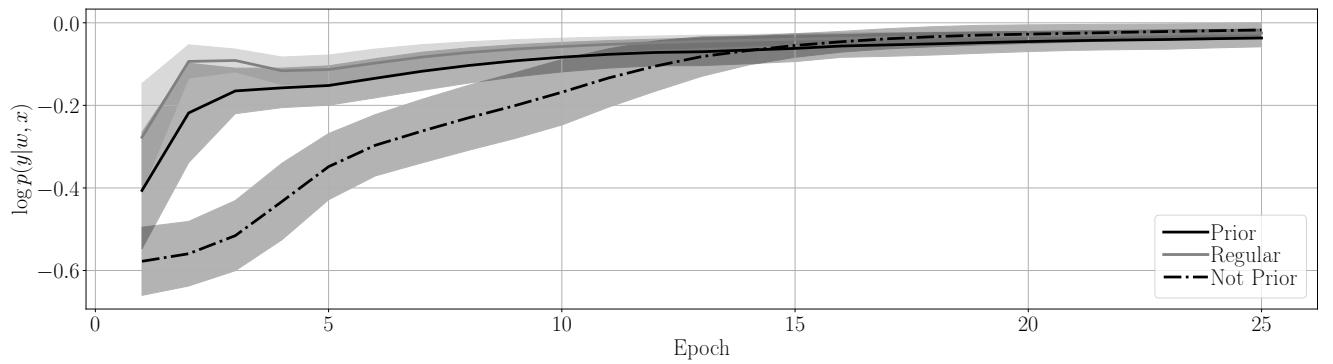


Рис. 4.6: График зависимости логарифма правдоподобия (4.5) от номера итерации

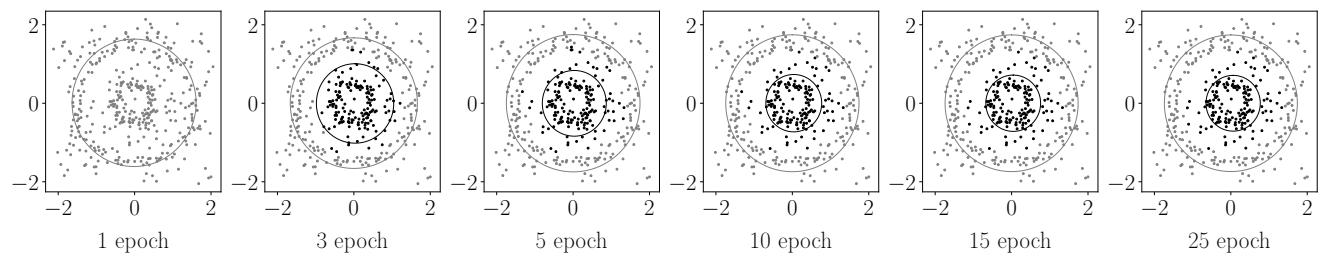


Рис. 4.7: Визуализации процесса сходимости мультимодели с использованием априорной регуляризации

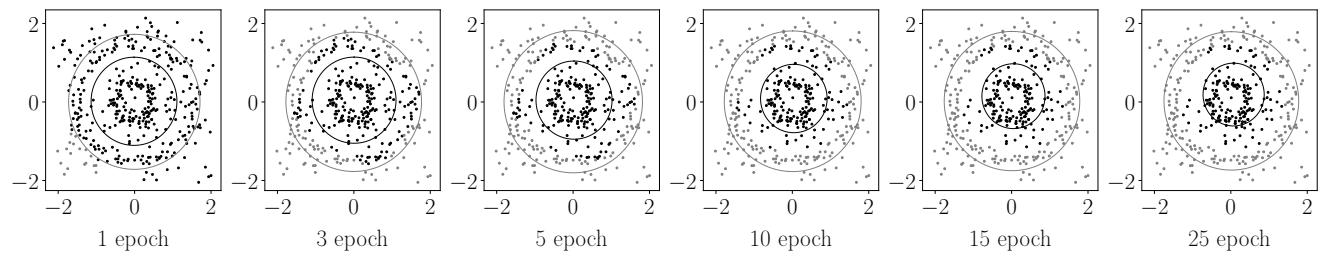


Рис. 4.8: Визуализации процесса сходимости мультимодели с использованием априорного распределения

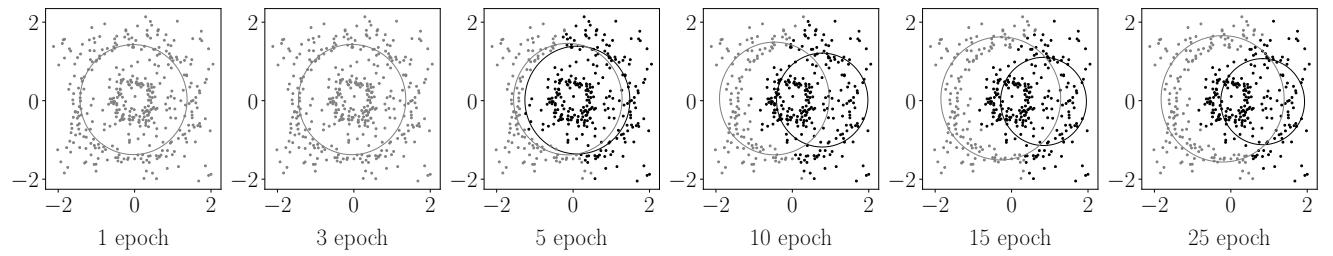


Рис. 4.9: Визуализации процесса сходимости мультимодели без использования априорного распределения

Анализ мультимелей в зависимости от уровня шума.

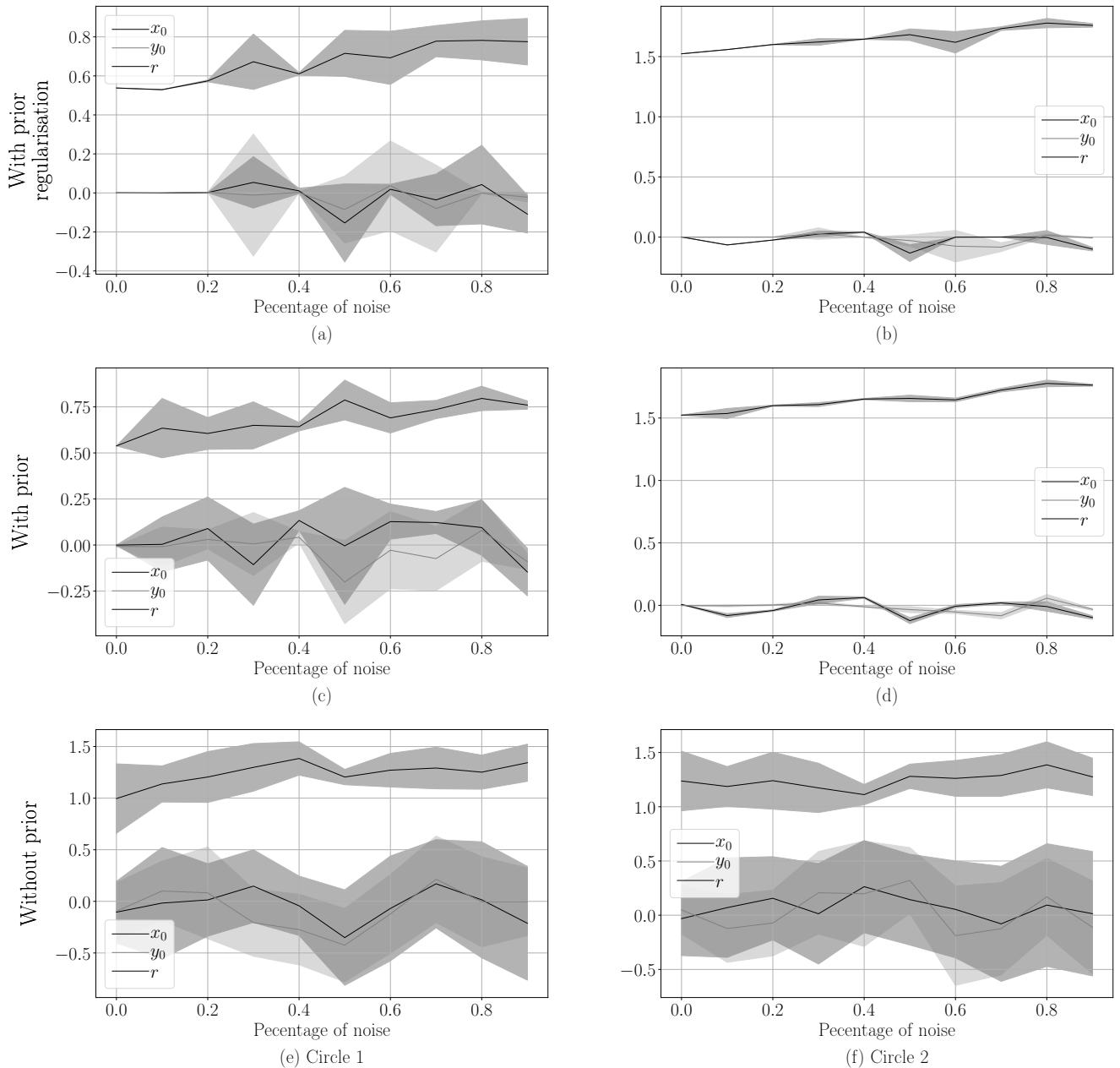


Рис. 4.10: График зависимости центра и радиуса окружностей от номера итерации: (a)–(b) модель с регуляризацией априорных распределений; (c)–(d) модель с заданными априорными распределениями на параметры моделей; (e)–(f) модель без задания априорных распределений

Данная часть эксперимента анализирует зависимость разных мультимелей f_1, f_2, f_3 от уровня шума. Анализ всех мультимелей 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. 4.10 показан график зависимости центра окружности и ее радиус в зависимости от уровня шума. Из графика видно, что радиус окружности

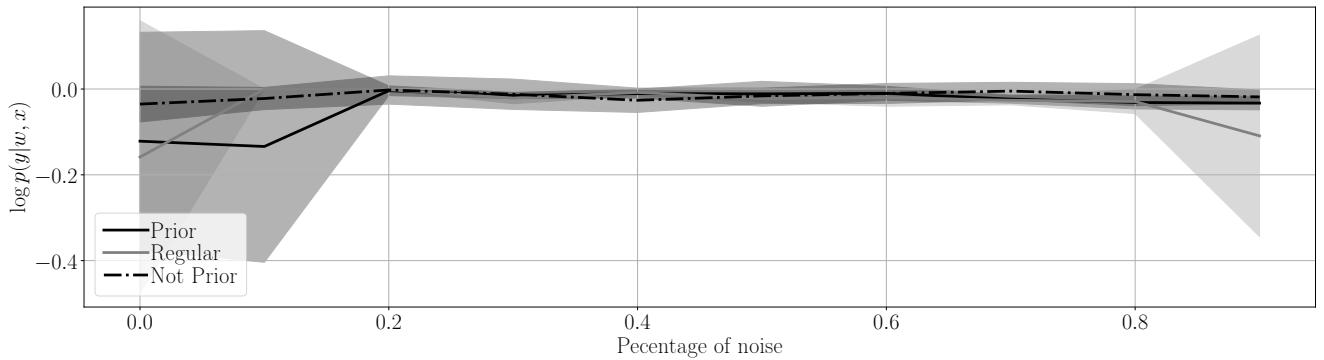


Рис. 4.11: График зависимости логарифма правдоподобия (4.5) от уровня шума

увеличивается при увеличении уровня шума. Мульти модели f_2, f_3 аппроксимируют центр окружности верно, но мульти модель f_3 более устойчива к шуму . На рис. 4.11 показана зависимость логарифма правдоподобия (4.5) от уровня шума. Из графика видно, что логарифм правдоподобия (4.5) эквивалентный для всех мульти моделей, но на рис. 4.10 видно, что качество аппроксимации (4.21) зависит от мульти модели. Данная часть вычислительного эксперимента показывает, что мульти модель f_3 с регуляризацией априорного распределения является более устойчива к шуму, чем остальные.

Реальные данные.

Данная часть эксперимента анализирует разные мульти модели f_1, f_2, f_3 на реальной выборке. На рис. 4.12 показан результат работы разных мульти моделей. Мульти модель f_1 не верно аппроксимирует меньшую окружность. Мульти модели f_2, f_3 аппроксимируют обе окружности верно.

На рис. 4.13-4.15 показан процесс аппроксимации для разных мульти моделей f_1, f_2, f_3 .

Данная часть эксперимента показывает, что мульти модели f_2, f_3 аппроксимируют окружности на реальных изображениях лучше, чем мульти модель f_1 .

Проведен вычислительный эксперимент по анализу качества моделей кривых второго порядка на изображении. Эксперимент разделен на несколько частей. Первая часть - это эксперимент с несколькими окружностями на изображении. Во второй части анализируется сходимость метода в зависимости от уровня шума в данных и от указанной экспертной информации. В третьей части проводится эксперимент по аппроксимации радужной оболочки глаза.

В этой части эксперимента показан пример обучения мульти модели для аппроксимации нескольких фигур второго порядка одновременно. В качестве данных используется синтетическая выборка с тремя произвольными непересекающимися окружностями, а также добавления шума к этим окружностям. Шум добавлен к радиусу круга для каждой точки, а также случайные точки добавлены к выборке.

На рисунке 4.16 показан результат построения ансамбля локально аппроксимирующих моделей, которые аппроксимируют образец. Каждая локальная мо-

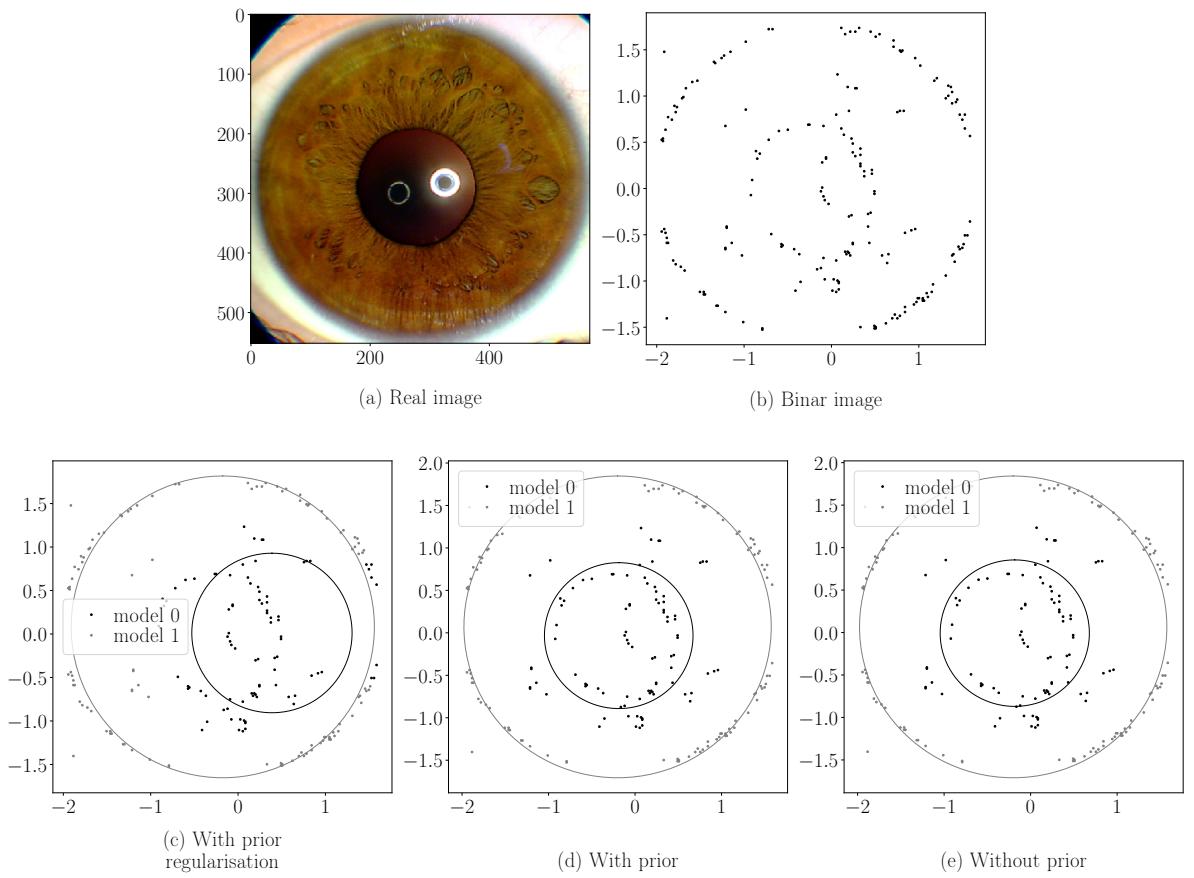


Рис. 4.12: Мульти модель в зависимости от разных априорных предположений на реальном изображении: (а) исходное изображение; (б) бинаризованое изображение; (с) мульти модель без априорных предположений; (д) мульти модель с априорными распределениями на параметрах локальных моделей; (е) мульти модель с регуляризацией на априорных распределениях параметров локальных моделей

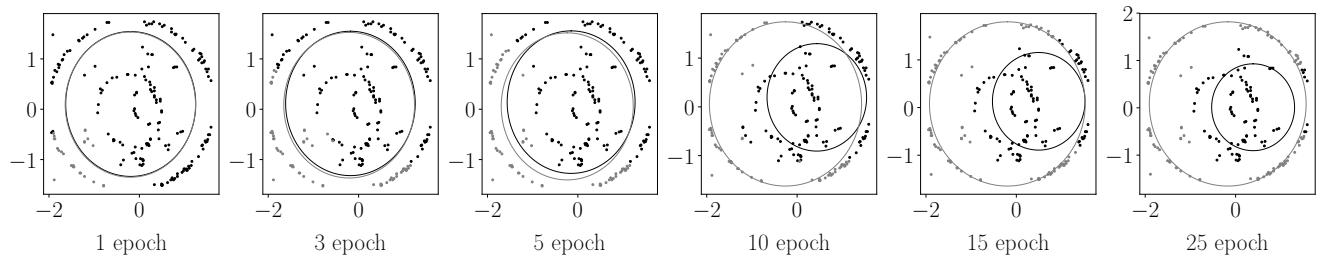


Рис. 4.13: Визуализации процесса сходимости мульти модели без использования априорного распределения

дель аппроксимирует одну окружность, а при добавлении еще большего шума качество аппроксимации падает. На рисунке 4.17 показан график зависимости радиуса окружностей r и их центров (x_0, y_0) от номера итерации.

В этой части эксперимента анализируется качество аппроксимации S в зависимости от уровня шума β в данных и по параметру априорных распреде-

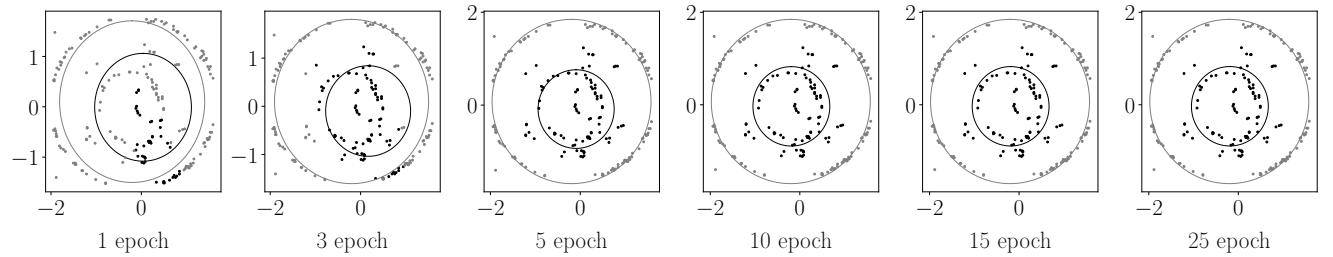


Рис. 4.14: Визуализации процесса сходимости мультимодели с использованием априорного распределением

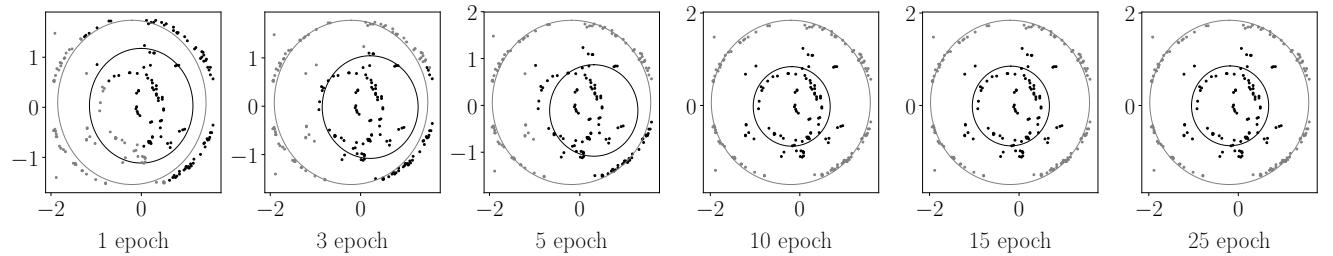


Рис. 4.15: Визуализации процесса сходимости мультимодели с использованием априорной регуляризации

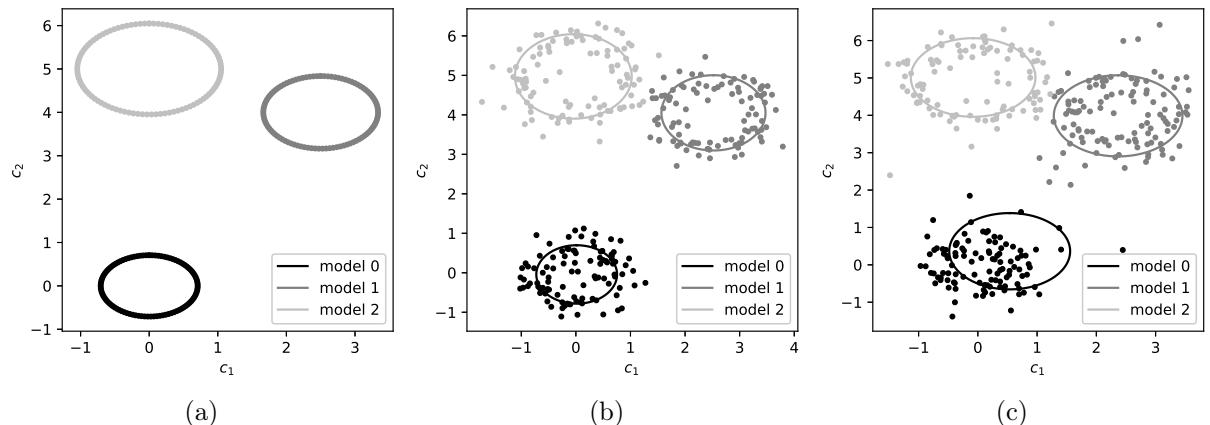


Рис. 4.16: Мультимодель в зависимости от различных предварительных предположений и уровня шума. Слева направо: окружности без шума; шум в радиусе круга; шум в радиусе круга, а также произвольные точки по всему изображению

лений γ . Алгоритм генерации: сначала случайным образом выбираются два вектора параметров: $\mathbf{w}_1^{\text{true}}$ и $\mathbf{w}_2^{\text{true}}$ коэффициенты двух парабол. Векторы $\mathbf{w}_1^{\text{true}}$ и $\mathbf{w}_2^{\text{true}}$ используются для создания точек x_i и y_i с добавлением нормального шума $\varepsilon \sim \mathcal{N}(0, \beta)$. При обучении мультимодели, априорное распределение параметров считается $\mathbf{w}_1 \sim \mathcal{N}(\mathbf{w}_1^{\text{true}}, \gamma \mathbf{I})$, $\mathbf{w}_2 \sim \mathcal{N}(\mathbf{w}_2^{\text{true}}, \gamma \mathbf{I})$.

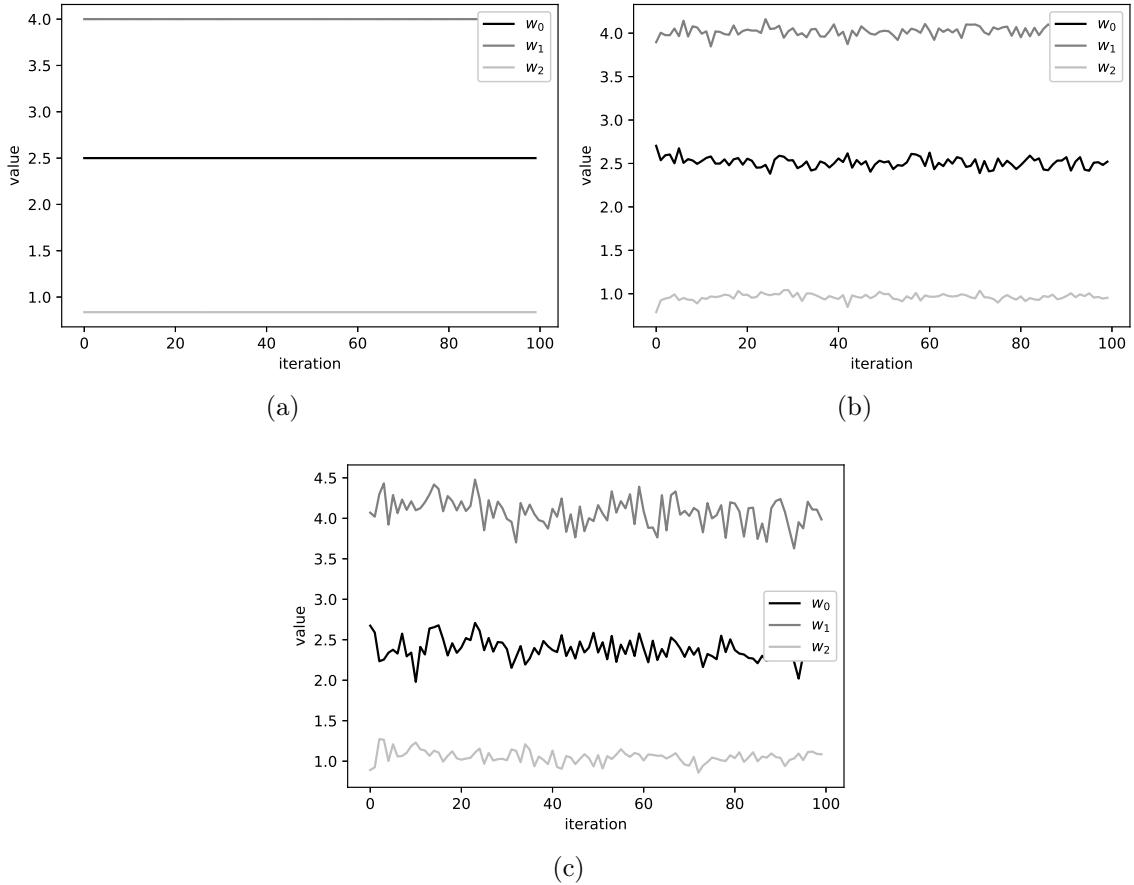


Рис. 4.17: Зависимость параметров r , x_0 и y_0 от номера итерации для различных априорных распределений. Слева направо: окружности без шума; шум в радиусе круга; шум в радиусе круга, а также произвольные точки по всему изображению

Рассмотрим критерий качества:

$$S = \|\mathbf{w}_1^{\text{pred}} - \mathbf{w}_1^{\text{true}}\|_2^2 + \|\mathbf{w}_2^{\text{pred}} - \mathbf{w}_2^{\text{true}}\|_2^2,$$

где $\mathbf{w}_1^{\text{pred}}$ аппроксимация вектора параметров первой локальной модели, а $\mathbf{w}_2^{\text{pred}}$ аппроксимация вектора параметров второй локальной модели.

На рисунке 4.18 показана зависимость критерия качества S от уровня шума β и параметра априорного распределения γ . График показывает, что при низком уровне шума β качество приближения не зависит от параметра γ , а с увеличением шума β качество приближения S снижается.

На рисунке 4.19 показан пример того, как алгоритм работает с разными параметрами β и γ . Видно, что в отсутствие шума β обе локальные модели аппроксимируют образец. С увеличением уровня шума качество аппроксимации снижается: при $\beta = 0,2$ при увеличении γ первая локальная модель от параболы переходит в эллипс; при $\beta = 0,4$ при увеличении γ первая локальная модель из параболы переходит в эллипс, а вторая модель из параболы переходит в гиперболу.

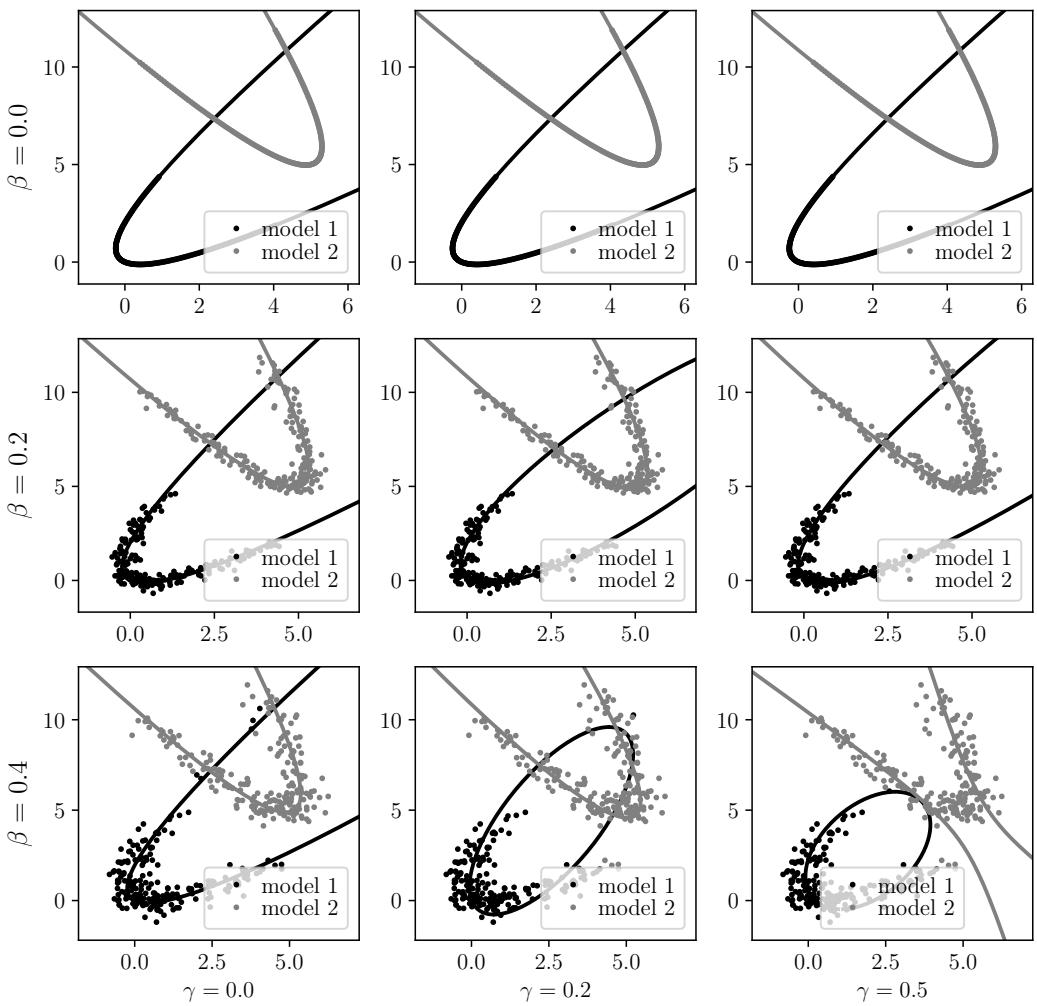


Рис. 4.18: Результат аппроксимации для данных с разными уровнями шума β и дисперсией априорного распределения γ

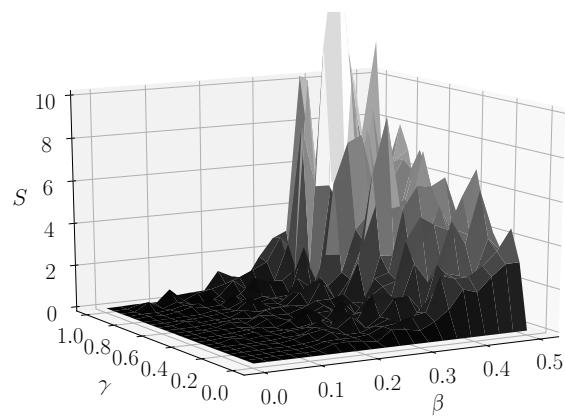


Рис. 4.19: Зависимость моделей от уровня шума β в данных, а также от дисперсии априорного распределения γ

Анализ качества аппроксимации проводится для задачи аппроксимации радужной оболочки глаза на изображении. Радужная оболочка глаза состоит из

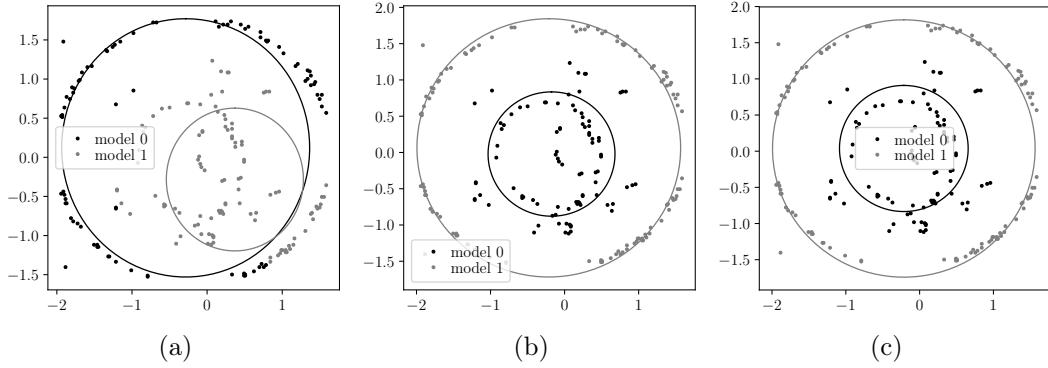


Рис. 4.20: Визуализация приближения радужной оболочки: а) если указан регуляризатор R_0 ; б) если указан регуляризатор R_1 ; в) если указан регуляризатор R_2

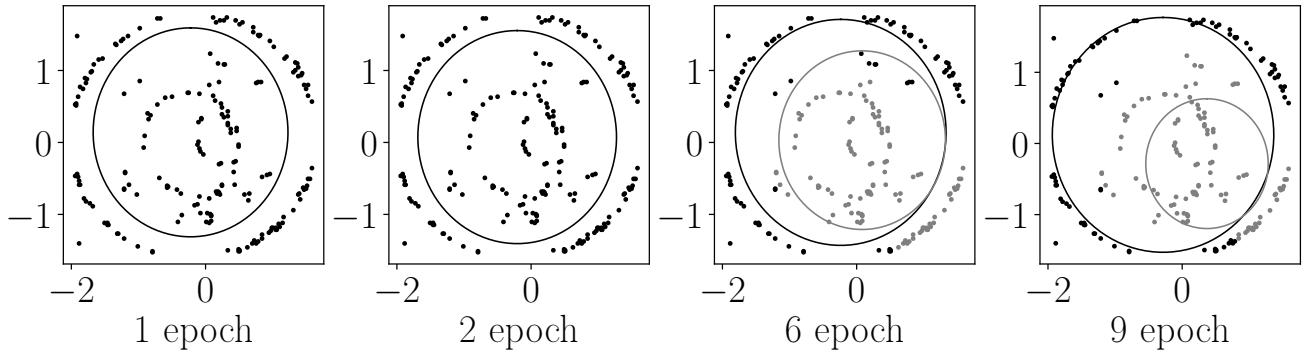


Рис. 4.21: Визуализация процесса сходимости параметров мульти модели в случае регуляризатора R_0

двух концентрических окружностей, поэтому рассматривается мульти модель, состоящая из двух экспертов: каждый эксперт аппроксимирует одно из обстоятельств. В вычислительном эксперименте сравнивается качество аппроксимации окружностей при задании разных регуляризаторов R_0, R_1, R_2 . Регуляризатор $R_0(\mathbf{V}, \mathbf{W}, E(\Omega)) = 0$, то есть регуляризатора нет. Регуляризатор

$$R_1(\mathbf{V}, \mathbf{W}, E(\Omega)) = - \sum_{k=1}^K \mathbf{w}_k^\top \mathbf{w}_k,$$

удаляет оклонулевые параметры локальных моделей. Регуляризатор

$$R_2(\mathbf{V}, \mathbf{W}, E(\Omega)) = - \sum_{k=1}^K \mathbf{w}_k^\top \mathbf{w}_k + \sum_{k=1}^K \sum_{k'=1}^K \sum_{j=1}^2 (w_k^j - w_k'^j)^2$$

способствует совпадению центров окружностей и близким к нулю параметрам модели. На рисунке 4.20 показан результат алгоритма аппроксимации радужной оболочки глаза после 10 итераций. Видно, что при отсутствии регуляризатора

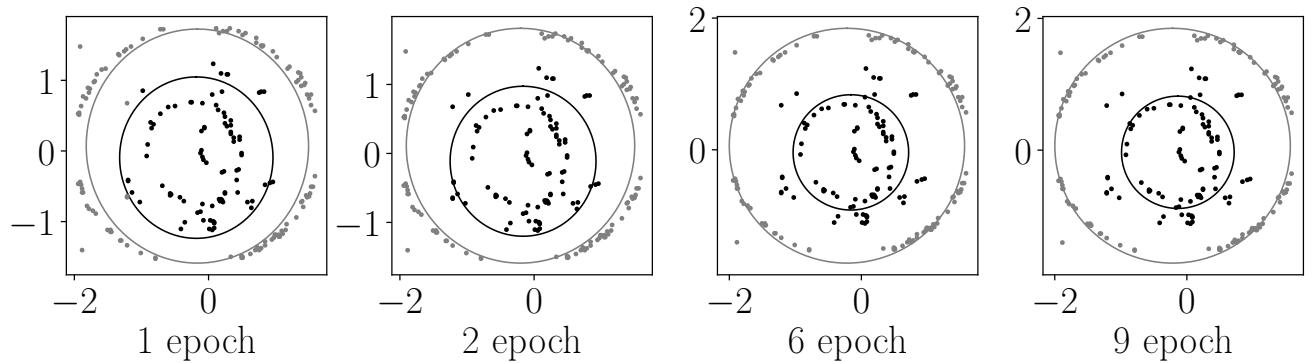


Рис. 4.22: Визуализация процесса сходимости параметров мультимодели в случае регуляризатора R_1

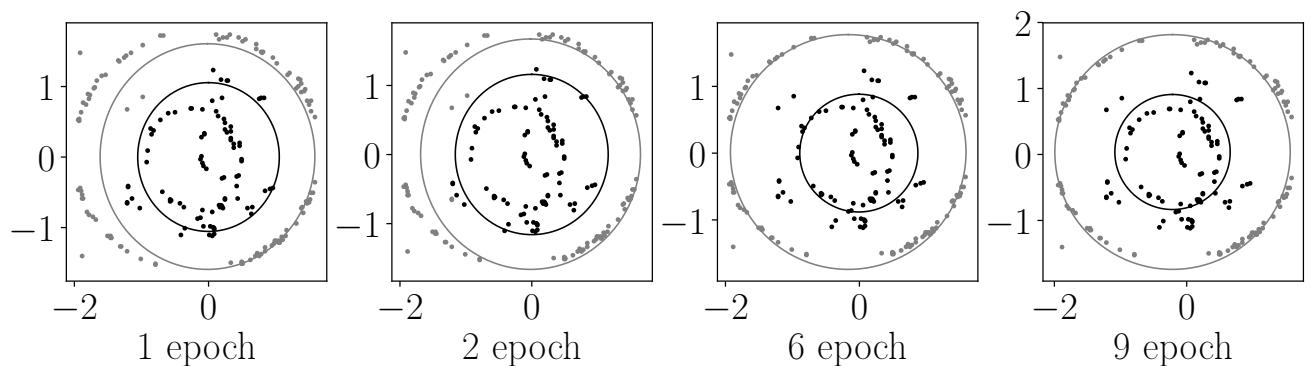


Рис. 4.23: Визуализация процесса сходимости параметров мультимодели в случае регуляризатора R_2

ра одна из окружностей находится некорректно. Если задан регуляризатор R_1 , модель аппроксимирует обе окружности с хорошим качеством, но окружности не концентрические. В случае задания регуляризатора R_2 мы получим концентрические окружности на изображении.

На рисунках 4.21-4.23 показан процесс сходимости мультимоделей в случае задания разных регуляризаторов R_0, R_1, R_2 . Видно, что модели с типом регуляризатора R_1 и R_2 аппроксимируют обе окружности, а мультимодель с регуляризатором R_0 аппроксимирует только большую окружность.

Глава 5

Введение отношения порядка на множестве параметров аппроксимирующих моделей

Рассматривается проблема введения отношения порядка на множестве параметров сложных аппроксимирующих моделей. В качестве параметрических моделей исследуются линейные и нейросетевые модели. Предполагается, что число параметров нейросети можно существенно снизить без значимой потери качества и значимого повышения дисперсии функции ошибки. Предлагаются методы задания порядка на основе ковариационной матрицы градиентов функции ошибки по параметрам модели и на основе ковариационной матрицы апостериорного распределения параметров. Показано, что полученный порядок указывает на релевантность параметров в рамках обученной нейросетевой модели: от наименее релевантных параметров до наиболее релевантных параметров.

Множество параметров упорядочивается по возрастанию дисперсии: от параметра с минимальной дисперсией до параметра с максимальной дисперсией градиента функции ошибки по соответствующему параметру модели. Предполагается, что малая дисперсия градиента указывает на то, что соответствующий параметр можно зафиксировать.

Для задания порядка на множестве параметров при помощи ковариационной матрицы вводится предположение о том, что фиксация параметров происходит в момент, когда все параметры модели находятся в некоторой окрестности локального минимума функции ошибки. Данное условие накладывается для корректного использования итерационного метода поиска ковариационной матрицы градиентов.

Заданный порядок на множестве параметров модели используется для фиксации тех параметров модели, которые оказываются предстоящими с точки зрения заданного порядка. Сначала фиксируются те параметры, которые имеют минимальную дисперсию градиента в окрестности локального минимума функции ошибки.

Другим подходом к определению релевантности параметров предлагается анализ ковариационной матрицы параметров модели на основе метода Белсли. Метод Белсли [66] позволяет оценить мультиколлинеарность параметров, после чего удалить наиболее коррелирующие параметры.

Для анализа свойств предложенного метода задания порядка на множестве параметров проводился вычислительный эксперимент. В качестве моделей рассматривались модели различной структурной сложности: линейные модели, нейросетевые модели. Предложенный метод задания порядка сравнивается с методом, в котором порядок задан произвольным образом.

5.1. Задача упорядочивания параметров аппроксимирующих моделей

Задана выборка:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \quad \mathbf{x}_i \in \mathbb{X} = \mathbb{R}^n, \quad y_i \in \mathbb{Y},$$

где n — размерность признакового пространства, m — число объектов в выборке. Пространство ответов $\mathbb{Y} = \mathbb{R}$ в случае задачи регрессии и $\mathbb{Y} = \{1, \dots, R\}$ в случае задачи классификации, где R — число классов.

Задано семейство моделей параметрических функций с наперед заданной структурой:

$$\begin{aligned} \mathfrak{F} &= \{f(\mathbf{w}) : \mathbb{X} \rightarrow \mathbb{Y} | \mathbf{w} \in \mathbb{R}^p\}, \\ \mathbf{h}(\mathbf{w}, \mathbf{x}) &= \mathbf{W}_1 \boldsymbol{\sigma}(\mathbf{W}_2 \boldsymbol{\sigma}(\dots \boldsymbol{\sigma}(\mathbf{W}_r \mathbf{x}) \dots)), \\ f_{\text{cl}}(\mathbf{w}, \mathbf{x}) &= \arg \max_{j \in \{1, \dots, R\}} \text{softmax}(\mathbf{h}(\mathbf{w}, \mathbf{x}))_j, \\ f_{\text{reg}}(\mathbf{w}, \mathbf{x}) &= \mathbf{h}(\mathbf{w}, \mathbf{x}), \end{aligned} \tag{5.1}$$

где p — размерность пространства параметров, r — число слоев нейросети, $\mathbf{w} = \text{vec}[\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_r]$, а $\boldsymbol{\sigma}$ — функция активации. В случае задачи регрессии структура модели имеет вид f_{reg} , а в случае классификации имеет вид f_{cl} . Задана функция потерь:

$$\begin{aligned} \mathcal{L}(\mathbf{w}, \mathfrak{D}) &= \frac{1}{m} \sum_{i=1}^m l(\mathbf{x}_i, y_i, \mathbf{w}), \\ l_{\text{reg}}(\mathbf{x}, y, \mathbf{w}) &= (y - f(\mathbf{w}, \mathbf{x}))^2, \\ l_{\text{cl}}(\mathbf{x}, y, \mathbf{w}) &= - \sum_{j=1}^R ([y = j] \ln \text{softmax}_j(\mathbf{h}(\mathbf{w}, \mathbf{x}))), \end{aligned} \tag{5.2}$$

где l_{reg} — это функция ошибки на одном элементе для задачи регрессии, l_{cl} — для задачи классификации. Оптимальный вектор параметров $\hat{\mathbf{w}}$ получим минимизацией функции потерь:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^p} \mathcal{L}(\mathbf{w}, \mathfrak{D}).$$

Для поиска оптимальных параметров модели используется градиентный метод оптимизации:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \Delta \mathbf{w}(\mathbf{g}_{S,t}, \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \dots), \quad \mathbf{g}_{S,t} = \frac{\partial \mathcal{L}(\mathbf{w}_t, \mathbf{X}_S, \mathbf{Y}_S)}{\partial \mathbf{w}}, \tag{5.3}$$

где t — номер итерации, $\mathbf{g}_{S,t}$ — значение градиента на подвыборке размера S , $\Delta \mathbf{w}$ — приращение вектора параметров.

Порядок на множестве параметров модели задается при помощи ковариационной матрицы \mathbf{C} градиентов функции ошибки \mathcal{L} по параметрам модели \mathbf{w} . Для вычисления ковариационной матрицы \mathbf{C} используется итерационная формула [2], которая вычисляется на каждой итерации (5.3) градиентного метода оптимизации параметров:

$$\mathbf{C}_t = (1 - \kappa_t) \mathbf{C}_{t-1} + \kappa_t (\mathbf{g}_{1,t} - \mathbf{g}_{S,t}) (\mathbf{g}_{1,t} - \mathbf{g}_{S,t})^\top,$$

где t — номер итерации, $\mathbf{g}_{S,t}$ — значение градиента на подвыборке размера S , $\mathbf{g}_{1,t}$ — значение градиента на первом элементе подвыборки, $\kappa_t = \frac{1}{t}$ — параметр сглаживания, \mathbf{C}_0 инициализируются из равномерного распределения.

Пусть известно t_0 — число итераций, после которого все параметры находятся в некоторой локальной окрестности минимума, тогда, как показано в работе [2], матрица \mathbf{C}_{t_0} аппроксимирует истинную ковариационную матрицу \mathbf{C} . Ковариационная матрица \mathbf{C}_{t_0} используется для упорядочения параметров модели \mathbf{w}_{t_0} .

Пусть \mathcal{I} — упорядоченный вектор индексов $[1, 2, \dots, p]$. Обозначим $\mathcal{I}_{\mathbf{w}_{t_0}}$ вектор индексов, порядок которого задан при помощи ковариационной матрицы \mathbf{C}_{t_0} .

Например, если ковариационная матрица \mathbf{C}_{t_0} имеет вид

$$\begin{bmatrix} 0,3 & 0 & 0 \\ 0 & 0,2 & 0 \\ 0 & 0 & 0,25 \end{bmatrix},$$

то вектор индексов $\mathcal{I}_{\mathbf{w}_{t_0}} = [3, 1, 2]$.

Фиксация параметров модели в процессе обучения. Для фиксации параметров \mathbf{w}_{t_0} при помощи вектора индексов $\mathcal{I}_{\mathbf{w}_{t_0}}$ используется бинарный вектор $\boldsymbol{\alpha}(\zeta)$:

$$\alpha_i(\zeta) = \begin{cases} 1, & \text{если } \mathcal{I}_{\mathbf{w}_{t_0}}[j] \leq \zeta; \\ 0 & \text{иначе,} \end{cases} \quad (5.4)$$

где ζ — число фиксирующих параметров.

Учитывая (5.4), уравнение (5.3) приводится к виду

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \boldsymbol{\alpha}(\zeta) \cdot \Delta \mathbf{w}(\mathbf{g}_{S,t}, \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \dots), \quad (5.5)$$

где t — номер итерации, $\mathbf{g}_{S,t}$ — значение градиента на подвыборке размера S , $\Delta \mathbf{w}$ — приращение вектора параметров. После умножения на бинарный вектор $\boldsymbol{\alpha}$ часть параметров не оптимизируется, что приводит к фиксации параметров.

5.2. Определение релевантности на основе метода Белсли

Предлагается метод основанный на модификации метода Белсли. Пусть \mathbf{w} — вектор параметров доставляющий минимум функционалу потерь \mathcal{L} на множестве $\mathbb{W}_{\mathcal{A}}$, а \mathbf{A}_{ps} соответствующая ему ковариационная матрица.

Выполним сингулярное разложение матрицы

$$\mathbf{A}_{\text{ps}} = \mathbf{U} \Lambda \mathbf{V}^T.$$

Индекс обусловленности η_j определим как отношение максимального элемента к j -му элементу матрицы Λ . Для нахождения мультиколлинеарных признаков требуется найти индекс ξ вида:

$$\xi = \arg \max_{j \in \mathcal{A}} \eta_j.$$

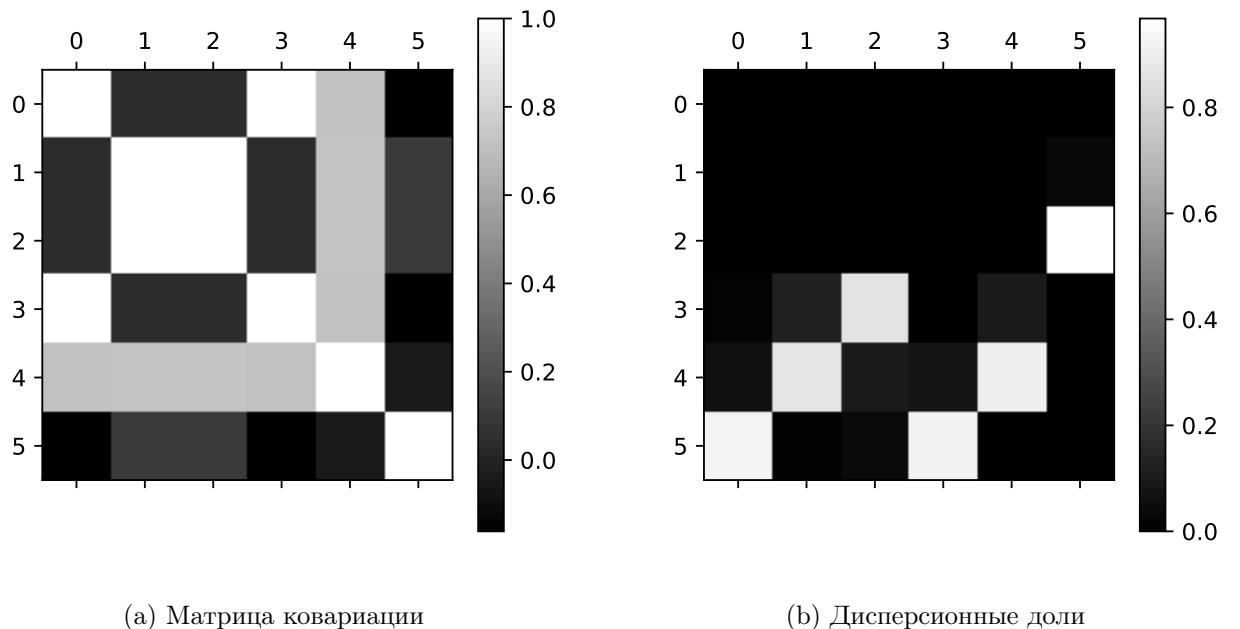


Рис. 5.1: Иллюстрация метода Белсли для анализа мультиколлинеарности параметров

Таблица 5.1: Иллюстрация метода Белсли для анализа мультиколлинеарности параметров

η	q_1	q_2	q_3	q_4	q_5	q_6
1.0	$2 \cdot 10^{-17}$	$4 \cdot 10^{-17}$	$1 \cdot 10^{-16}$	$2 \cdot 10^{-17}$	$6 \cdot 10^{-17}$	$3 \cdot 10^{-4}$
1.5	$5 \cdot 10^{-17}$	$9 \cdot 10^{-17}$	$2 \cdot 10^{-16}$	$5 \cdot 10^{-17}$	$3 \cdot 10^{-20}$	$3 \cdot 10^{-2}$
3.3	$9 \cdot 10^{-18}$	$1 \cdot 10^{-17}$	$2 \cdot 10^{-17}$	$9 \cdot 10^{-18}$	$2 \cdot 10^{-19}$	$9 \cdot 10^{-1}$
$2 \cdot 10^{15}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-1}$	$8 \cdot 10^{-1}$	$2 \cdot 10^{-3}$	$9 \cdot 10^{-2}$	$1 \cdot 10^{17}$
$8 \cdot 10^{15}$	$6 \cdot 10^{-2}$	$8 \cdot 10^{-1}$	$9 \cdot 10^{-2}$	$8 \cdot 10^{-2}$	$9 \cdot 10^{-1}$	$2 \cdot 10^{17}$
$1 \cdot 10^{16}$	$9 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$9 \cdot 10^{-1}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-21}$

Дисперсионный долевой коэффициент q_{ij} определим как вклад j -го признака в дисперсию i -го элемента вектора параметра \mathbf{w} :

$$q_{ij} = \frac{u_{ij}^2 / \lambda_{jj}}{\sum_{j=1}^n u_{ij}^2 / \lambda_{jj}}.$$

Большие значение дисперсионных долей указывают на наличие зависимости между параметрами. Находим долевые коэффициенты, которые вносят максимальный вклад в дисперсию параметра w_ξ :

$$\zeta = \arg \max_{j \in \mathcal{A}} q_{\xi j}.$$

Параметр с индексом ζ определим как наименее релевантный параметр нейросети.

Проиллюстрируем принцип работы метода Белсли на примере. Гипотеза порождения данных:

$$\mathbf{w} = \begin{bmatrix} \sin(x) \\ \cos(x) \\ 2+\cos(x) \\ 2+\sin(x) \\ \cos(x) + \sin(x) \\ x \end{bmatrix}$$

с матрицей ковариации на рис. 5.1.a, где $x \in [0.0, 0.02, \dots, 20.0]$.

В табл. 5.1 приведены индексы обусловленности и соответствующие им дисперсионные доли, которые также изображены на рис. 5.1.b. Согласно этим данным, максимальный индекс обусловленности $\eta_6 = 1.2 \cdot 10^{16}$. Ему соответствуют максимальные дисперсионные доли признаков с индексами 1 и 4, которые, как видно из построения выборки, являются линейно зависимые.

5.3. Анализ разных подходов к определению релевантности

Для анализа свойств предложенного алгоритма и сравнения его с существующими проведен вычислительный эксперимент в котором параметры нейросети удалялись методами, которые описаны в разделах 3.1–3.3 и методом Белсли.

В качестве данных использовались три выборки. Выборки Wine [67] и Boston Housing [68] — это реальные данные. Синтетические данные сгенерированы таким образом чтобы параметры сети мультиколлинеарными. Генерация данных состояла из двух этапов. На первом этапе генерировался вектор параметров $\mathbf{w}_{\text{synthetic}}$:

$$\mathbf{w}_{\text{synthetic}} \sim \mathcal{N}(\mathbf{m}_{\text{synthetic}}, \mathbf{A}_{\text{synthetic}}),$$

$$\text{где } \mathbf{m}_{\text{synthetic}} = \begin{bmatrix} 1.0 \\ 0.0025 \\ \dots \\ 0.0025 \end{bmatrix}, \mathbf{A}_{\text{synthetic}} = \begin{bmatrix} 1.0 & 10^{-3} & \dots & 10^{-3} & 10^{-3} \\ 10^{-3} & 1.0 & \dots & 0.95 & 0.95 \\ \dots & \dots & \dots & \dots & \dots \\ 10^{-3} & 0.95 & \dots & 0.95 & 1.0 \end{bmatrix}.$$

На втором этапе генерировалась выборка $\mathfrak{D}_{\text{synthetic}}$:

$$\mathfrak{D}_{\text{synthetic}} = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \sim \mathcal{N}(\mathbf{1}, \mathbf{I}), y_i = x_{i0}, i = 1 \dots 10000\}.$$

В приведенном выше векторе параметров $\mathbf{w}_{\text{synthetic}}$ для выборки $\mathfrak{D}_{\text{synthetic}}$, наиболее релевантным является первый параметр, а все остальные параметры являются нерелевантными. Матрица ковариации выбрана таким образом, чтобы все нерелевантные параметры являлись зависимыми величинами, что приводит к максимальной эффективности метода Белсли.

Таблица 5.2: Описание выборок для анализа метода задания порядка методом Белсли

Выборка	Тип задачи	Размер выборки	Число признаков
Wine	классификация	178	13
Boston Housing	регрессия	506	13
Synthetic data	регрессия	10000	100

Для алгоритмов тренировочная и тестовая выборки составили 80% и 20% соответственно. Критерием качества прореживания служит процент параметров нейросети, удаление которого не влечет значимой потери качества прогноза. Также критерием качества служит устойчивость нейросети к зашумленности данных.

Качеством прогноза R_{cl} модели для задачи классификации является точность прогноза модели:

$$R_{\text{cl}} = \frac{\sum_{(\mathbf{x}, y) \in \mathfrak{D}} [f(\mathbf{x}, \mathbf{w}) = y]}{|\mathfrak{D}|},$$

Качеством прогноза R_{rg} модели для задачи регрессии является среднеквадратическое отклонение результата модели от точного:

$$R_{\text{rg}} = \frac{\sum_{(\mathbf{x}, y) \in \mathfrak{D}} (f(\mathbf{x}, \mathbf{w}) - y)^2}{|\mathfrak{D}|},$$

Wine. Рассмотрим нейронную сеть с 13 нейронами на входе, 13 нейронами в скрытом слое и 3 нейронами на выходе.

На рис. 5.2 показано как меняется точность прогноза R_{cl} при удалении параметров указанными методами. Из графика видно, что метод оптимального прореживания, вариационный метод и метод Белсли позволяют удалить $\approx 80\%$

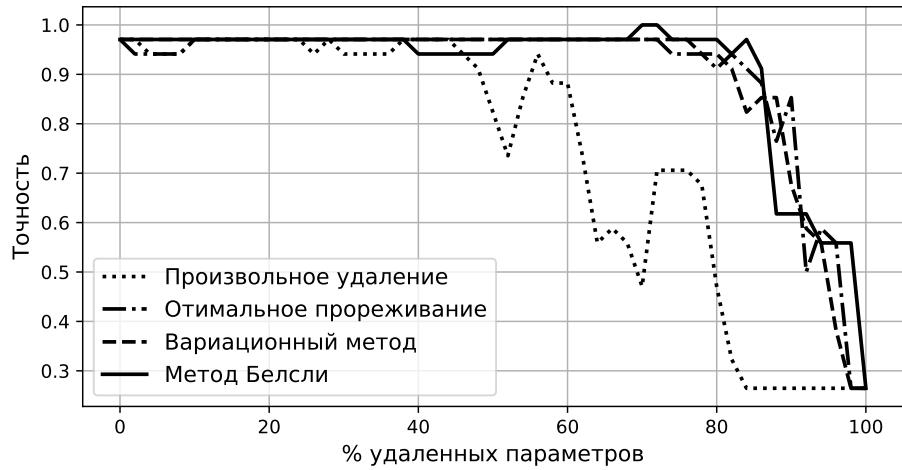


Рис. 5.2: Качество прогноза при удаление параметров на выборке Wine

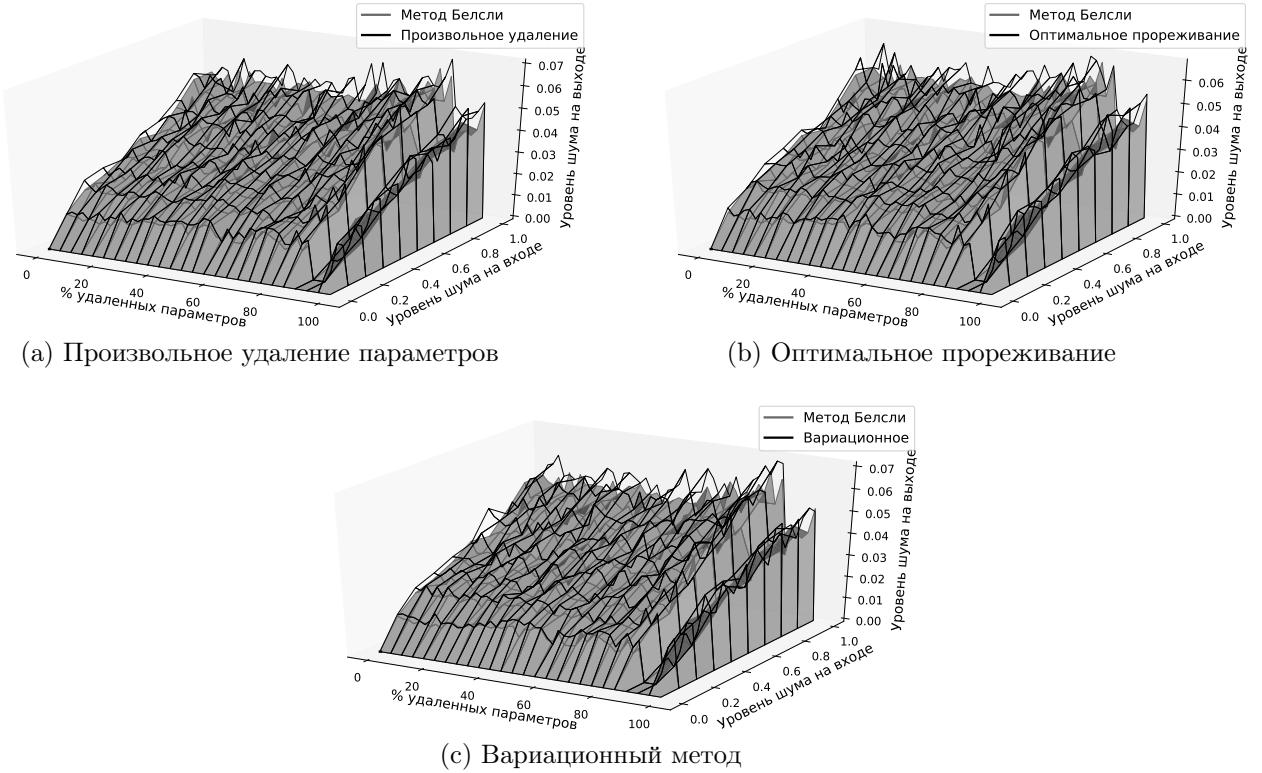


Рис. 5.3: Влияние шума в начальных данных на шум выхода нейросети на выборке Wine

параметров и качество всех этих методов падает при удалении $\approx 90\%$ параметров нейросети.

На рис. 5.3 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. На графиках показано, что при удалении параметров нейросети методом Белсли шум меньше, чем при удалении

параметров другими методами, на это указывает то что поверхность которая соответствует методу Белсли ниже других поверхностей.

Boston Housing. Рассмотрим нейронную сеть с 13 нейронами на входе, 39 нейронами в скрытом слое и одним нейроном на выходе.

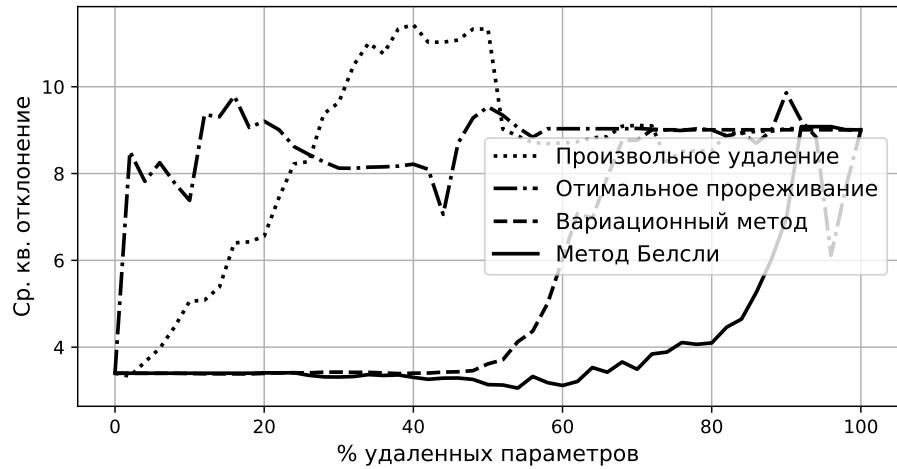


Рис. 5.4: Качество прогноза при удаление параметров на выборке Boston

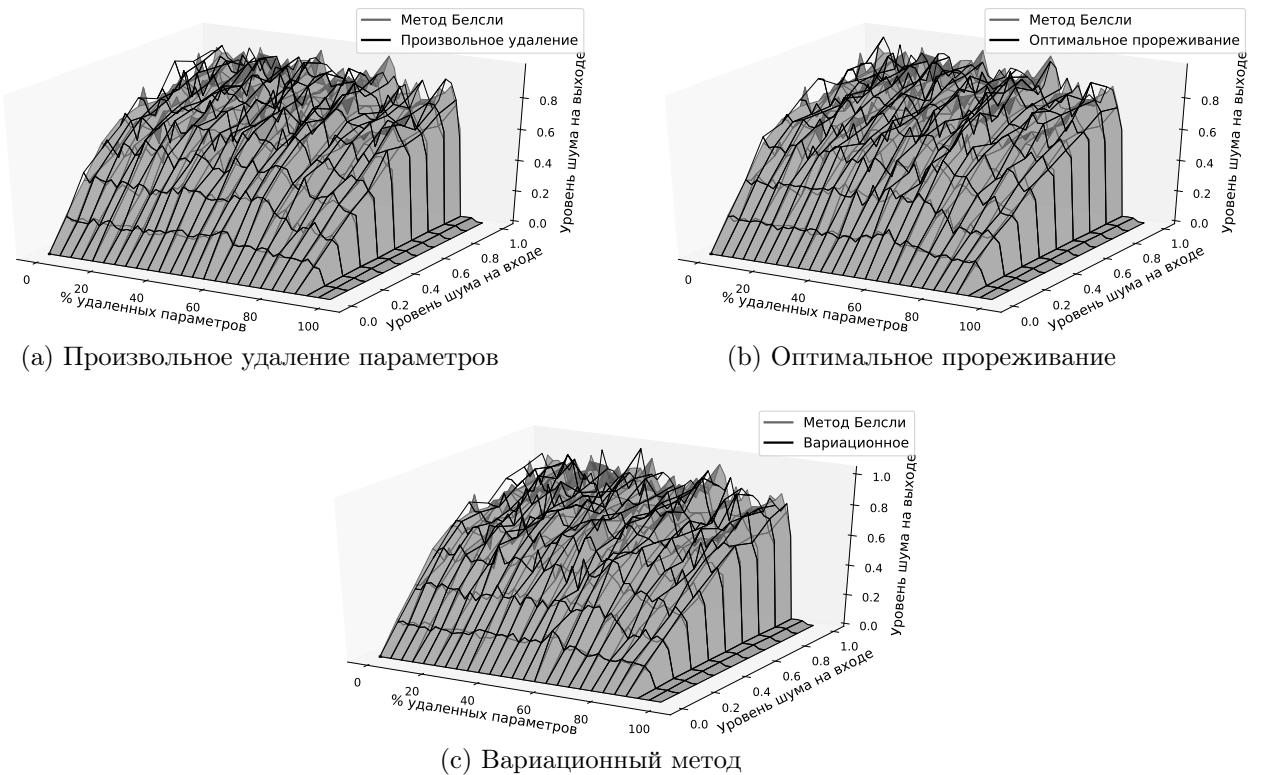


Рис. 5.5: Влияние шума в начальных данных на шум выхода нейросети на выборке Boston

На рис. 5.4 показано как меняется среднеквадратическое отклонение прогноза R_{rg} от точного ответа при удалении параметров указанными методами.

График показывает, что метод Белсли является более эффективным, чем другие методы, так как позволяет удалить больше параметров нейросети без потери качества.

На рис. 5.5 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. График показывает, что уровень шума всех методов одинаковый, так как поверхности всех методов находятся на одном уровне.

Синтетические данные. Рассмотрим нейронную сеть с 100 нейронами на входе и одним нейроном на выходе.

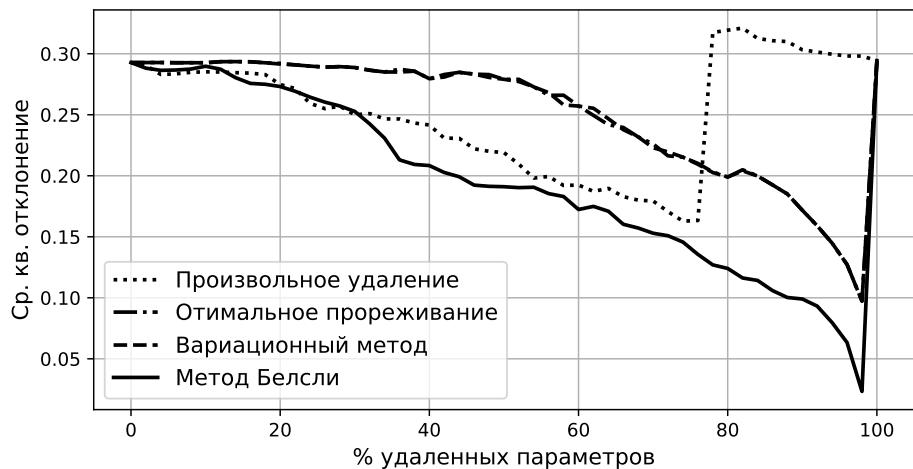


Рис. 5.6: Качество прогноза при удаление параметров на синтетической выборке

На рис. 5.6 показано как меняется среднеквадратическое отклонение прогноза от R_{rg} точного ответа при удалении параметров указанными методами. График показывает, что удаление параметров методом Белсли является более эффективным чем другие методы прореживания, так как качество прогноза нейросети повышается при удалении шумовых параметров.

На рис. 5.7 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. На графиках показано, что при удалении параметров нейросети методом Белсли шум меньше, чем при удалении параметров другими методами, так как поверхность которая соответствует методу Белсли ниже других поверхностей.

5.4. Вычислительный эксперимент по упорядочиванию параметров

Для анализа результатов, полученных предложенным алгоритмом, проводится вычислительный эксперимент. В качестве данных используются синтетические и реальные данные, которые описаны в табл. 5.3. Выборки MNIST [15] и Boston

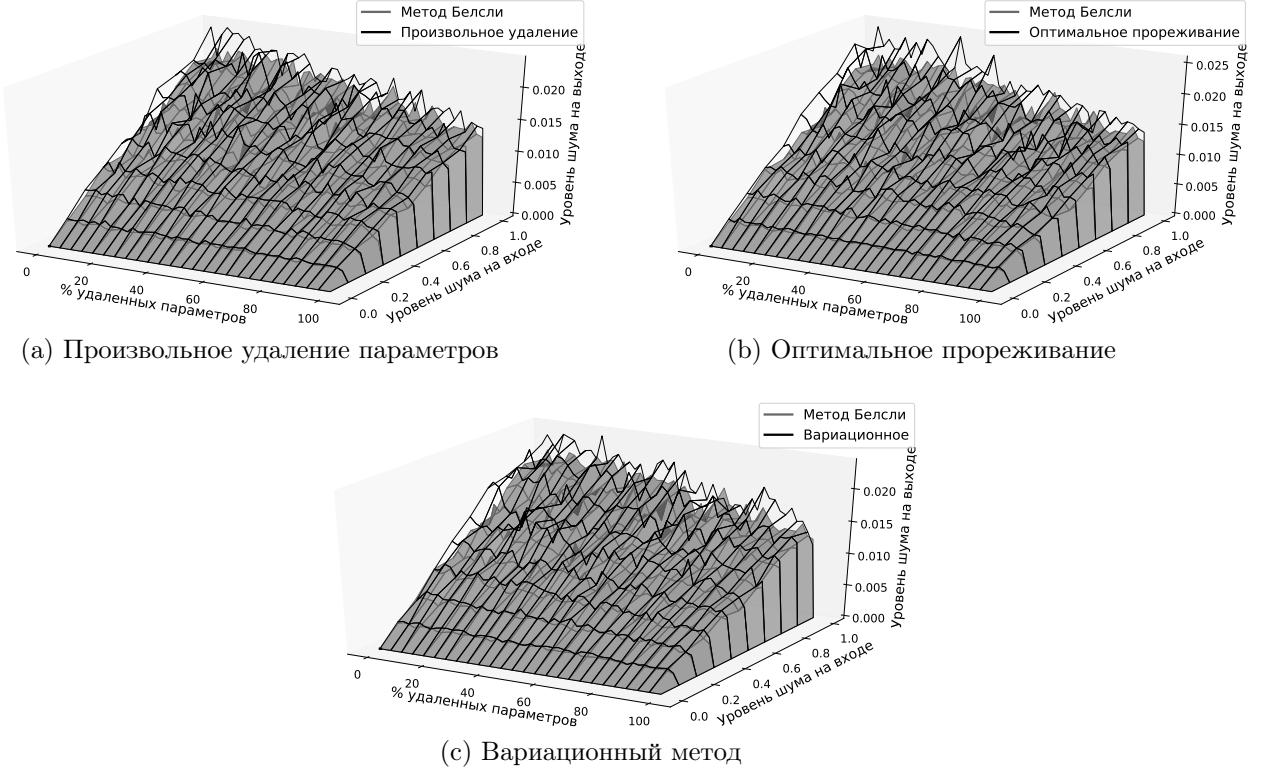


Рис. 5.7: Влияние шума в начальных данных на шум выхода нейросети на синтетической выборке

Таблица 5.3: Описание выборок, используемых в эксперименте по анализу метода задания порядка на основе анализа ковариационной матрицы градиентов

Выборка, \mathfrak{D}	Тип	Число признаков, n	Модель	Число параметров, p
Boston Housing	Регрессия	13	Нейросеть	301
MNIST	Классификация	784	Нейросеть	7960
Synthetic 3	Регрессия	200	Линейная	200
Synthetic 2	Классификация	200	Линейная	200
Synthetic 1	Регрессия	200	Нейросеть	4041

Housing [68] рассматриваются в качестве реальных данных, для которых решается задача классификации и регрессии соответственно. Алгоритм генерации синтетической выборки:

$$\begin{aligned} \mathfrak{D}_{\text{reg}} &= \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n), y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2), \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)\}, \\ \mathfrak{D}_{\text{cl}} &= \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n), y_i \sim \text{Be}(\mathbf{w}^\top \mathbf{x}_i), \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)\}. \end{aligned}$$

В качестве аппроксимирующих моделей рассматриваются линейные и нейросетевые модели (5.1). В качестве функции ошибки для задачи регрессии рассматривается `MSELoss`, а для задачи классификации — `CrossEntropyLoss` (5.2).

Предварительно для каждой модели и выборки определяется число t_0 — номер итерации, после которой все параметры модели находятся в некоторой окрестности локального минимума. Параметр t_0 устанавливается экспериментальным путем для каждой модели и выборки отдельно из условия, что качество модели меняется незначительно при числе итераций $t > t_0$.

После t_0 шагов алгоритма оптимизации часть параметров модели фиксируется в соответствии с формулами (5.4), (5.5). Результат получается усреднением по 25 независимым запускам оптимизации модели. Значение функции ошибки \mathcal{L} усредняется по разным запускам алгоритма оптимизации. В ходе эксперимента проводится анализ вектора α , который также усредняется по разным запускам алгоритма оптимизации. Усредненное значение бинарного вектора α обозначим $\hat{\alpha}$.

Выборка Synthetic 1.

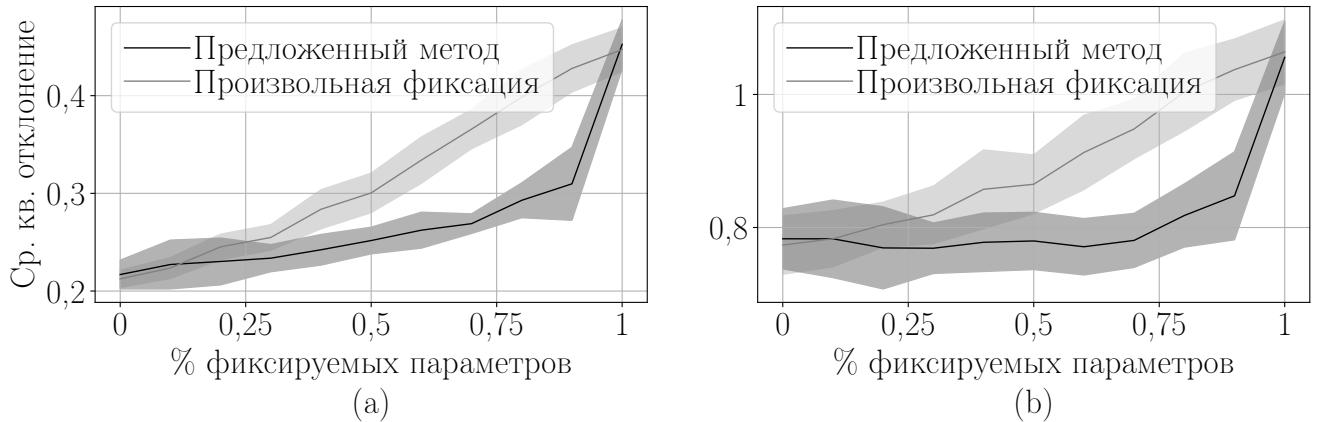


Рис. 5.8: Зависимость качества модели от числа зафиксированных параметров:
а) на обучающей выборке; б) на тестовой выборке

Эксперимент проводился на синтетически построенных данных. В качестве модели использовалась двухслойная нейросеть — перцентрон. На рис. 5.8 показаны графики зависимости функции потерь \mathcal{L} от числа фиксируемых параметров. В случае фиксации параметров предложенным методом функция потерь \mathcal{L} растет медленней, чем в случае фиксации параметров произвольным образом.

На рис. 5.9 показана зависимость векторов $\hat{\alpha}(\zeta)$ от числа фиксируемых параметров. Каждый столбец соответствует одному вектору $\hat{\alpha}(\zeta)$. На рис. 5.9a, 5.9c видно, что $\hat{\alpha}(\zeta)$ имеет большое число компонент вектора, близких к 1. Так как $\hat{\alpha}(\zeta)$ является усреднением вектора с компонентами 0 или 1, то предложенный порядок задает некоторый устойчивый порядок на множестве параметров модели. На рис. 5.9b, 5.9d видно, что в случае произвольной фиксации параметров компоненты вектора $\hat{\alpha}(\zeta)$ имеют одинаковые значения, следовательно, никакого порядка на множестве параметров нет.

Выборка Boston Housing. Эксперимент проводился на реальных данных. На рис. 5.10 показаны графики зависимости функции потерь \mathcal{L} от числа фик-

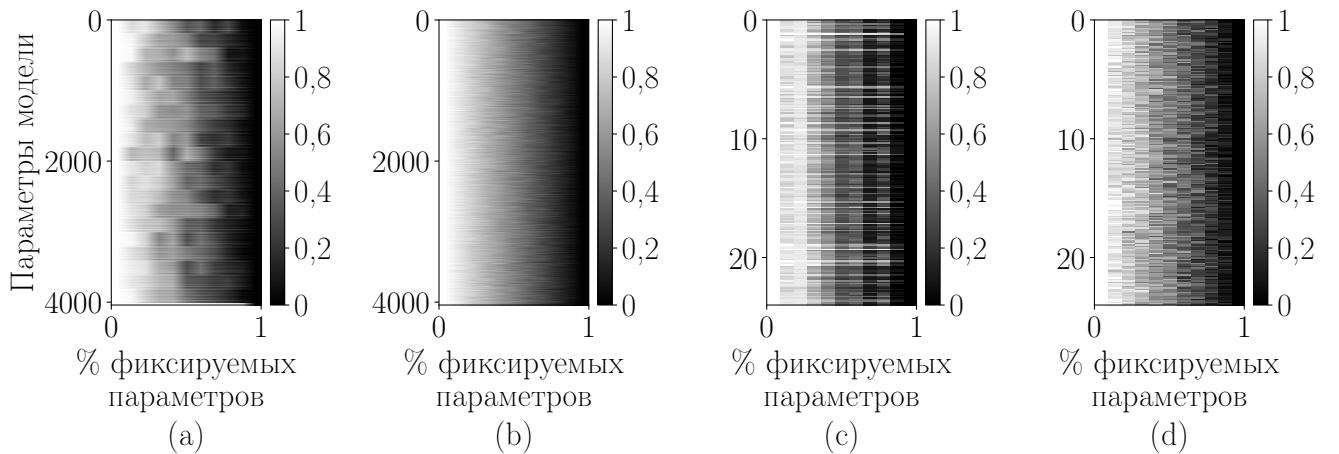


Рис. 5.9: Визуализация векторов $\hat{\alpha}(\zeta)$ в зависимости от числа фиксируемых параметров: а) все параметры модели упорядочены предложенным методом; б) все параметры модели упорядочены произвольным образом; в) часть параметров модели упорядочена предложенным методом; г) часть параметров модели упорядочена произвольным образом

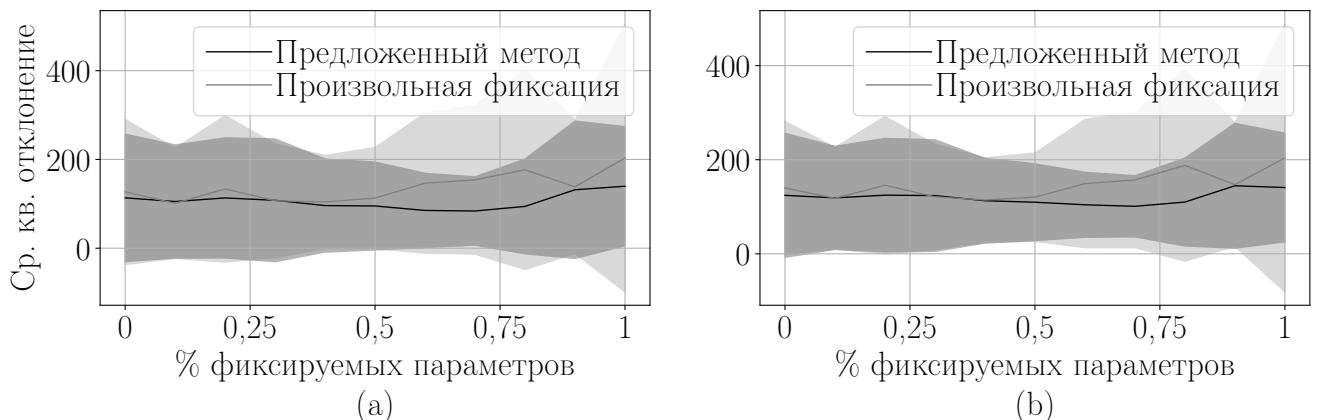


Рис. 5.10: Зависимость качества модели от числа зафиксированных параметров: а) на обучающей выборке; б) на тестовой выборке

сируемых параметров. В случае фиксации параметров предложенным методом, функция потерь \mathcal{L} растет так же, как и в случае фиксации параметров произвольным образом. Данный результат следует из того, что нейросеть оказалась избыточно сложной моделью с большим числом параметров. После фиксации значимого числа параметров у модели оставалась большое число параметров для дообучения.

На рис. 5.11 показана зависимость векторов $\hat{\alpha}(\zeta)$ от числа фиксируемых параметров. На рис. 5.11а, 5.11с видно, что $\hat{\alpha}(\zeta)$ меняется незначительно от запуска к запуску алгоритма. Следовательно, предложенный порядок задает устойчивый к разным запускам порядок на множестве параметров модели. На рис. 5.11б, 5.11д видно, что в случае произвольной фиксации параметров вектор

$\hat{\alpha}(\zeta)$ является произвольным и никакого порядка на множестве параметров нет.

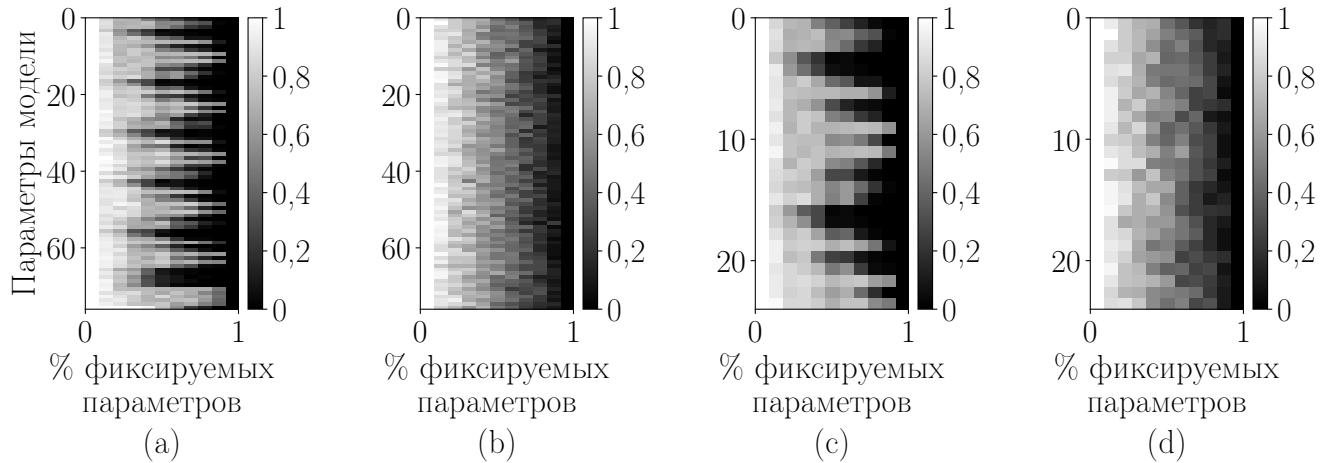


Рис. 5.11: Визуализация векторов $\hat{\alpha}(\zeta)$ в зависимости от числа фиксируемых параметров: а) все параметры модели упорядочены предложенным методом; б) все параметры модели упорядочены произвольным образом; в) часть параметров модели, упорядочена предложенным методом; г) часть параметров модели упорядочена произвольным образом

Выборка Synthetic 3.

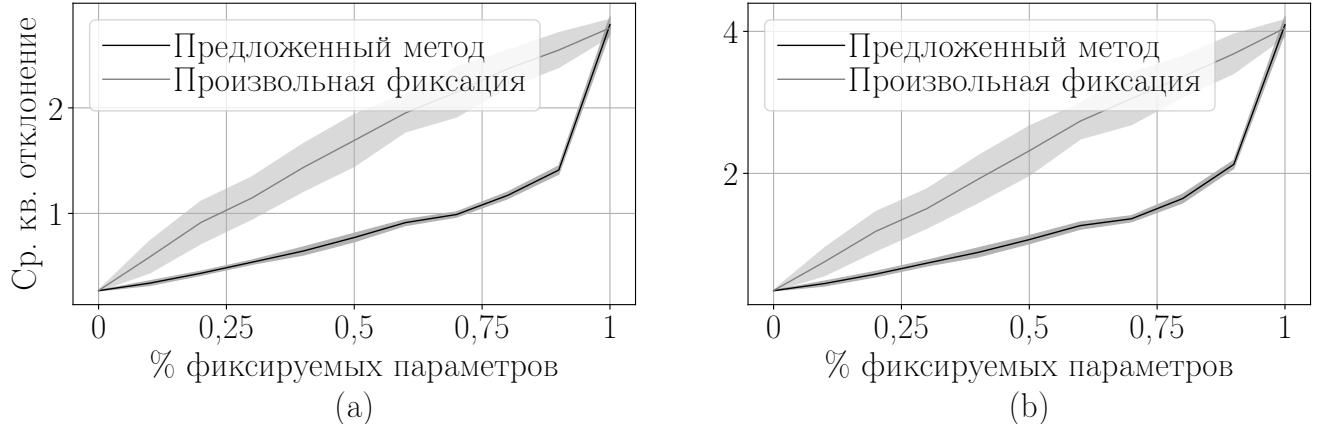


Рис. 5.12: Зависимость качества модели от числа зафиксированных параметров: а) на обучающей выборке; б) на тестовой выборке

Эксперимент проводился на синтетически построенных данных. В качестве модели использовалась линейная модель регрессии.

На рис. 5.12 показаны графики зависимости функции потерь \mathcal{L} от числа фиксируемых параметров. В случае фиксации параметров предложенным методом функция потерь \mathcal{L} растет значительно медленней в сравнении со случаем фиксации параметров произвольным образом. Дисперсия функции ошибки также значительно меньше в случае фиксации параметров предложенным методом.

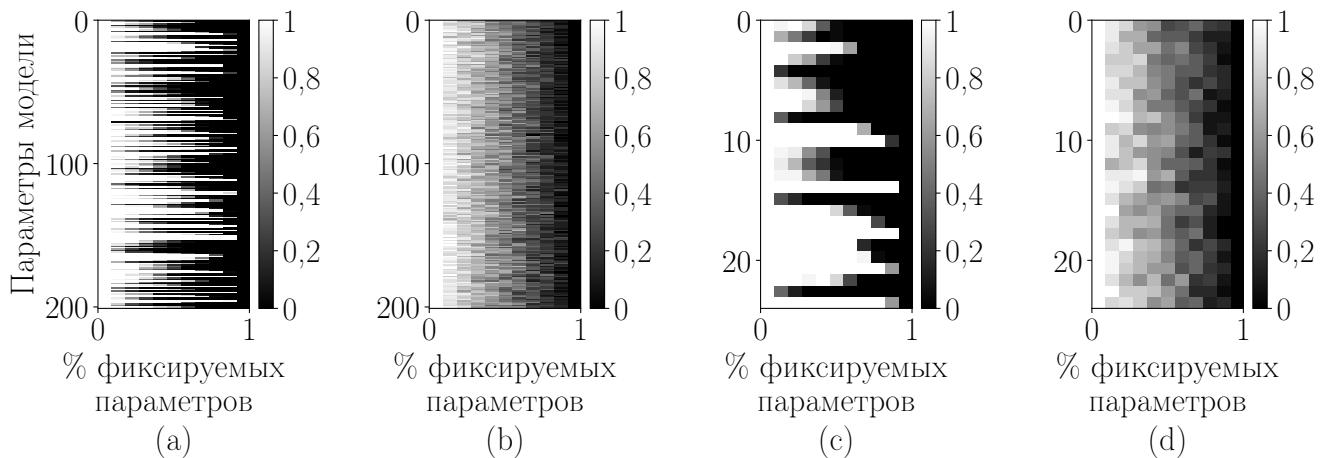


Рис. 5.13: Визуализация векторов $\hat{\alpha}(\zeta)$ в зависимости от числа фиксируемых параметров: а) все параметры модели упорядочены предложенным методом; б) все параметры модели упорядочены произвольным образом; в) часть параметров модели, упорядочены предложенным методом; г) часть параметров модели упорядочена произвольным образом

На рис. 5.13 показано, что вектора $\hat{\alpha}(\zeta)$ не меняются от запуска к запуску. Так как данная модель линейная, то порядок на параметрах модели индуцирует некоторый порядок на множестве признаков.

Выборка MNIST.

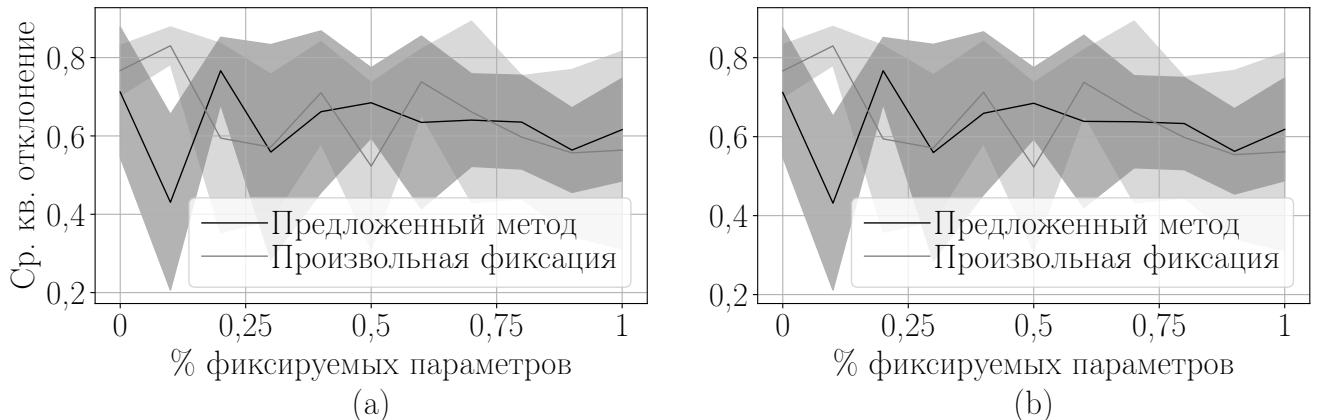


Рис. 5.14: Зависимость качества модели от числа зафиксированных параметров: а) на обучающей выборке; б) на тестовой выборке

В эксперименте рассматривался двухслойный перцептрон для классификации изображений. В качестве входных данных рассматривались изображения размера 28×28 , на которых изображены цифры.

На рис. 5.14 показано, что графики функции ошибки похожи в случае фиксации параметров предложенным методом и в случае произвольной фиксации. Данный результат есть следствие того факта, что нейросеть является заведомо переусложненной моделью с большим числом параметров. После

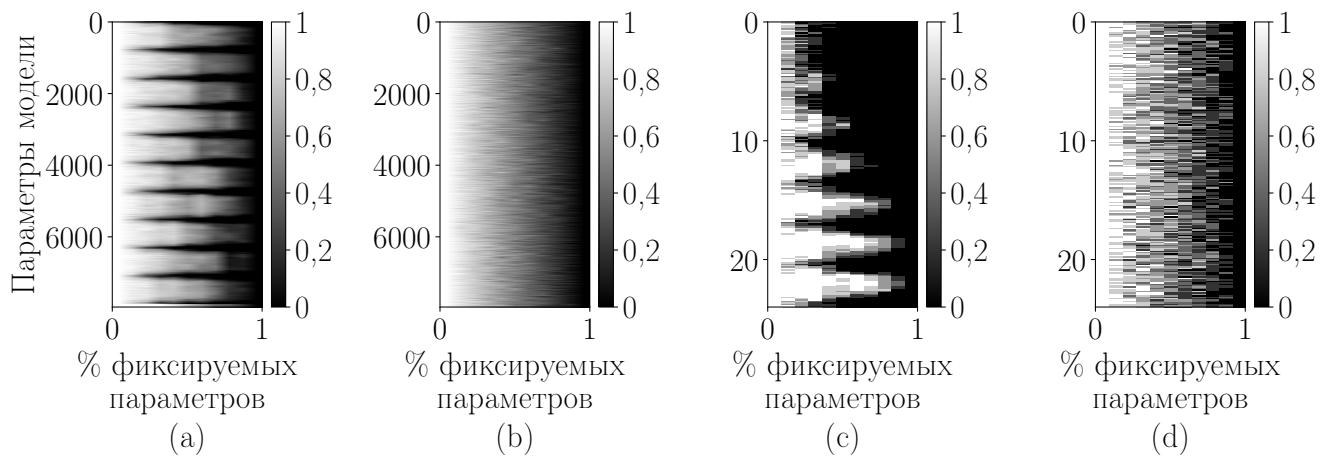


Рис. 5.15: Визуализация векторов $\hat{\alpha}(\zeta)$ в зависимости от числа фиксируемых параметров: а) все параметры модели упорядочены предложенным методом; б) все параметры модели упорядочены произвольным образом; в) часть параметров модели упорядочена предложенным методом; г) часть параметров модели упорядочена произвольным образом

фиксации большого числа параметров у нейросети все еще остается значимое число параметров модели для дообучения.

На рис. 5.15 показано, что в случае модели со значимым числом оптимизационных параметров, предложенный метод упорядочения параметров устойчив от запуску к запуску.

Глава 6

Анализ прикладных задач выбора моделей машинного обучения

Планирование эксперимента требует оценки минимального размера выборки: числа выполненных измерений набора характеристик, необходимых для построения сформулированных условий. Выбор метода оценки размера выборки зависит от решаемой задачи, которая определяет формулировку статистической гипотезы и статистики для ее проверки. В таблице 6.1 представлены десять методов оценки размера выборки. Они включают как статистические, так и байесовский методы оценки размера выборки.

Статистические методы предполагают, что выборка соответствует некоторым предварительным условиям, сформулированным ранее. Эти условия сформулированы как статистический критерий [69, 70, 71, 72]. Метод оценки размера выборки, связанный с этим критерием, гарантирует достижение фиксированной статистической мощности $1 - \beta$ со степенью ошибки первого рода, не превышающей установленное значение α . Такой размер выборки называется достаточным.

Однако практическое применение методов оценки размера выборки предполагает, что модель соответствует измеренным данным [79]. Эти модели выбираются в соответствии с постановкой задачи регрессии или классификации. В этой статье представлены обобщенные линейные модели. В статье [70] предложен подход к оценке мощности и размера выборки, связанной с ней, на основе теста отношения максимального правдоподобия. Этот подход оказался более точным для ряда независимых переменных. Кроме того, в статье [73] предложен метод оценки мощности для статистики Вальда. В статье [74] в случае логистической регрессии предлагается использовать метод, использующий кривую ROC-AUC и концепцию сдвига. Классические методы [69, 70, 71, 73, 72] имеют ряд ограничений, связанных с практическим применением этих методов. Чтобы оценить размер выборки, необходимо знать дисперсию оценки параметра или, в более общем случае, иметь оценку параметра нецентральности в распределении статистики, используемой, когда альтернативная гипотеза верна. Эти методы не показывают как получить эти значения. Кроме того, дисперсия оценки и параметр нецентральности не будут получены с определенной дисперсией, влияние которой на результат оценки размера выборки не имеет значения.

Статистические методы позволяют оценить размер выборки на основе предположений о распределении данных и информации о соответствии между наблюдаемыми значениями и предположениями нулевой гипотезы. Когда размер исследуемой выборки является достаточным или чрезмерным, можно использовать методы, основанные на наблюдении изменения определенной характеристики процедуры построения модели при увеличении размера выборки. В частности, наблюдая за соотношением качества прогнозирования с контрольной выборкой и обучающей выборкой [74], определяется достаточный размер выборки, который соответствует началу переобучения. В статье [75] для оцен-

Таблица 6.1: Сводная таблица методов определения оптимального размера выборка для линейных моделей

Метод	Описание	Ссылка
Lagrange multipliers test	Правдоподобие выборки имеет вид: $p(y \mathbf{x}, \mathbf{w}) = \exp(y\theta - b(\theta) + c(y))$. Достаточный размер выборки m^* : $m^* = \frac{\gamma^*}{\gamma_0}$, где γ^* и γ_0 задаются выражениям (6.3) и (6.2).	[69]
Likelihood ratio test	Правдоподобие выборки имеет вид: $p(y \mathbf{x}, \mathbf{w}) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$. Достаточный размер выборки m^* : $m^* = \frac{\gamma^*}{\Delta^*}$, где γ^* и Δ^* задаются выражениями (6.5) и (6.6).	[71]
Wald statistic	Правдоподобие выборки имеет вид: $p(y \mathbf{x}, \mathbf{w}) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$. Достаточный размер выборки m^* : $m^* = \frac{\gamma^*}{\delta}$, где γ^* и δ задаются выражением (6.8).	[73]
Cross-validation	Достаточный размер выборки m^* : $\forall m \geq m^* RS(m) \geq 1 - \varepsilon$, где ε задается экспериментально, RS определяется выражением (6.9).	[74]
Bootstrap	Достаточный размер выборки m^* : $\forall m \geq m^* \max_i (b_i^m - a_i^m) < l$, где (a_i^m, b_i^m) является квантилью доверительного интервала i -й бутстррап подвыборки размера m	[75]
Kullback-Leibler	Достаточный размер выборки m^* : $\forall \mathfrak{D}_{\mathcal{B}_1} : \mathfrak{D}_{\mathcal{B}_1} \geq m^* \mathbb{E}_{\mathfrak{D}_{\mathcal{B}_2}} D_{\text{KL}}(p_1, p_2) \leq \varepsilon$, где $\mathcal{B}_1, \mathcal{B}_2$ удовлетворяет (6.13).	[74]
Average posterior variance criterion	Достаточный размер выборки m^* : $\forall m \geq m^* \mathbb{E}_{\mathfrak{D}_m} D[\hat{\mathbf{w}} \mathfrak{D}_m] \leq \varepsilon$, где ε достаточно малое значение.	[76, 77]
Average coverage criterion	Достаточный размер выборки m^* : $\forall m \geq m^* \mathbb{E}_{\mathfrak{D}_m} \mathbb{P}\{\mathbf{w} \in A(\mathfrak{D}_m)\} \geq 1 - \alpha$, где α достаточно малое значение.	[76, 77]
Average length criterion	Достаточный размер выборки m^* : $\forall m \geq m^* \mathbb{E}_{\mathfrak{D}_m} r_m \leq l$, where r_m описанного в (6.11)	[76, 77]
Utility maximization	Достаточный размер выборки m^* : $m^* = \arg \max_m \mathbb{E}_{\mathfrak{D}_m} \int_{\mathbf{w}} u(\mathfrak{D}, \mathbf{w}) p(\mathbf{w} \mathfrak{D}) d\mathbf{w}$, где функция полезности $u(\mathfrak{D}, \mathbf{w})$ задается выражением (6.12).	[78]

ки достаточного размера выборки используется процедуру бутстррап. Превышение текущего размера выборки проверяется на основе анализа доверительных интервалов оцениваемого параметра. Ширина доверительного интервала с различными значениями объема выборки оценивается с помощью метода бутстрата. Для этого выборки меньшего размера отбираются заданное число раз и вычисляется доверительный интервал ошибки при оценке параметра модели. Размер выборки считается достаточным, если ширина доверительного интервала не превышает заранее установленного значения.

Перечисленные выше ограничения статистических методов оценки размера выборки подробно исследуются в байесовской процедуре [78, 80, 81], где оценка размера выборки определяется на основе максимизации ожидаемого значение некоторой функции качества [78]. Функция качества может включать в себя явные функции распределения параметров и штрафы за увеличение размера выборки. Альтернативой подходам [81], основанным на функции качества, является выборка размера выборки путем установления ограничений на определенный критерий качества оценки параметров модели. Примеры критериев: критерий средней апостериорной дисперсии (AVPC), критерий средней длины (ALC), критерий среднего покрытия (ACC). Для каждого перечисленного критерия оценка размера выборки определяется как минимальное значение размера выборки, для которого ожидаемое значение выбранного критерия не превышает какого-либо фиксированного порога. В статье [74] предлагается считать размер выборки достаточным, если расстояние Кульбака-Лейблера между распределениями, оцененными на основе подвыборок такого размера, достаточно мало. Такой подход не требует дальнейшего обобщения в случае нескольких переменных. Кроме того, оценка может производиться как при наличии предположений о распределении данных, так и при их отсутствии. Недостаток этого подхода заключается в том, что количественная оценка может быть получена только при чрезмерно большом размере выборки.

6.1. Постановка задачи определения достаточного размера выборки

Задана выборка размера m :

$$\mathfrak{D}_m = \{\mathbf{x}_i, y_i\}_{i=1}^m,$$

где $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{Y}$. Вектор признаков $\mathbf{x} = [\mathbf{u}, \mathbf{v}]$ соединяет $\mathbf{u}_i \in \mathbb{R}^k$ and $\mathbf{v}_i \in \mathbb{R}^{n-k}$. Выборка \mathfrak{D}_m случайным образом делится на обучающую и тестовую части:

$$\mathfrak{D}_{\mathcal{T}_m} = \{\mathbf{x}_i, y_i\}_{i \in \mathcal{T}_m}, \quad \mathfrak{D}_{\mathcal{L}_m} = \{\mathbf{x}_i, y_i\}_{i \in \mathcal{L}_m}, \quad \mathcal{T}_m \sqcup \mathcal{L}_m = \{1, \dots, m\}.$$

Введем параметрическое семейство функций для аппроксимации неизвестного распределения $p(y|\mathbf{x}, \mathfrak{D}_{\mathcal{L}_m})$:

$$\mathfrak{F} = \left\{ f(y, \mathbf{x}, \mathbf{w}) \mid \mathbf{w} \in \mathbb{W}, \int_{y \in \mathbb{Y}, \mathbf{x} \in \mathbb{R}^n} f(y, \mathbf{x}, \mathbf{w}) dy d\mathbf{x} = 1 \right\}.$$

Для модели f с вектором параметров \mathbf{w} определим функцию правдоподобия и логарифм функции правдоподобия выборки \mathfrak{D} :

$$L(\mathfrak{D}, \mathbf{w}) = \prod f(y, \mathbf{x}, \mathbf{w}), \quad l(\mathfrak{D}, \mathbf{w}) = \sum \log f(y, \mathbf{x}, \mathbf{w}),$$

где $f(y, \mathbf{x}, \mathbf{w})$ является оценкой правдоподобия выборки $\mathfrak{D}_{\mathcal{L}}$ с заданным вектором параметров \mathbf{w} . Используя принцип максимального правдоподобия для оценки параметров \mathbf{w}

$$\hat{\mathbf{w}} = \arg \max_{\mathbf{w} \in \mathbb{W}} L(\mathfrak{D}_{\mathcal{L}}, \mathbf{w}).$$

Информационная матрица Фишера имеет вид:

$$\mathbf{I}(\mathfrak{D}, \mathbf{w}) = -\nabla \nabla^T l(\mathfrak{D}, \mathbf{w}), \quad \mathbf{V} = \mathbf{I}^{-1}(\mathfrak{D}, \mathbf{m}), \quad (6.1)$$

статистические методы и байесовские методы используют информационную матрицу Фишера для оценки размера выборки.

6.2. Байесовский подход к определению достаточного размера выборки

Статистические методы определения достаточного размера выборки. Основным преимуществом методов, основанных на статистике, является их способность оценивать достаточный размер выборки при недостаточном наборе выборки. Они позволяют прогнозировать необходимое число образцов на ранней стадии эксперимента.

Плотность распределения целевой переменной

$$p(y|\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v) = \exp(y\theta - b(\theta) + c(y)),$$

где θ является параметром распределения, полученный с помощью функции связи $\theta = \theta(\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v)$.

Тестируемая гипотеза

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad H_1 : \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Пусть статистики $S_{m,u}(\mathbf{w}_u, \mathbf{w}_v)$ и $S_{m,v}(\mathbf{w}_u, \mathbf{w}_v)$ являются производными логарифма правдоподобия выборки \mathfrak{D}_m в точках \mathbf{w}_u и \mathbf{w}_v . Рассмотрим $\mathbf{s}_m = S_{m,u}(\mathbf{m}_u^0, \hat{\mathbf{w}}_v^0)$, где $\hat{\mathbf{w}}_v^0$ получается из уравнения

$$S_{m,v}(\mathbf{m}_u^0, \mathbf{w}_v) = 0.$$

Статистика Лагранджа равняется

$$LM = \mathbf{s}_m^T \mathbf{Q}_m^{-1} \mathbf{s}_m.$$

где \mathbf{Q}_m ковариационная матрица вектора \mathbf{s}_m .

В случае истинности гипотезы H_0 статистика LM асимптотически имеет распределения $\chi^2(k)$. В [69] показано, что при альтернативной гипотезе H_1 статистика LM асимптотически имеет распределения $\chi^2(k, \gamma)$, где γ является параметром нецентральности

$$\gamma = \boldsymbol{\xi}_m^\top \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\xi}_m = m \boldsymbol{\xi}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi} = m\gamma^0, \quad (6.2)$$

где $\boldsymbol{\xi}_m$ и $\boldsymbol{\Sigma}_m$ матрицы математического ожидания и ковариации \mathbf{s}_m . Обозначим $\boldsymbol{\xi}_1 = \boldsymbol{\xi}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$.

Альтернативный метод получения γ включает условия на уровне значимости α и вероятность ошибки II рода β :

$$\gamma^* : \chi_{k,1-\alpha}^2 = \chi_{k,\beta}^2(\gamma). \quad (6.3)$$

Используя (6.2) и (6.3) получаем

$$m^* = \frac{\gamma^*}{\gamma^0}.$$

Это достаточный минимальный размер выборки, чтобы различить вектор \mathbf{m}_u от \mathbf{m}_u^0 .

Пусть правдоподобие выборки задается выражением

$$p(y|\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v) = \exp \left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right), \quad (6.4)$$

где θ является параметром распределения, который вычисляются с помощью функции связи $\theta = \theta(\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v)$.

Проверяемая гипотеза

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad H_1 : \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Введем логарифм статистики отношения правдоподобий:

$$LR = 2 \left(l(\mathfrak{D}, \hat{\mathbf{w}}) - l(\mathfrak{D}, \hat{\mathbf{w}}^0) \right),$$

где $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_u, \hat{\mathbf{w}}_v]$ является вектором, который максимизирует правдоподобие (6.4), $\hat{\mathbf{w}}^0 = [\mathbf{m}_u^0, \hat{\mathbf{w}}_v^0]$ является вектором, который максимизирует правдоподобие (6.4) с фиксируемым подвектором параметров \mathbf{m}_u^0 .

В случае истинности гипотезы H_0 статистика LR асимптотически имеет распределения $\chi^2(k)$. В [71] показано, что при альтернативной гипотезе H_1 статистика LR асимптотически имеет распределения $\chi^2(k, \gamma)$, где γ является параметром нецентральности

$$\gamma = m\Delta^*, \quad \Delta^* = \mathbb{E} [2a^{-1}(\phi) \{ (\theta - \theta^*) \nabla b(\theta) - b(\theta) + b(\theta^*) \}], \quad (6.5)$$

где параметры θ и θ^* рассчитываются с использованием параметров $\mathbf{w} = [\mathbf{w}_u, \mathbf{w}_v]$ и $\mathbf{w}^* = [\mathbf{w}_u^0, \mathbf{w}_v^*]$. Параметры \mathbf{w}_v^* вычисляются на основе решения уравнения

$$\lim_{m \rightarrow \infty} m^{-1} \mathsf{E} \left(\frac{\partial l(\mathfrak{D}, [\mathbf{m}_u^0, \mathbf{w}_v])}{\partial \mathbf{w}_v} \right) = 0.$$

Тогда с учетом α и β достаточный размер выборки m^* вычисляется

$$m^* = \frac{\gamma^*}{\Delta^*}, \quad \gamma^* : \chi_{k,1-\alpha}^2 = \chi_{k,\beta}^2(\gamma), \quad (6.6)$$

где $\chi_{k,1-\alpha}^2$, $\chi_{k,\beta}^2(\gamma^*)$ квантили распределений χ_k^2 and $\chi_k^2(\gamma^*)$. Правдоподобие выборки:

$$p(y|\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v) = \exp \left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right), \quad (6.7)$$

где θ является параметром распределения, который вычисляются с помощью функции связи $\theta = \theta(\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v)$.

Тестируемая гипотеза:

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad H_1 : \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Тест Вальда для гипотезы:

$$W = (\hat{\mathbf{w}}_u - \mathbf{m}_u^0)^T \hat{\mathbf{V}}_u^{-1} (\hat{\mathbf{w}}_u - \mathbf{m}_u^0),$$

где $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_u, \hat{\mathbf{w}}_v]$ вектор параметров, который максимизирует правдоподобие выборки (6.7), где матрица $\hat{\mathbf{V}}_u$ задается в выражении (6.1).

В случае истинности гипотезы H_0 статистика Вальда W асимптотически имеет распределение χ^2 . В [73] показано, что в случае истинности альтернативной гипотезы H_1 статистика Вальда W асимптотически имеет распределение $\chi^2(k, \gamma)$ с параметром нецентральности γ :

$$\gamma = m\delta, \quad \delta = (\hat{\mathbf{w}}_u - \mathbf{m}_u^0)^T \Sigma_u^{-1} (\hat{\mathbf{w}}_u - \mathbf{m}_u^0), \quad \Sigma_u = m\hat{\mathbf{V}}_u.$$

Использую заданный уровень значимости α и заданную ошибку второго рода β определим оптимальный размер выборки:

$$m^* = \frac{\gamma^*}{\delta}, \quad \gamma^* : \chi_{k,1-\alpha^*}^2 = \chi_{k,\beta}^2(\gamma),$$

где $\chi_{k,1-\alpha^*}^2$, $\chi_{k,\beta}^2(\gamma^*)$ квантили распределения, а параметр α^* это поправка на уровень значимости:

$$\alpha^* = P(\boldsymbol{\xi}^T \Sigma^{*-1} \boldsymbol{\xi} > \chi_{k,1-\alpha}^2), \quad \Sigma^* = \mathbf{I}^{-1}(\mathfrak{D}, \mathbf{w}^*),$$

где $\mathbf{w}^* = [\mathbf{m}_u^0, \mathbf{w}_v^*]$ является решением уравнения

$$\lim_{m \rightarrow \infty} m^{-1} \mathsf{E} \left(\frac{\partial l(\mathfrak{D}, [\mathbf{m}_u^0, \mathbf{w}_v])}{\partial \mathbf{w}_v} \right) = 0. \quad (6.8)$$

Эвристические методы определения достаточного размера. В методе, основанном на эвристике, используются популярные статистические эвристики, такие как бутстрэп, перекрестная проверка и задание функции полезности. Введем набор индексов \mathcal{A} для параметров логистической регрессии \mathbf{w} . Тестируется гипотеза

$$H_0 : j \notin \mathcal{A} (\mathbf{w}_j = 0), \quad H_1 : j \in \mathcal{A}^* (\mathbf{w}_j \neq 0),$$

где \mathbf{w}_j является j -м элементом вектора \mathbf{w} . Установим параметр отступа c_0 для задачи логистической регрессии:

$$H_0 : 1 - c_0 = p_0, \quad H_1 : 1 - c_0 = p_1,$$

где c_0 оптимальное решение, когда исключен j -й элемент вектора. Используя статистику

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{m}, \quad \hat{p} = \frac{1}{m} \sum_{i=1}^m y_i.$$

В случае истинности нулевой гипотезы H_0 статистика Z асимптотически имеет распределение $\mathcal{N}(0, 1)$. В случае истинности альтернативной гипотезы H_1 статистика Z асимптотически имеет распределение $\mathcal{N}\left(p_1 - p_0, \sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}\right)$.

Достаточный объем выборки задается выражением

$$m^* = \frac{p_0(1 - p_0) \left(Z_{1-\alpha/2} + Z_{1-\beta} \sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}} \right)^2}{(p_1 - p_0)^2},$$

где $Z_{1-\alpha/2}$ и $Z_{1-\beta}$ являются квантилями распределения $\mathcal{N}(0, 1)$.

Данный метод не рассматривается далее, поскольку его можно использовать только в задаче логистической регрессии.

Рассмотрим метод на основе кроссвалидации. Определим критерий переобучения как

$$RS(m) = \ln \frac{L(\mathfrak{D}_{\mathcal{L}(m)}, \hat{\mathbf{w}})}{L(\mathfrak{D}_{\mathcal{T}(m)}, \hat{\mathbf{w}})}, \quad \frac{|\mathcal{T}(m)|}{|\mathcal{L}(m)|} = \text{const} \leq 0.5. \quad (6.9)$$

Заметим, что

$$\lim_{m \rightarrow \infty} RS(m) \rightarrow 0.$$

Достаточный размер выборки m^* определяется согласно условию:

$$m^* : \forall m \geq m^* \mathsf{E}_{\mathfrak{D}_m} RS(m) \leq \varepsilon,$$

где ε некоторый параметр, который задается экспертизой.

Этот метод предполагает, что длины доверительных интервалов квантиля не превышают некоторого фиксированного значения l . Для некоторого размера выборки m вычисляются квантильные доверительные интервалы $(a_1^m, b_1^m), (a_2^m, b_2^m), \dots, (a_n^m, b_n^m)$ с уровнем значимости α с использованием начальной загрузки для каждого параметра модели. Достаточный размер выборки задается выражением:

$$m^* : \forall m \geq m^* \max_i (b_i^m - a_i^m) < l.$$

Важно, что этот метод является покоординатным и следовательно для повышения точности прогноза требуется значительное увеличение размера выборки.

Байесовские методы. Байесовские методы оценки размера выборки основаны на ограничении некоторых характеристик модели. Для анализа эффективности определяется функция размера выборки. Увеличение этой функции интерпретируется как снижение эффективности модели. Размер выборки m^* выбирается таким, чтобы исследуемая функция принимала значения меньше некоторого порогового значения ε .

Average posterior variance criterion. Размер выборки m^* определяется условием:

$$\forall m \geq m^* E_{\mathfrak{D}_m} D[\hat{\mathbf{w}} | \mathfrak{D}_m] \leq l.$$

где l некоторый заданный экспертизой параметр, который количественно определяет неопределенность оценки параметра.

Average coverage criterion. Обозначим через $A(\mathfrak{D}) \subset \mathbb{R}^n$ некоторый набор параметров модели \mathbf{w} :

$$A(\mathfrak{D}) = \{\mathbf{w} : \|\mathbf{w} - \hat{\mathbf{w}}\| \leq l\},$$

где l — некоторый фиксированный радиус шара. Размер выборки m^* определяется критерием среднего покрытия:

$$\forall m \geq m^* E_{\mathfrak{D}_m} P\{\mathbf{w} \in A(\mathfrak{D}_m)\} \geq 1 - \alpha, \quad (6.10)$$

где α некоторый параметр заданный экспертизой.

Average length criterion. Определим функцию $A(\mathfrak{D})$:

$$P(A(\mathfrak{D})) = 1 - \alpha.$$

Оценки критерия средней длины m^* заданный в (6.10):

$$\forall m \geq m^* E_{\mathfrak{D}_m} r_m \leq l, \quad (6.11)$$

где r_m является радиусом шара $A(\mathfrak{D}_m)$.

Следующие методы максимизируют ожидание некоторой функции полезности $u(\mathfrak{D}, \mathbf{w})$ по размеру выборки:

$$m^* = \arg \max_m \mathbb{E}_{\mathfrak{D}_m} \int_{\mathbf{w}} u(\mathfrak{D}_m, \mathbf{w}) p(\mathbf{w} | \mathfrak{D}_m) d\mathbf{w},$$

где функция полезности $u(\mathfrak{D}, \mathbf{w})$ задается в виде:

$$u(\mathfrak{D}_m, \mathbf{w}) = l(\mathfrak{D}_m, \mathbf{w}) - cm, \quad (6.12)$$

где c функция штрафов для каждого элемента в наборе выборки.

Назовем индексы $\mathcal{B}_1, \mathcal{B}_2 \subset \{1, \dots, m\}$ по соседству, если

$$|\mathcal{B}_1 \Delta \mathcal{B}_2| = 1. \quad (6.13)$$

Таким образом, \mathcal{B}_2 можно преобразовать в \mathcal{B}_1 путем удаления, замены или добавления одного элемента. В [74] показано, что если размер набора выборок $\mathfrak{D}_{\mathcal{B}_1}$ достаточно велик, чем параметры модели $\hat{\mathbf{w}}_1$, оптимизированные с помощью $\mathfrak{D}_{\mathcal{B}_1}$, должны находиться в окрестности параметров модели $\hat{\mathbf{w}}_2$, которые оптимизированы с помощью $\mathfrak{D}_{\mathcal{B}_2}$.

Используя дивергенцию Кульбака-Лейблера в качестве функции близости между распределениями параметров модели, оптимизированных с помощью $\mathfrak{D}_{\mathcal{B}_1}$ и $\mathfrak{D}_{\mathcal{B}_2}$:

$$D_{KL}(p_1, p_2) = \int_{\mathbf{w} \in \mathbb{W}} p_1(\mathbf{w}) \log \frac{p_1(\mathbf{w})}{p_2(\mathbf{w})} d\mathbf{w},$$

где p_1 and p_2 апостериорные вероятности вектора параметров \mathbf{w} рассчитаные на подвыборках $\mathfrak{D}_{\mathcal{B}_1}$ и $\mathfrak{D}_{\mathcal{B}_2}$ соответсвенно. Также предполагается, что $\mathfrak{D}_{\mathcal{B}_1}$ и $\mathfrak{D}_{\mathcal{B}_2}$ находятся по соседству. Достаточный размер выборки m^* оценивается:

$$\forall \mathfrak{D}_{\mathcal{B}_1} : |\mathfrak{D}_{\mathcal{B}_1}| \geq m^* \mathbb{E}_{\mathfrak{D}_{\mathcal{B}_2}} D_{KL}(p_1, p_2) \leq \varepsilon.$$

6.3. Анализ методов определения достаточного размера выборки

Проводится эксперимент для анализа свойств методов оценки достаточного размера выборки. Эксперимент состоит из трех частей. В первой части рассматриваются оценки достаточного размера выборки для разных наборов данных с фиксированным набором гиперпараметров различных методов. Во второй части исследуется зависимость достаточного размера выборки от имеющегося размера выборки. В третьей части исследуется поведение методов в зависимости от изменения гиперпараметров методов. В качестве данных использовались выборки, описанные в таблице 6.2. Методы в строках таблицы 6.3 показывают оценки размера выборки для соответствующих выборок.

В этой части вычислительного эксперимента анализируется сходимость разных методов на разных выборках. В эксперименте используются выборки:

Таблица 6.2: Описание выборок для анализа качества определения оптимального размера выборки

Выборка	Задача	Число признаков	Размер выборки
Boston Housing	regression	14	506
Diabets	regression	20	576
Forest Fires	regression	13	517
Servo	regression	4	167
NBA	classification	12	2235

Таблица 6.3: Эксперимент по оценке размера выборки для различных наборов выборок

Методы и наборы данных	Boston Housing	Diabetes	Forest Fires	Servo	NBA
Lagrange Multipliers Test	18	25	44	38	218
Likelihood Ratio Test	17	25	43	18	110
Wald Test	66	51	46	76	200
Cross Validation	178	441	172	120	–
Bootstrap	113	117	86	60	405
APVC	98	167	351	20	–
ACC	228	441	346	65	–
ALC	98	267	516	25	–
Utility Function	148	172	206	105	925

Boston Housing [68], Diabetes, Forest Fires, Servo [82], NBA. Результат анализа представлен в таблице 6.3. Символ “—” обозначает, что исходный размер выборки недостаточный для прогноза.

Гиперпараметры каждого метода для всех выборок описаны в таблице 6.4. Поскольку критерии Лагранжа, отношения правдоподобия и Вальда асимптотически эквивалентны, то параметры этих методов задавались одинаково. Параметры методов «Average Coverage» и «Average Length» также задаются одинаково.

Таблица 6.4: Экспертные оценки гиперпараметров для разных методов оценки объема выборки

Method	GLM parameters	l	ε	α	β
Lagrange Multipliers Test	\mathbf{w}_u^0	—	0.2	0.05	0.2
Likelihood Ratio Test	\mathbf{w}_u^0	—	0.2	0.05	0.2
Wald Test	\mathbf{w}_u^0	—	0.2	0.05	0.2
Cross Validation	—	—	0.05	—	—
Bootstrap	—	0.5	—	0.05	—
APVC	—	0.5	—	—	—
ACC	—	0.25	—	0.05	—
ALC	—	0.5	—	0.05	—
Utility function	—	—	0.005	—	—

Вычислительный эксперимент проводился для анализа описанных методов. Выбирается некоторый размер выборки t и методом бутстррап семплируется множество подвыборок размером t . Для разных значений t вычисляется t^* .

Рис. 6.1 демонстрирует зависимость статических значений каждого метода для разных выборок с фиксированным размером выборки t . Пороговые значения для каждого метода устанавливаются экспертно, что позволяет контролировать различные статистические характеристики выборки. Рис. 6.1 показывает адекватность различных методов определения достаточного размера выборки. Представленные функции монотонны и асимптотически стремятся к константе. На рис. 6.2 показаны результаты методов на выборках разных размера. Показано различие методов в дисперсии вычисленного t^* . Анализируются различные методы в случае небольшого размера выборки. Все представленные методы сходятся, причем результат предсказания в асимптотике не зависит от доступного размера выборки t .

Небольшое значение дисперсии интерпретируется как вычислительная устойчивость рассмотренных методов. Показано, что некоторые методы не дают оценку достаточного размера выборки, если у них нет соответствующего размера выборки. Это значит, что они не эффективны с точки зрения прогноза,

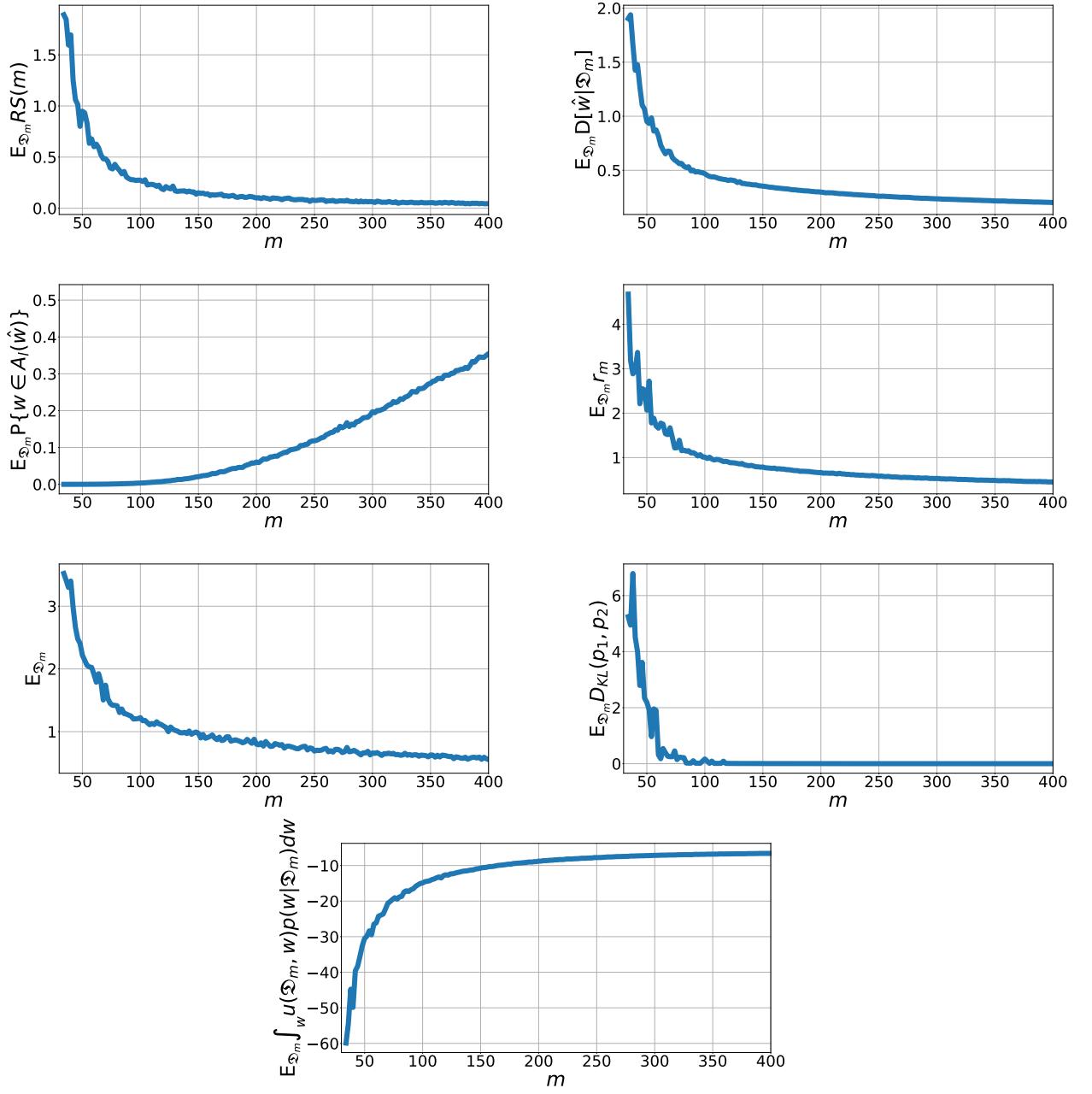


Рис. 6.1: Зависимость статистических значений различных методов

но могут быть использованы для ретроспективы и анализа уже проведенного эксперимента.

Анализируется оценка достаточного размера выборки в зависимости от гиперпараметров для байесовских методов, а также эвристических методов. Для анализа рассмотрена выборка Boston Housing. Байесовские методы используют решающее правило над скалярной функцией для определения достаточного размера выборки. На рис. 6.1 показана зависимость скалярных функций от размера подвыборки. На рис. 6.1 показано, что эти функции монотонны. Изменяя ограничения, установленные экспертом, можно изменить размер выборки, который будет соответствовать

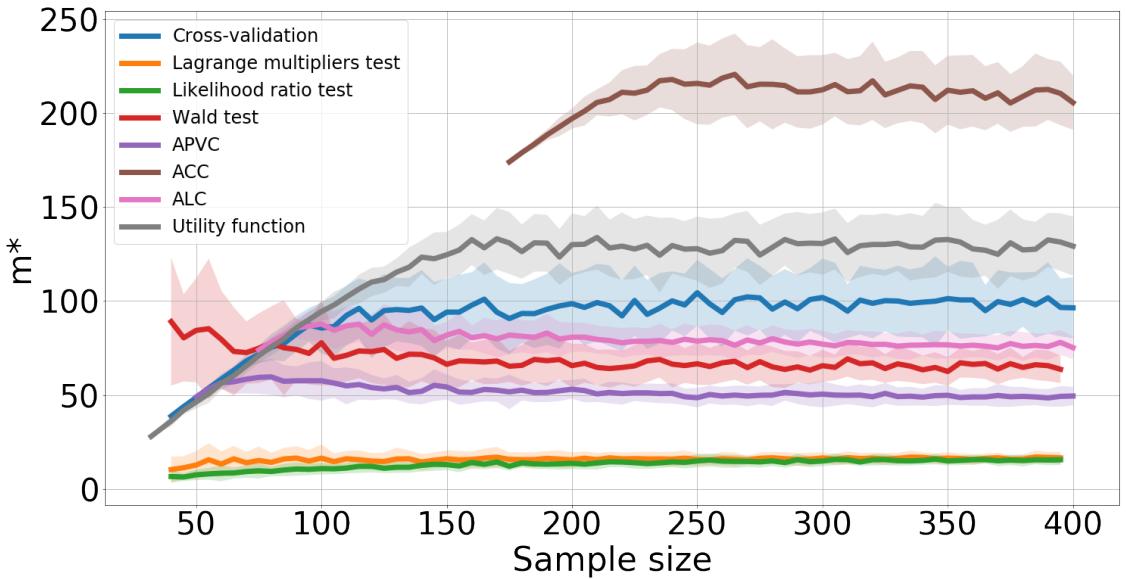


Рис. 6.2: Анализ методов в зависимости от доступного размера выборки

этим ограничениям.

6.4. Кластеризация точек квазипериодических временных рядов

Анализ физической активности человека производится при помощи мобильных телефонов, разумных часов [83, 84]. Эти устройства используют акселерометр, гироскоп и магнитометр. Цель данного исследования заключается в построении метода автоматической разметки и распознавании человеческой активности [85, 86, 87], а также поиска начала каждого действия [88]. Примерами одного сегмента действия служит шаг, шаг бега, приседание, прыжок и др. Исследуются последовательности, которые состоят не менее чем из двух подряд идущих сегментов, которые соответствуют одному и тому же типу человеческой активности.

При классификации временных рядов значимую роль играет модель построения признакового пространства. Объектом анализа и кластеризации является точка на оси времени. Решается задача кластеризации точек временного ряда. При *кластеризации* каждой точке временного ряда ставится в соответствие метка из конечного множества меток. Каждая метка соответствует одному характерному физическому действию. *Сегмент* это часть временного ряда, которая соответствует одному характерному физическому действию, например: шаг двумя ногами при ходьбе, или шаг двумя ногами при беге. Последовательность сегментов, которые соответствуют одному физическому действию образуют *цепочку* действий. Предполагается, что цепочка действий образует квазипериодическую последовательность значений временного ряда. Последовательность точек $\{b_t\}_{t=1}^N$ назовем *квазипериодической* с периодом T , если для всех t най-

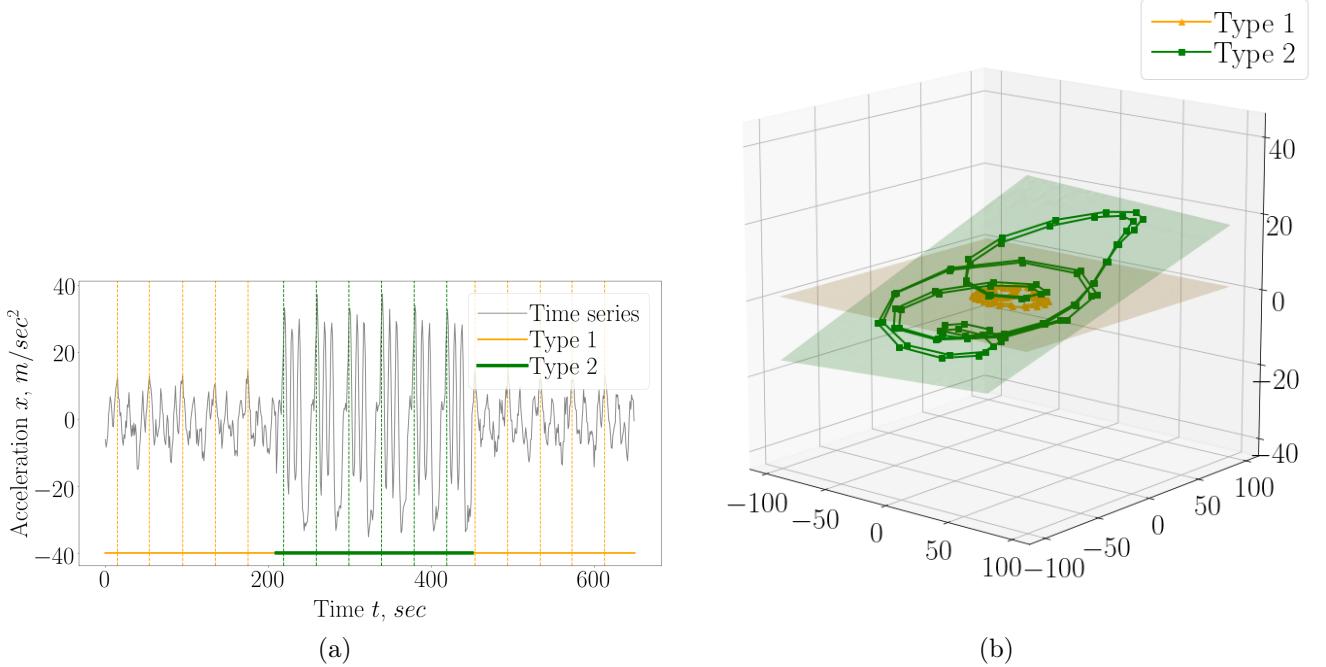


Рис. 6.3: Временной ряд, с разметкой на кластеры: а) временной ряд с аксессорской разметкой на кластеры и выделением начала квазипериодического сегмента; б) проекция фазовых траекторий на первые две главные компоненты

дется Δ , такое что:

$$b_t \approx b_{t+T+\Delta}, \quad |\Delta| \ll T.$$

Пример кластеризации и разбиения ряда на сегменты показан на рис. 6.3а. Данный ряд разбит на два характерных физических действия, которые обозначаются Type 1 и Type 2. Также данный ряд содержит в себе две квазипериодические цепочки действий.

Решение задачи кластеризации состоит из двух этапов. Во-первых, для получения признакового описания временного ряда предлагается алгоритм локальной аппроксимации временного ряда при помощи метода главных компонент [89]. Под локальной аппроксимацией временного ряда подразумевается, что для признакового описания его точки используется не весь ряд, а только некоторая окрестность данной точки. В качестве признакового описания точки временного ряда рассматриваются две главные компоненты *сегмента фазовой траектории* в окрестности данной точки. На рис. 6.3б показаны две первые главные компоненты *фазовых траекторий*, а также проекция фазовых траекторий на эти компоненты. Они соответствуют разным физическим действиям, которые обозначаются Type 1 и Type 2, внутри одного временного ряда. Как видно, плоскости, порожденные главными компонентами не совпадают. Это говорит о том, что наблюдаются различные действия. Во-вторых, вводится функция расстояния в построенном пространстве признакового описания. Данная функция является расстоянием между двумя базисами некоторых под-

пространств внутри всего фазового пространства временного ряда. На рис. 6.3б данная функция является некоторым расстояниям между двумя плоскостями. Получив расстояния между точками временного ряда, выполним кластеризацию данных точек. Задача сегментации внутри каждого кластера решается при помощи метода, который рассмотрен в [88].

Для решения задачи кластеризации точек временного ряда вводятся предположения. Предполагается, что периоды различных сегментов различаются незначительно, причем известны минимальный и максимальный периоды сегмента и число различных сегментов внутри временного ряда. Также предполагается, что тип активности во времени не меняется часто, а также что фазовые траектории разных сегментов являются различными.

Проверка и анализ метода кластеризации проводится на синтетической и реальной выборках. Синтетическая выборка построена при помощи суммы нескольких первых членов ряда Фурье со случайными коэффициентами. Эксперимент по сегментации временного ряда проводился на простых синусоидальных сигналах с произвольной амплитудой и частотой. Реальные данные получены при помощи мобильного акселерометра, который снимал показания во время некоторой физической активности человека.

В [83] рассматривает метод построения признакового описания на основе экспертно-заданных порождающих функций. В [90] рассматривается метод построения признаков на основе гипотезы порождения данных. В [91] рассматривается комбинированное признаковое описание на основе данных методов. В [92] рассматривается проблема построение признакового пространства и предлагаются критерий избыточности выбранных признаков.

Работа [88] является ближайшей работой по данной теме. Она заключается в поиске начала сегмента внутри квазипериодического сигнала, который состоит, только из одной цепочки действий. Этот метод основан на исследовании фазового пространства, а именно поиска устойчивой гиперплоскости, которая делит фазовое пространство на две равные части. В качестве начала сегмента выбираются точки, которые находятся близко к данной гиперплоскости. В [88] предлагается выполнить проекцию фазового пространства на первые две главные компоненты, после чего провести устойчивую прямую, выделив начала каждого сегмента. Данный метод имеет недостаток в том, что позволяет находить начало только для временного ряда, который состоит из квазипериодического сигнала единственного типа.

Также близкой является работа [87]. Требуется найти периодическую структуру внутри ряда при помощи модели LSTM с модифицированным механизмом внимания. Предполагается, что механизм внимания будет давать максимальное значение качества в точках, которые удалены от данной на целое число периодов.

Задан временной ряд

$$\mathbf{x} \in \mathbb{R}^N,$$

где N число точек временного ряда. Он состоит из последовательности сегментов:

$$\mathbf{x} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M],$$

где \mathbf{v}_i некоторый сегмент из множества сегментов \mathbf{V} , которые встречаются в данном ряде. Причем для всех i либо $[\mathbf{v}_{i-1}, \mathbf{v}_i]$ либо $[\mathbf{v}_i, \mathbf{v}_{i+1}]$ является цепочкой действий. Пусть множество \mathbf{V} удовлетворяет свойствам:

$$|\mathbf{V}| = K, \quad \mathbf{v} \in \mathbf{V} \quad |\mathbf{v}| \leq T,$$

где $|\mathbf{V}|$ число различных действий в множестве сегментов \mathbf{V} , $|\mathbf{v}|$ длина сегмента, а K и T это число различных действий во временном ряде и длина максимального сегмента соответственно.

Рассматривается отображение

$$a : t \rightarrow \mathbb{Y} = \{1, \dots, K\}, \quad (6.14)$$

где $t \in \{1, \dots, N\}$ некоторый момент времени, на котором задан временной ряд. Требуется, чтобы отображение a удовлетворяло свойствам:

$$\begin{cases} a(t_1) = a(t_2), & \text{если в моменты } t_1, t_2 \text{ совершаются один тип действий,} \\ a(t_1) \neq a(t_2), & \text{если в моменты } t_1, t_2 \text{ совершаются разные типы действий.} \end{cases}$$

Пусть задана некоторая асессорская разметка временного ряда:

$$\mathbf{y} \in \{1, \dots, K\}^N.$$

Тогда ошибка алгоритма a на временном ряде \mathbf{x} представляется в виде:

$$S = \frac{1}{N} \sum_{t=1}^N [y_t = a(t)],$$

где t — момент времени, y_t асессорская разметка t -го момента времени для заданного временного ряда.

Кластеризация точек в фазовом пространстве. Рассмотрим фазовую траекторию временного ряда \mathbf{x} :

$$\mathbf{H} = \{\mathbf{h}_t | \mathbf{h}_t = [x_{t-T}, x_{t-T+1}, \dots, x_t], T \leq t \leq N\},$$

где \mathbf{h}_t — точка фазовой траектории.

Информация об длине максимального сегмента T внутри временного ряда позволяет разбить фазовую траекторию на сегменты из $2T$ векторов:

$$\mathbf{S} = \{\mathbf{s}_t | \mathbf{s}_t = [\mathbf{h}_{t-T}, \mathbf{h}_{t-T+1}, \dots, \mathbf{h}_{t+T-1}], T \leq t \leq N - T\},$$

где \mathbf{s}_t — это сегмент фазовой траектории. Данные сегменты имеют всю локальную информацию об временном ряде, так как содержит всю информацию на периоде до момента времени t и информацию о периоде после момента времени t .

В качестве признакового описания точки временного ряда t рассматривают главные компоненты \mathbf{W}_t для T -мерных сегментов \mathbf{s}_t . Сегмент \mathbf{s}_t проектируется на подпространство размерности два при помощи метода главных компонент $\mathbf{z}_t = \mathbf{W}_t \mathbf{s}_t$. Получаем:

$$\mathbf{W} = \{\mathbf{W}_t | \mathbf{W}_t = [\lambda_t^1 \mathbf{w}_t^1, \lambda_t^2 \mathbf{w}_t^2]\}, \quad \Lambda = \{\boldsymbol{\lambda}_t | \boldsymbol{\lambda}_t = [\lambda_t^1, \lambda_t^2]\},$$

где $[\mathbf{w}_t^1, \mathbf{w}_t^2]$ и $[\lambda_t^1, \lambda_t^2]$ это базисные векторы и соответствующие им собственные для сегмента фазовой траектории \mathbf{s}_t .

Для кластеризации точек временного ряда рассмотрим функцию расстояния между элементами $\mathbf{W}_{t_1}, \mathbf{W}_{t_2}$:

$$\rho(\mathbf{W}_1, \mathbf{W}_2) = \max \left(\max_{\mathbf{e}_2 \in \mathbf{W}_2} d_1(\mathbf{e}_2), \max_{\mathbf{e}_1 \in \mathbf{W}_1} d_2(\mathbf{e}_1) \right),$$

где \mathbf{e}_i это базисный вектор пространства \mathbf{W}_i , а $d_i(\mathbf{e})$ является расстоянием от вектора \mathbf{e} до пространства \mathbf{W}_i .

В случае, когда все подпространства \mathbf{W}_t имеют размерность два, расстояние $\rho(\mathbf{W}_1, \mathbf{W}_2)$ имеет интерпретацию:

$$\rho(\mathbf{W}_1, \mathbf{W}_2) = \max_{\{\mathbf{a}, \mathbf{b}, \mathbf{c}\} \subset \mathbf{W}_1 \cup \mathbf{W}_2} V(\mathbf{a}, \mathbf{b}, \mathbf{c}), \quad (6.15)$$

где $\mathbf{W}_1 \cup \mathbf{W}_2$ это объединение базисных векторов первого и второго пространства, $V(\mathbf{a}, \mathbf{b}, \mathbf{c})$ — объем параллелепипеда построенного на векторах $\mathbf{a}, \mathbf{b}, \mathbf{c}$, которые являются столбцами матрицы $\mathbf{W}_1 \cup \mathbf{W}_2$.

Рассмотрим расстояние между собственными числами:

$$\rho(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \sqrt{(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^T (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)}. \quad (6.16)$$

Используя выражения (6.15-6.16) введем расстояние между двумя точками t_1, t_2 временного ряда, а также рассмотрим матрицу попарных расстояний \mathbf{M} между точками данного ряда:

$$\rho(t_1, t_2) = \rho(\mathbf{W}_1, \mathbf{W}_2) + \rho(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2), \quad \mathbf{M} = \mathbb{R}^{N \times N}, \quad (6.17)$$

где матрица \mathbf{M} является матрицей попарных расстояний между всеми парами точек t временного ряда \mathbf{x} . Используя матрицу попарных расстояний \mathbf{M} выполним кластеризацию моментов времени t временного ряда (6.14):

$$a : t \rightarrow \{1, \dots, K\},$$

где t некоторый момент времени временного ряда \mathbf{x} .

6.5. Анализ фазовых траекторий в задаче кластеризации

Для анализа свойств предложенного алгоритма кластеризации проведен вычислительный эксперимент в котором кластеризация точек временного ряда проводилась используя матрицы попарных расстояний (6.17).

В качестве данных использовались две выборки временных рядов, которые описаны в таблице 6.5. Выборка Physical Motion это реальные временные ряды полученные при помощи мобильного акселерометра. Синтетические временные ряды построены при помощи нескольких первых слагаемых ряда Фурье со случайными коэффициентами из стандартного нормального распределения. Генерация данных состояла из двух этапов. На первом этапе генерировались короткие сегменты \mathbf{v} для построения множества \mathbf{V} . Вторым этапом генерации выборки \mathbf{x} является случайнм процессом:

$$\mathbf{x} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M] + \boldsymbol{\epsilon}, \quad \begin{cases} \mathbf{v}_1 \sim \mathcal{U}(\mathbf{V}), \\ \mathbf{v}_i = \mathbf{v}_{i-1}, & \text{с вероятностью } \frac{3}{4}, \\ \mathbf{v}_i \sim \mathcal{U}(\mathbf{V}), & \text{с вероятностью } \frac{1}{4} \end{cases}$$

где $\mathcal{U}(\mathbf{V})$ — равномерное распределение на объектах из \mathbf{V} , а $\boldsymbol{\epsilon}$ является шумом из нормального распределения.

Таблица 6.5: Описание временных рядов в эксперименте кластеризации точек временного ряда

Ряд, \mathbf{x}	Длина ряда, N	Число сегментов, K	Длина сегмента, T
Physical Motion 1	900	2	40
Physical Motion 2	900	2	40
Synthetic 1	2000	2	20
Synthetic 2	2000	3	20

На рис. 6.4 приведен пример синтетических временных рядов. На рис. 6.4a показан пример ряда в котором число различных сегментов $K = 2$, а длина каждого сегмента $T = 20$. На рис. 6.4b показан пример ряда в котором число различных сегментов $K = 3$, а длина каждого сегмента $T = 20$.

Рис. 6.5 иллюстрирует матрицы попарных расстояний \mathbf{M} между всеми парами точек t временного ряда, которые построены при помощи (6.17). Используя матрицу попарных расстояний и метод многомерного шкалирования [93] визуализируем точки временного ряда на плоскости. На рис. 6.6 показана визуализация точек на плоскости и выполнена их кластеризация при помощи метода иерархической кластеризации. Иллюстрация кластеров точек временного ряда продемонстрирована на рис. 6.7.

На рис. 6.8 приведен пример реальных временных рядов полученных при помощи взятия одной из координат мобильного акселерометра.

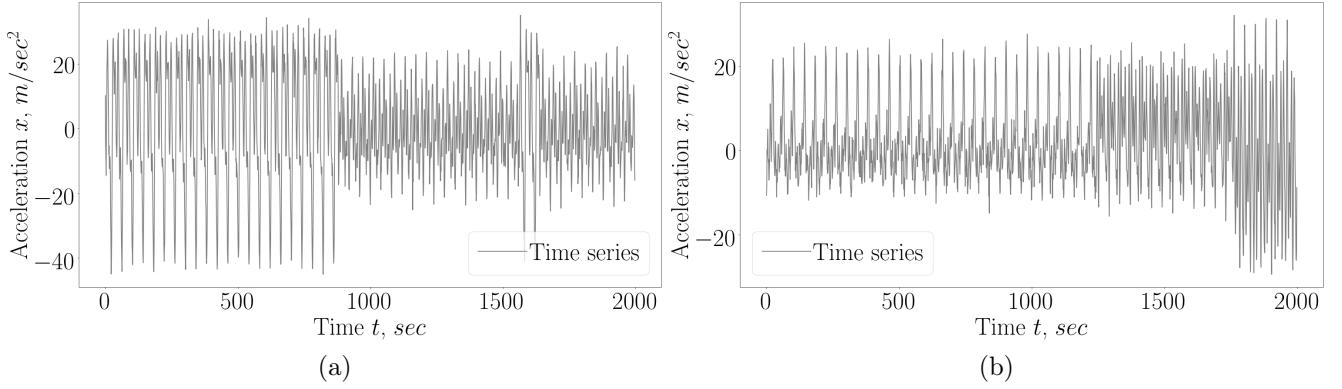


Рис. 6.4: Пример синтетически построенных временных рядов: а) для временного ряда Synthetic 1; б) для временного ряда Synthetic 2

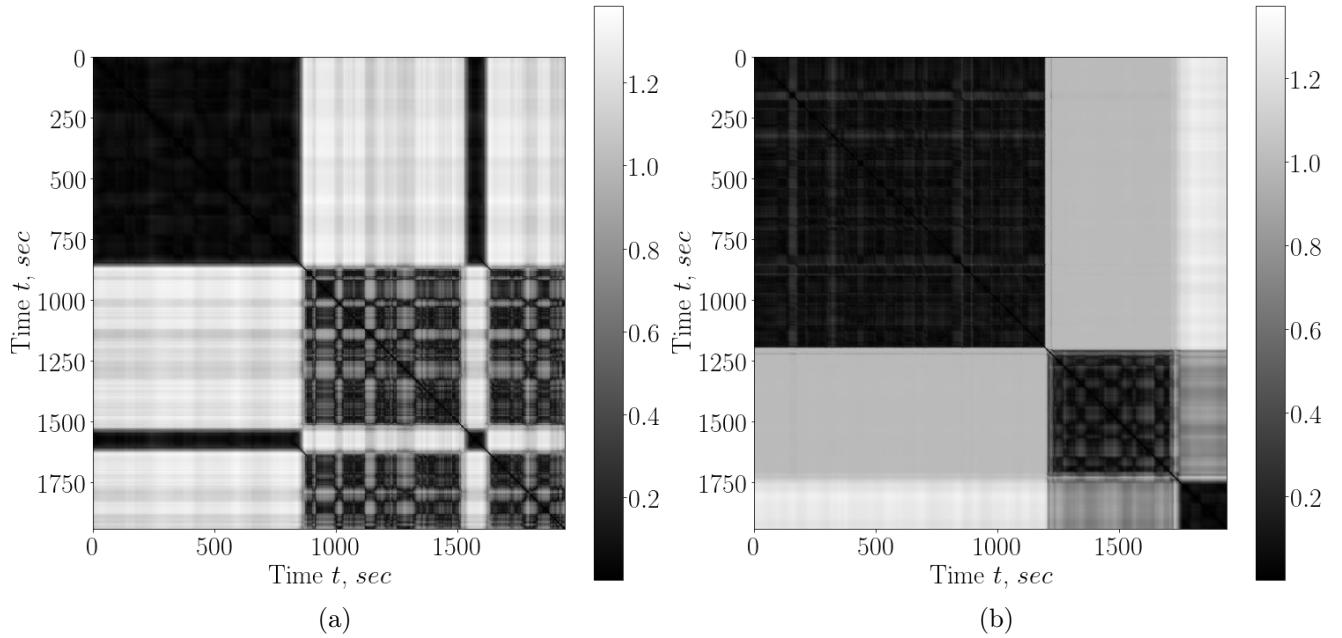


Рис. 6.5: Матрица попарных расстояний M между точками временного ряда: а) для временного ряда Synthetic 1; б) для временного ряда Synthetic 2

Рис. 6.9 иллюстрирует матрицы попарных расстояний M между всеми парами точек t временного ряда, которые построены при помощи (6.17). Используя матрицу попарных расстояний и метод многомерного шкалирования [93] визуализируем точки временного ряда на плоскости. На рис. 6.10 показана визуализация точек на плоскости и выполнена их кластеризация при помощи метода иерархической кластеризации. Иллюстрация кластеров точек временного ряда продемонстрирована на рис. 6.11.

Сегментация временных рядов проводится на синтетических и реальных данных. Для данного эксперимента в качестве синтетического ряда рассматривается ряд построенный из двух синусов с произвольной частотой и амплитудой.

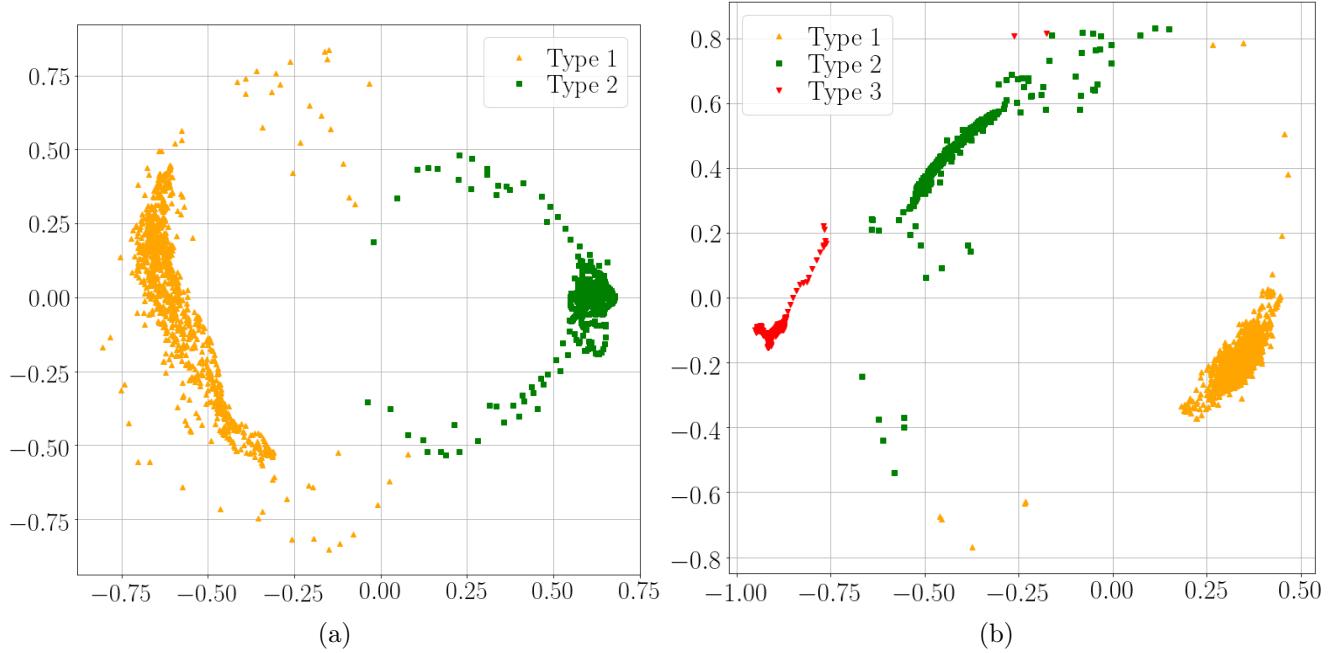


Рис. 6.6: Проекция точек временного ряда на плоскость при помощи матрицы попарных расстояний \mathbf{M} : а) для временного ряда Synthetic 1; б) для временного ряда Synthetic 2

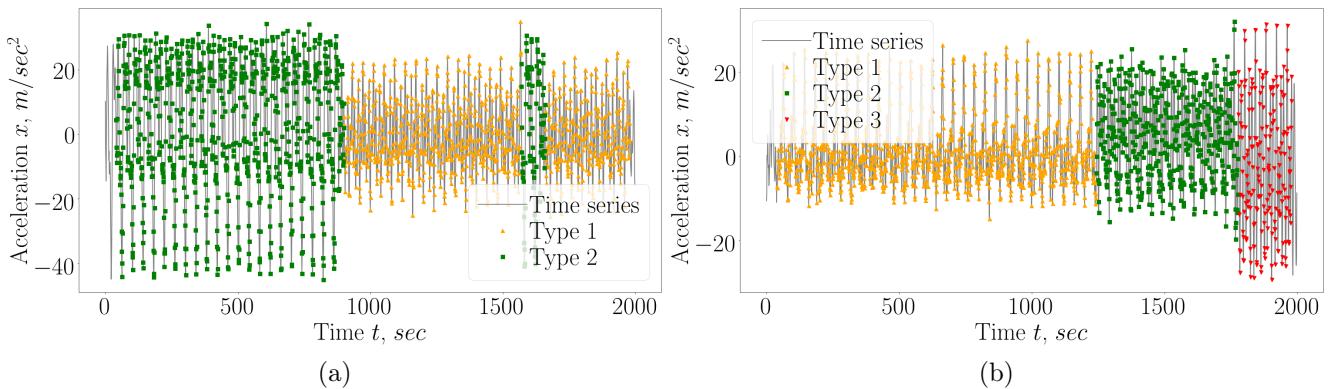


Рис. 6.7: Кластеризация точек временного ряда: а) для временного ряда Synthetic 1; б) для временного ряда Synthetic 2

дой. Описание временных рядов, которые используются в данном эксперименте представлены в таблице 6.6.

Сегментация проводится при помощи метода, который представлен в работе [88]. Данный метод применяется для каждого действия внутри временного ряда по отдельности.

На рис. 6.12 показан результат сегментации для временного ряда Simple 1. Данный алгоритм хорошо выделил начала сегментов. Также на рис. 6.12 показаны проекции фазовых пространств для обеих кластеров на их первые две главные компоненты.

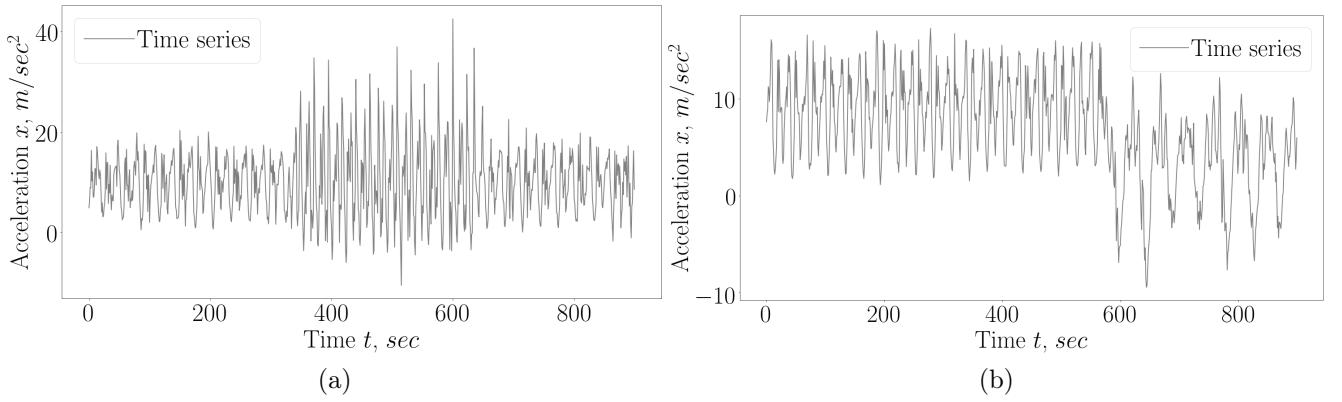


Рис. 6.8: Пример синтетически построенных временных рядов: а) для временного ряда Physical Motion 1; б) для временного ряда Physical Motion 2

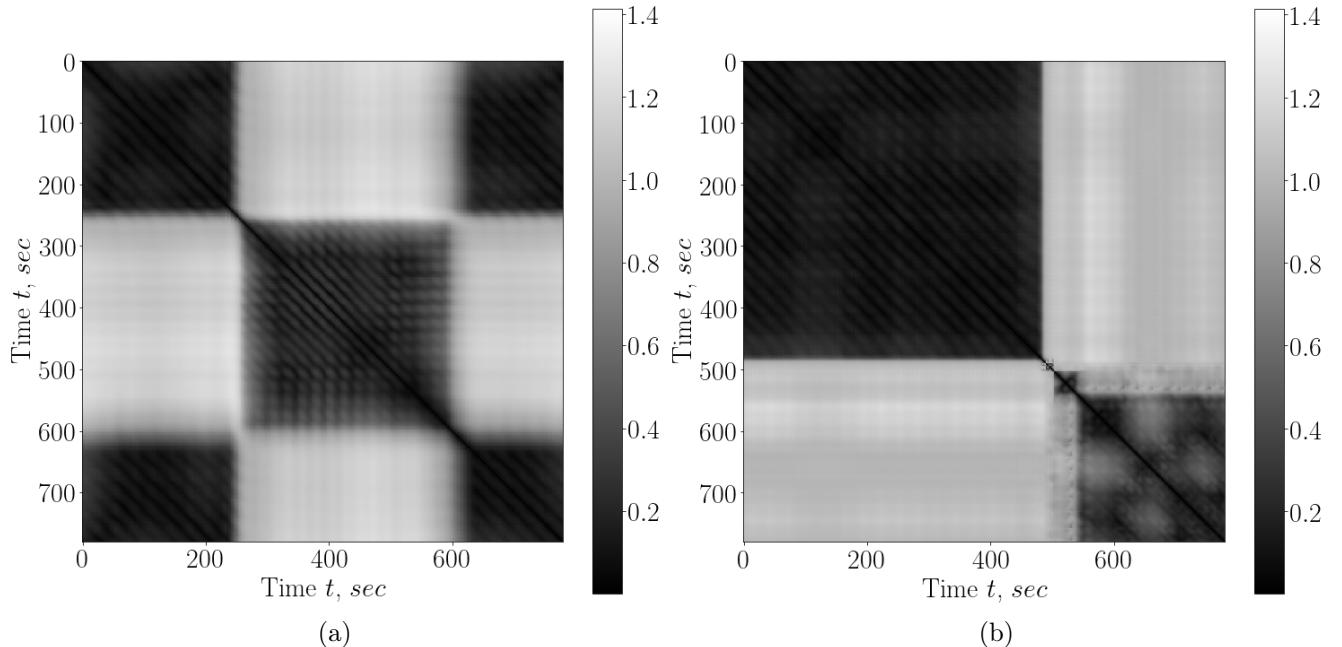


Рис. 6.9: Матрица попарных расстояний \mathbf{M} между точками временного ряда: а) для временного ряда Physical Motion 1; б) для временного ряда Physical Motion 2

Таблица 6.6: Описание временных рядов в эксперименте сегментации временных рядов

Ряд, \mathbf{x}	Длина ряда, N	Число сегментов, K	Длина сегмента, T
Simple 1	1000	2	100
Physical Motion 2	900	2	40

Реальные данные. На рис. 6.13 показан результат сегментации для временного ряда Physical Motion 2. Данный алгоритм хорошо выделил начала сег-

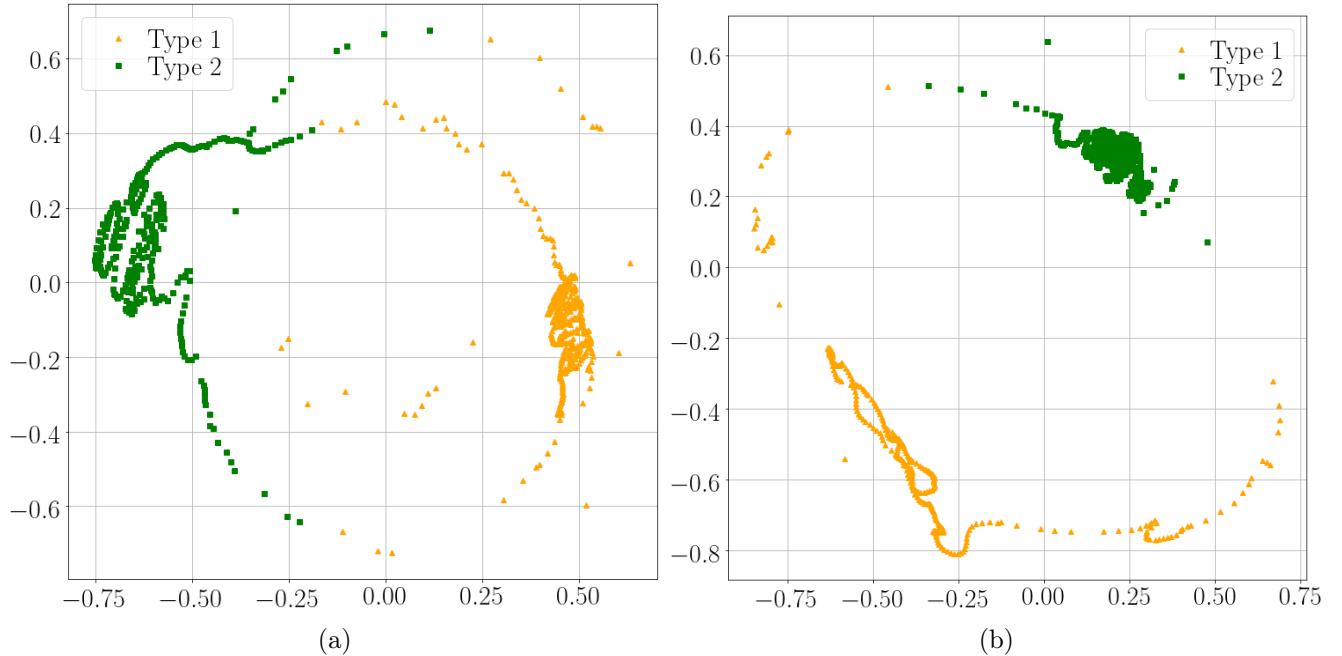


Рис. 6.10: Проекция точек временного на плоскость при помощи матрицы попарных расстояний \mathbf{M} : а) для временного ряда Physical Motion 1; б) для временного ряда Physical Motion 2

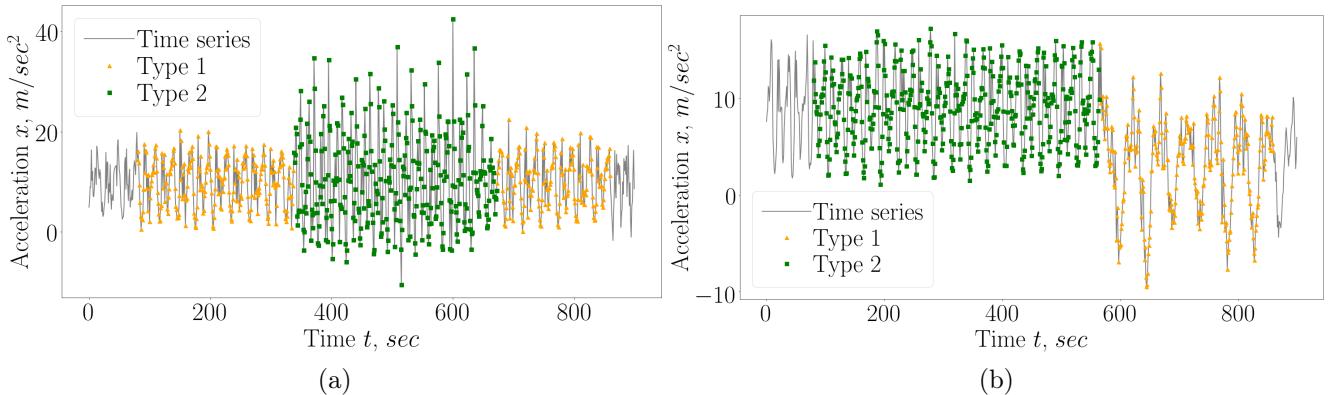


Рис. 6.11: Кластеризация точек временного ряда: а) для временного ряда Physical Motion 1; б) для временного ряда Physical Motion 2

ментов для Type 1 и плохо для Type 2. Также на рис. 6.13 показаны проекции фазовых пространств для обеих кластеров на их первые две главные компоненты. Видно, что в случае проекции фазового пространства для части ряда, который относится к Type 2 получаем, что фазовая траектория имеет само-пересечение внутри одного сегмента, что влечет нахождения ложного начала сегмента.

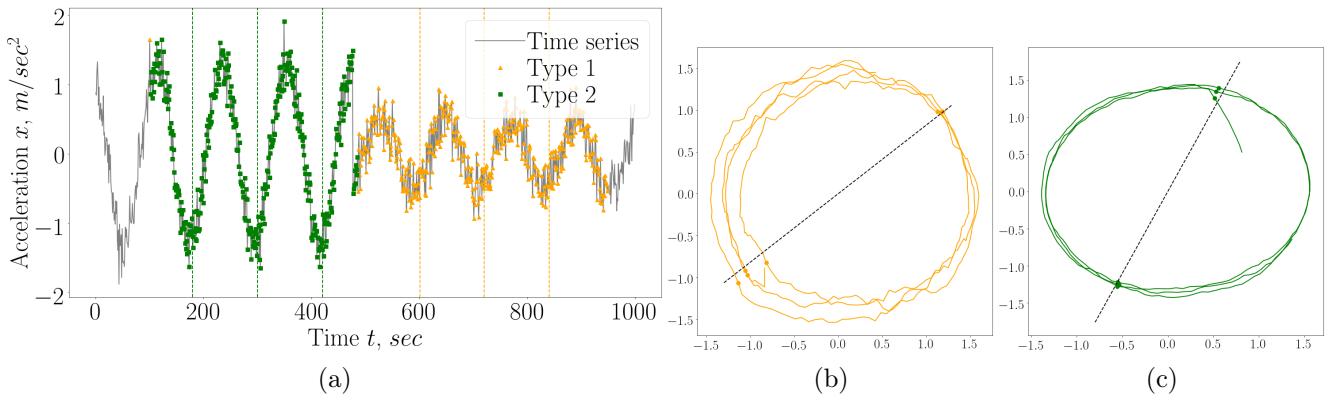


Рис. 6.12: Сегментация точек временного ряда Simple 1: а) сегментация временного ряда; б) проекция фазового пространства на первые две главные компоненты для первого кластера; в) проекция фазового пространства на первые две главные компоненты для второго кластера

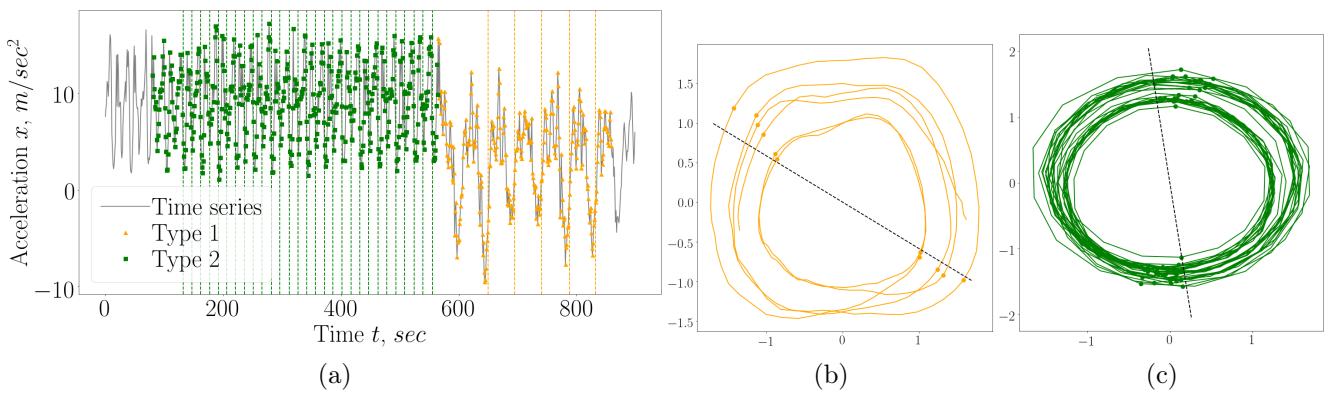


Рис. 6.13: Сегментация точек временного ряда Physical Motion 2: а) сегментация временного ряда; б) проекция фазового пространства на первые две главные компоненты для первого кластера; в) проекция фазового пространства на первые две главные компоненты для второго кластера

Заключение

Основные результаты диссертационной работы заключаются в следующем.

В главе 1 введены основные понятия, поставлены задачи выбора априорного распределения параметров моделей машинного обучения. Проанализированы методы дистилляции и привилегированного обучения предложенные Владимиром Наумовичем Вапником и Джеком Хинтоном. Проанализированы методы задания порядка на множестве параметров нейросетевых моделей. Последние включают в себя как эвристические методы, так и методы, основанные на байесовском выводе и вероятностных предположениях о распределении параметров моделей.

В главе 2 были предложены методы обобщения дистилляции и привилегированного обучения на основе вероятностного подхода. Введены вероятностные предположения, описывающие дистилляцию моделей. В рамках данных вероятностных предположений проанализированы модели для задачи классификации и регрессии. Результат анализа сформулирован в виде соответствующих *теорем об эквивалентности*. Теорема для задачи регрессии показала, что обучение линейной регрессии с учителем эквивалентно замене обучающей выборки и вероятностных предположений о распределении истинных ответов. Для задачи классификации ответы учителя дают же дополнительную информацию. Она задается в виде распределения классов для каждого объекта из обучающей выборки. Для задачи классификации проведен вычислительный эксперимент. Из вычислительного эксперимента видно, что дистилляция влияет на распределение классов для каждого объекта. Вероятности классов для каждого объекта являются более разреженными, а не концентрируются в одном классе. Данное свойство представлено в синтетической выборке, так как она генерировалась с максимальной дисперсией в вероятностях классов.

В главе 3 был предложен байесовский метод для дистилляции моделей глубокого обучения на основе вариационного вывода. Дистилляция основывается на задании априорного распределения параметров модели ученика. Априорное распределение параметров модели ученика задается на основе апостериорного распределения параметров модели учителя. Представлен механизм выравнивания структуры модели учителя в структуру модели ученика. Результаты анализа данного механизма сформулированы в *теоремах о виде априорного распределения*. В теоремах получен вид распределения после выравнивания пространства параметров модели учителя к пространству параметров модели ученика.

В главе 4 предложены методы задания априорного распределения параметров локальных моделей в задаче обучения смеси экспертов. Рассмотрен подход задания гиперпараметров априорного распределения на основе экспертной информации о рассматриваемой задачи. Вводится понятие регуляризации гиперпараметров априорных распределений согласно экспертной информации. Предлагается метод оптимизации гиперпараметров и параметров смеси экспертов в

единой процедуре на основе ЕМ-алгоритма. Представлены результаты экспериментов по сравнению различных способов регуляризации гиперпараметров априорных распределений.

В главе 5 предложены методы введения отношения порядка на множестве параметров аппроксимирующих моделей. Введены вероятностные предположения в рамках которых решается поставленная задача. Предложен метод задания порядка на основе метода Белсли. Метод анализирует мультиколлинеарность параметров при обучении моделей глубокого обучения. Предложен метод задания порядка на основе анализа стохастических свойств градиента функции ошибки по параметрам модели. Данный метод использует ковариационную матрицу градиентов параметров. Предложенные методы проанализированы в вычислительном эксперименте. Показано, что предложенный порядок задает порядок на параметрах модели и соответствует влиянию параметров на функцию ошибки при оптимизации.

В главе 6 проведен анализ прикладных задач. Рассмотрена задача определения оптимального размера выборки. В ней задание экспертной информации о гипотезе порождения данных и классе аппроксимирующих моделей указывает достаточный объем выборки. Анализируется метод кластеризации точек квазипериодических временных рядов в случае, когда известна экспертная информация о структуре рассматриваемых рядов.

Список основных обозначений

- $\mathbf{x}_i \in \mathbf{X}$ — вектор признакового описания i -го объекта
 $y_i \in \mathbf{y}$ — метка i -го объекта
 \mathfrak{D} — выборка для аппроксимации
 $\mathbb{X} = \mathbb{R}^m$ — признаковое пространство объектов
 \mathbb{Y}' — множество оценок меток
 \mathbb{Y} — множество меток объектов
 $\mathbf{X} \subset \mathbb{X}$ — матрица, содержащая признаковое описание объектов выборки
 \mathbf{I} — индексное множество объектов с привилегированной информацией
 $\mathbf{y} \subset \mathbb{Y}$ — вектор меток объектов выборки
 $\mathbf{s} \subset \mathbb{Y}'$ — вектор оценок меток учителем
 m — число объектов в выборке
 n — число признаков в признаковом описании объекта
 R — число классов в задаче классификации
 K — число локальных моделей в ансамбле
 \mathfrak{F} — параметрическое семейство моделей учителя
 \mathfrak{G} — параметрическое семейство моделей ученика
 \mathbf{f} — модель учителя
 \mathbf{g} — модель ученика
 \mathbf{u} — вектор параметров модели учителя
 \mathbf{w} — вектор параметров модели ученика
 $\mathbf{w} \in \mathbb{W}$ — параметры модели ученика
 \mathbb{W} — пространство параметров модели ученика
 $\mathbf{u} \in \mathbb{U}$ — параметры модели учителя
 \mathbb{U} — пространство параметров модели учителя
 \mathbf{n} — структура модели
 \mathcal{L} — функция потерь
 $p(\mathbf{u})$ — априорное распределение параметров учителя
 $p(\mathbf{u}|\mathfrak{D})$ — апостериорное распределение параметров учителя
 $p(\mathbf{w})$ — априорное распределение параметров ученика
 $\mathbf{m}, \boldsymbol{\mu}$ — мат. ожидание параметров учителя до и после выравнивания
 $\Sigma, \boldsymbol{\Xi}$ — ковариационная матрица параметров учителя до и после выравнивания
 $\pi(\mathbf{x}, \mathbf{V})$ — шлюзовая функция в смеси экспертов
 $\mathbf{V} \in \mathbb{V}$ — параметры шлюзовой функции
 \mathbb{V} — пространство параметров шлюзовой функции
 E — экспертная информация о выборке
 $K_x(\cdot, E), K_y(\cdot, E)$ — отображение в признаковое описание объектов на основе экспертной информации

Список иллюстраций

2.1	Вероятностная модель в графовой нотации.	20
2.2	Зависимость кросс-энтропии между истинными метками и предсказанными учеников вероятностями классов: а) на обучающей выборке; б) на тестовой выборке	25
2.3	Вероятностная модель используемая в синтетическом эксперименте	26
2.4	Истинное распределение объектов по классам	26
2.5	Распределение предсказанное моделью без использования информации об истинном распределении на классах	27
2.6	Распределение предсказанное моделью с использования информации об истинном распределении на классах	27
2.7	Вероятности классов для разных объектов	28
3.1	Структура (3.10) модели ученика g . Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: KL-дивергенция между вариационным и априорным распределениями параметров модели	43
3.2	Структура (3.11) модели ученика g . Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: KL-дивергенция между вариационным и априорным распределениями параметров модели	43
3.3	Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: KL-дивергенция между вариационным и априорным распределениями параметров модели	44
4.1	Пример окружностей с разным уровнем шума: (а) окружности без шума; (б) окружности с зашумленным радиусом; (с) окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению	46
4.2	Пример: а) экспертная информация первого эксперта; б) исходные данные; в) экспертная информация второго эксперта	47
4.3	Пример изображения радужной оболочки глаза и ее контурное изображение: а) изображение радужной оболочки глаза; б) контурное изображение радужной оболочки и аппроксимация заданного изображения окружностей	48
4.4	Мультимодель в зависимости от разных априорных предположений и в зависимости от разного уровня шума: (а)–(с) модель с регуляризацией априорных распределений; (д)–(г) модель с заданными априорными распределениями на параметрах локальных моделей; (е)–(ж) модель без заданных априорных предположений .	59

4.5 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	60
4.6 График зависимости логарифма правдоподобия (4.5) от номера итерации	61
4.7 Визуализации процесса сходимости мультимодели с использованием априорной регуляризации	61
4.8 Визуализации процесса сходимости мультимодели с использованием априорного распределением	61
4.9 Визуализации процесса сходимости мультимодели без использования априорного распределения	61
4.10 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	62
4.11 График зависимости логарифма правдоподобия (4.5) от уровня шума	63
4.12 Мультимодель в зависимости от разных априорных предположений на реальном изображении: (a) исходное изображение; (b) бинаризованное изображение; (c) мультимодель без априорных предположений; (d) мультимодель с априорными распределениями на параметрах локальных моделей; (e) мультимодель с регуляризацией на априорных распределениях параметров локальных моделей	64
4.13 Визуализации процесса сходимости мультимодели без использования априорного распределения	64
4.14 Визуализации процесса сходимости мультимодели с использованием априорного распределением	65
4.15 Визуализации процесса сходимости мультимодели с использованием априорной регуляризации	65
4.16 Мультимодель в зависимости от различных предварительных предположений и уровня шума. Слева направо: окружности без шума; шум в радиусе круга; шум в радиусе круга, а также произвольные точки по всему изображению	65
4.17 Зависимость параметров r , x_0 и y_0 от номера итерации для различных априорных распределений. Слева направо: окружности без шума; шум в радиусе круга; шум в радиусе круга, а также произвольные точки по всему изображению	66
4.18 Результат аппроксимации для данных с разными уровнями шума β и дисперсией априорного распределения γ	67

4.19	Зависимость моделей от уровня шума β в данных, а также от дисперсии априорного распределения γ	67
4.20	Визуализация приближения радужной оболочки: а) если указан регуляризатор R_0 ; б) если указан регуляризатор R_1 ; б) если указан регуляризатор R_2	68
4.21	Визуализация процесса сходимости параметров мультимодели в случае регуляризатора R_0	68
4.22	Визуализация процесса сходимости параметров мультимодели в случае регуляризатора R_1	69
4.23	Визуализация процесса сходимости параметров мультимодели в случае регуляризатора R_2	69
5.1	Иллюстрация метода Белсли для анализа мультиколлинеарности параметров	74
5.2	Качество прогноза при удаление параметров на выборке Wine . .	77
5.3	Влияние шума в начальных данных на шум выхода нейросети на выборке Wine	77
5.4	Качество прогноза при удаление параметров на выборке Boston .	78
5.5	Влияние шума в начальных данных на шум выхода нейросети на выборке Boston	78
5.6	Качество прогноза при удаление параметров на синтетической выборке	79
5.7	Влияние шума в начальных данных на шум выхода нейросети на синтетической выборке	80
5.8	Зависимость качества модели от числа зафиксированных параметров: а) на обучающей выборке; б) на тестовой выборке . . .	81
5.9	Визуализация векторов $\hat{\alpha}(\zeta)$ в зависимости от числа фиксируемых параметров: а) все параметры модели упорядочены предложенным методом; б) все параметры модели упорядочены произвольным образом; с) часть параметров модели упорядочена предложенным методом; д) часть параметров модели упорядочена произвольным образом	82
5.10	Зависимость качества модели от числа зафиксированных параметров: а) на обучающей выборке; б) на тестовой выборке . . .	82
5.11	Визуализация векторов $\hat{\alpha}(\zeta)$ в зависимости от числа фиксируемых параметров: а) все параметры модели упорядочены предложенным методом; б) все параметры модели упорядочены произвольным образом; с) часть параметров модели, упорядочена предложенным методом; д) часть параметров модели упорядочена произвольным образом	83
5.12	Зависимость качества модели от числа зафиксированных параметров: а) на обучающей выборке; б) на тестовой выборке . . .	83

5.13 Визуализация векторов $\hat{\alpha}(\zeta)$ в зависимости от числа фиксируемых параметров: а) все параметры модели упорядочены предложенным методом; б) все параметры модели упорядочена произвольным образом; в) часть параметров модели, упорядочены предложенным методом; г) часть параметров модели упорядочена произвольным образом	84
5.14 Зависимость качества модели от числа зафиксированных параметров: а) на обучающей выборке; б) на тестовой выборке	84
5.15 Визуализация векторов $\hat{\alpha}(\zeta)$ в зависимости от числа фиксируемых параметров: а) все параметры модели упорядочены предложенным методом; б) все параметры модели упорядочены произвольным образом; в) часть параметров модели упорядочена предложенным методом; г) часть параметров модели упорядочена произвольным образом	85
6.1 Зависимость статистических значений различных методов	97
6.2 Анализ методов в зависимости от доступного размера выборки	98
6.3 Временной ряд, с разметкой на кластеры: а) временной ряд с аксессорской разметкой на кластеры и выделением начала квазипериодического сегмента; б) проекция фазовых траекторий на первые две главные компоненты	99
6.4 Пример синтетически построенных временных рядов: а) для временного ряда Synthetic 1; б) для временного ряда Synthetic 2	104
6.5 Матрица попарных расстояний \mathbf{M} между точками временного ряда: а) для временного ряда Synthetic 1; б) для временного ряда Synthetic 2	104
6.6 Проекция точек временного ряда на плоскость при помощи матрицы попарных расстояний \mathbf{M} : а) для временного ряда Synthetic 1; б) для временного ряда Synthetic 2	105
6.7 Кластеризация точек временного ряда: а) для временного ряда Synthetic 1; б) для временного ряда Synthetic 2	105
6.8 Пример синтетически построенных временных рядов: а) для временного ряда Physical Motion 1; б) для временного ряда Physical Motion 2	106
6.9 Матрица попарных расстояний \mathbf{M} между точками временного ряда: а) для временного ряда Physical Motion 1; б) для временного ряда Physical Motion 2	106
6.10 Проекция точек временного на плоскость при помощи матрицы попарных расстояний \mathbf{M} : а) для временного ряда Physical Motion 1; б) для временного ряда Physical Motion 2	107
6.11 Кластеризация точек временного ряда: а) для временного ряда Physical Motion 1; б) для временного ряда Physical Motion 2	107

6.12 Сегментация точек временного ряда Simple 1: а) сегментация временного ряда; б) проекция фазового пространства на первые две главные компоненты для первого кластера; в) проекция фазового пространства на первые две главные компоненты для второго кластера	108
6.13 Сегментация точек временного ряда Physical Motion 2: а) сегментация временного ряда; б) проекция фазового пространства на первые две главные компоненты для первого кластера; в) проекция фазового пространства на первые две главные компоненты для второго кластера	108

Список таблиц

1.1	Анализ роста числа параметров при развитии моделей глубокого обучения	9
2.1	Сводная таблица результатов вычислительного эксперимента	29
3.1	Сводная таблица результатов анализа байесовской дистилляции .	45
4.1	Качество аппроксимации мультимодели в зависимости от априорных распределений	58
5.1	Иллюстрация метода Белсли для анализа мультиколлинеарности параметров	74
5.2	Описание выборок для анализа метода задания порядка методом Белсли	76
5.3	Описание выборок, используемых в эксперименте по анализу метода задания порядка на основе анализа ковариационной матрицы градиентов	80
6.1	Сводная таблица методов определения оптимального размера выборки для линейных моделей	87
6.2	Описание выборок для анализа качества определения оптимального размера выборки	95
6.3	Эксперимент по оценке размера выборки для различных наборов выборок	95
6.4	Экспертные оценки гиперпараметров для разных методов оценки объема выборки	96
6.5	Описание временных рядов в эксперименте кластеризации точек временного ряда	103
6.6	Описание временных рядов в эксперименте сегментации временных рядов	106

Список литературы

1. *Sutskever I., Vinyals O., Le Q.* Sequence to Sequence Learning with Neural Networks // Advances in Neural Information Processing Systems. — Vol. 27. — Curran Associates, Inc., 2014.
2. Preconditioned stochastic gradient Langevin dynamics for deep neural networks / C. Li, C. Chen, D. Carlson, L. Carin // AAAI. — 2016. — Pp. 1788–1794.
3. *Krizhevsky A., Sutskever I., Hinton G.* ImageNet Classification with Deep Convolutional Neural Networks // Advances in Neural Information Processing Systems. — Vol. 25. — Curran Associates, Inc., 2012.
4. *Simonyan K., Zisserman A.* Very Deep Convolutional Networks for Large-Scale Image Recognition // ICLR. — 2015.
5. Deep Residual Learning for Image Recognition / K. He, X. Zhang, S. Ren, J. Sun // IEEE Conference on Computer Vision and Pattern Recognition. — 2016. — Pp. 770–778.
6. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M. Chang, K. Lee, K. Toutanova // NAACL-HLT. — Vol. 1. — Association for Computational Linguistics, 2019. — Pp. 4171–4186.
7. Attention is All you Need / A. Vaswani, N. Shazeer, N. Parmar et al. // Advances in Neural Information Processing Systems. — Vol. 30. — Curran Associates, Inc., 2017.
8. *Xue L., Constant N., Roberts A. et al.* mT5: A massively multilingual pre-trained text-to-text transformer. — 2021.
9. Language Models are Few-Shot Learners / T. Brown, B. Mann, N. Ryder et al. // Advances in Neural Information Processing Systems. — Vol. 33. — Curran Associates, Inc., 2020. — Pp. 1877–1901.
10. *Madala H., A. Ivakhnenko.* Inductive Learning Algorithms for Complex Systems Modeling. — Boca Raton: CRC Press, 1993. — 380 pp.
11. Adversarial Attacks and Defenses in Deep Learning / K. Ren, T. Zheng, Z. Qin, X. Liu // Engineering. — 2020. — Vol. 6, no. 3. — Pp. 346–360.
12. *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. — 2015.
13. *Vapnik V., Izmailov R.* Learning Using Privileged Information: Similarity Control and Knowledge Transfer // *J. Mach. Learn. Res.* — 2015. — Vol. 16, no. 1. — Pp. 2023–2049.
14. Unifying distillation and privileged information / D. Lopez-Paz, L. Bottou, B. Schölkopf, V. Vapnik // ICLR. — 2016.
15. *LeCun Y., Cortes C.* MNIST handwritten digit database. — <http://yann.lecun.com/exdb/mnist/>. — 2010. <http://yann.lecun.com/exdb/mnist/>.

16. *Huang Z., Wang N.* Like What You Like: Knowledge Distill via Neuron Selectivity Transfer. — 2017.
17. *Тамарчук А.* Байесовские методы опорных векторов для обучения распознаванию образов с управляемой селективностью отбора признаков: Ph.D. thesis / Вычислительный центр РАН. — 2014.
18. *Krizhevsky A., Nair V., Hinton G.* CIFAR-10 (Canadian Institute for Advanced Research). — <https://www.cs.toronto.edu/~kriz/cifar.html>. <https://www.cs.toronto.edu/~kriz/cifar.html>.
19. ImageNet: A large-scale hierarchical image database. / J. Deng, W. Dong, R. Socher et al. // CVPR. — IEEE Computer Society, 2009. — Pp. 248–255.
20. *MacLaurin D., Duvenaud D., Adams R.* Gradient-based Hyperparameter Optimization through Reversible Learning // Proceedings of the 32nd International Conference on Machine Learning. — JMLR Workshop and Conference Proceedings, 2015. — Pp. 2113–2122.
21. Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters / J. Luketina, T. Raiko, M. Berglund, K. Greff // Proceedings of the 33rd International Conference on Machine Learning. — Vol. 48. — JMLR.org, 2016. — Pp. 2952–2960.
22. *Molchanov D., Ashukha A., Vetrov D.* Variational Dropout Sparsifies Deep Neural Networks // Proceedings of the 34th International Conference on Machine Learning. — Vol. 70. — JMLR.org, 2017. — Pp. 2498–2507.
23. *Neal R.* Bayesian Learning for Neural Networks. — Berlin, Heidelberg: Springer-Verlag, 1996. — 289 pp.
24. *LeCun Y., Denker J., Solla S.* Optimal Brain Damage // Advances in Neural Information Processing Systems. — Vol. 2. — Morgan-Kaufmann, 1990.
25. *Louizos C., Ullrich K., Welling M.* Bayesian compression for deep learning // Advances in Neural Information Processing Systems. — 2017. — Pp. 3290–3300.
26. *Graves A.* Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems. — Vol. 24. — Curran Associates, Inc., 2011.
27. *Грабовой А. Б., Бахтееев О. Ю., Стрижов В. В.* Определение релевантности параметров нейросети // Информ. и её примен. — 2019. — Vol. 13, no. 2. — Pp. 62–70.
28. *Tibshirani R.* Regression Shrinkage and Selection Via the Lasso // Journal of the Royal Statistical Society. — 1994. — Vol. 58. — Pp. 267–288.
29. *Zou H., Hastie T.* Regularization and variable selection via the Elastic Net // Journal of the Royal Statistical Society. — 2005. — Vol. 67. — Pp. 301–320.
30. Dropout: A Simple Way to Prevent Neural Networks from Overfitting / N. Srivastava, G. Hinton, A. Krizhevsky et al. // J. Mach. Learn. Res. — 2014. — Vol. 15, no. 1. — Pp. 1929–1958.

31. Грабовой А. Б., Бахтееев О. Ю., Стрижсов В. В. Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информ. и её примен. — 2020. — Vol. 14, no. 2. — Pp. 58–65.
32. MacKay D. Information Theory, Inference & Learning Algorithms. — New York, NY, USA: Cambridge University Press, 2002. — 392 pp.
33. Bishop C. Pattern Recognition and Machine Learning (Information Science and Statistics). — Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. — 416 pp.
34. Ширяев А. Н. Вероятность. — М.: Наука, 1980. — 640 pp.
35. Кобзарь А. И. Прикладная математическая статистика: для инженеров и научных работников. — М.: Физматлит, 2012. — 813 pp.
36. Бахтееев О. Ю., Стрижсов В. В. Выбор моделей глубокого обучения субоптимальной сложности // Автомат. и телемех. — 2018. — Vol. 79, no. 8. — Pp. 129–147.
37. Mandt S., Hoffman M., Blei D. Stochastic Gradient Descent as Approximate Bayesian Inference // J. Mach. Learn. Res. — 2017. — Vol. 18, no. 1. — Pp. 4873–4907.
38. Kingma D., Ba J. Adam: A Method for Stochastic Optimization // ICLR. — 2015.
39. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — New York, NY, USA: Association for Computing Machinery, 2016. — Pp. 785–794.
40. Chen X., Ishwaran H. Random forests for genomic data analysis // Genomics. — 2012. — Vol. 99, no. 6. — Pp. 323–329.
41. Yuksel S., Wilson J., Gader P. Twenty Years of Mixture of Experts // IEEE Transactions on Neural Networks and Learning Systems. — 2012. — Vol. 23, no. 8. — Pp. 1177–1193.
42. Rasmussen C., Ghahramani Z. Infinite Mixtures of Gaussian Process Experts // Advances in Neural Information Processing Systems. — Vol. 14. — MIT Press, 2002.
43. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer / N. Shazeer, A. Mirhoseini, K. Maziarz et al. // 5th International Conference on Learning Representations. — OpenReview.net, 2017.
44. Jordan M., Jacobs R. Hierarchical mixtures of experts and the EM algorithm // Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan). — Vol. 2. — 1993. — Pp. 1339–1344.
45. Jordan M., Jacobs R. Hierarchies of Adaptive Experts // Proceedings of the 4th International Conference on Neural Information Processing Systems. — San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1991. — Pp. 985–992.

46. *Lima C., Coelho A., Von Zuben F.* Hybridizing Mixtures of Experts with Support Vector Machines: Investigation into Nonlinear Dynamic Systems Identification // *Inf. Sci.* — 2007. — Vol. 177, no. 10. — Pp. 2049–2074.
47. *Cao L.* Support vector machines experts for time series forecasting // *Neurocomputing*. — 2003. — Vol. 51. — Pp. 321–339.
48. *Yümlü M., Gürgen F., Okay N.* Financial Time Series Prediction Using Mixture of Experts // Lecture Notes in Computer Science. — Vol. 2869. — Springer, 2003. — Pp. 553–560.
49. *Cheung Y., Leung W., Xu L.* Application of Mixture of Experts Model to Financial Time Series Forecasting // Series Forecasting, submitted to International Conference on Neural Networks and Signal Processing. — 1995.
50. *Weigend A., Shi S.* Predicting Daily Probability Distributions of S&P500 Returns // *J. Forecast.* — 2000. — Vol. 19, no. 4. — Pp. 375–392.
51. Recognition of Persian handwritten digits using characterization Loci and mixture of experts / R. Ebrahimpour, M. Moradian, A. Esmkhani, F. Jafarlou // International Journal of Digital Content Technology and its Applications. — Vol. 3. — 2009.
52. *Estabrooks A., Japkowicz N.* A mixture-of-experts framework for text classification // Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL). — 2001.
53. A bayesian hierarchical mixture of experts approach to estimate speech quality / S. Mossavat, O. Amft, B. de Vries et al. // 2010 Second International Workshop on Quality of Multimedia Experience (QoMEX). — 2010. — Pp. 200–205.
54. *Peng F., Jacobs R., M. Tanner.* Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models with an Application to Speech Recognition // *Journal of the American Statistical Association*. — 1996. — Vol. 91, no. 435. — Pp. 953–960.
55. *Tuerk A.* The State Based Mixture of Expert HMM with Applications to the Recognition of Spontaneous Speech: Ph.D. thesis / University of Cambridge. — 2001.
56. *Sminchisescu C., Kanaujia A., Metaxas D.* BM³E : Discriminative Density Propagation for Visual Tracking // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. — 2007. — Vol. 29, no. 11. — Pp. 2030–2044.
57. *Bowyer K., Hollingsworth K., Flynn P.* A Survey of Iris Biometrics Research: 2008–2010. — London: Springer, 2013. — 410 pp.
58. *Matveev I.* Detection of Iris in image by Interrelated Maxima of Brightness Gradient Projections // *Appl.Comput. Math.* — Vol. 9. — 2010. — Pp. 252–257.
59. *Matveev I., Simonenko I.* Detecting precise iris boundaries by circular shortest path method // *Pattern Recognition and Image Analysis*. — 2014. — Vol. 24. — Pp. 304–309.

60. *Xiao H., Rasul K., Vollgraf R.* Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. — 2017. <https://arxiv.org/abs/1708.07747>.
61. SemEval–2013 Task 2: Sentiment Analysis in Twitter / P. Nakov, S. Rosenthal, Z. Kozareva et al. // Proceedings of the 7th International Workshop on Semantic Evaluation. — The Association for Computer Linguistics, 2013. — Pp. 312–320.
62. Backpropagation Applied to Handwritten Zip Code Recognition / Y. LeCun, B. Boser, J. S. Denker et al. // *Neural Comput.* — 1989. — Vol. 1, no. 4. — Pp. 541–551.
63. *Hochreiter S., Schmidhuber J.* Long Short-Term Memory // *Neural Computation*. — 1997. — Vol. 9, no. 8. — Pp. 1735–1780.
64. Deep Generative Models for Fast Shower Simulation in ATLAS / D. Salamani, S. Gadatsch, T. Golling et al. // 2018 IEEE 14th International Conference on e-Science (e-Science). — 2018. — Pp. 348–348.
65. *Dempster A., Laird N., Rubin D.* Maximum likelihood from incomplete data via the EM algorithm // *Journal of the Royal Statistical Society*. — 1977. — Vol. 39. — Pp. 1–38.
66. *Neychev R., Katrutsa A., Strijov V.* Robust selection of multicollinear features in forecasting // *Factory Laboratory*. — 2016. — Vol. 82(3). — Pp. 68–74.
67. *Aeberhard S.* Wine Dataset. — <https://archive.ics.uci.edu/ml/datasets/Wine>.
68. *Harrison D., Rubinfeld D.* Hedonic housing prices and the demand for clean air // *Journal of environmental economics and management*. — 1978. — Vol. 5, no. 1. — Pp. 81–102.
69. *Self S., Mauritsen R.* Power sample size calculations for generalized linear models // *Biometrics*. — 1988. — Vol. 44. — Pp. 79–86.
70. *Self S., Mauritsen R., Ohara J.* Power calculations for likelihood ratio tests in generalized linear models // *Biometrics*. — 1992. — Vol. 48. — Pp. 31–39.
71. *Shieh G.* On power and sample size calculations for likelihood ratio tests in generalized linear models // *Biometrics*. — 2000. — Vol. 56. — Pp. 1192–1196.
72. *Demidenko E.* Sample size determination for logistic regression revisited // *Statistics in medicine*. — 2006. — Vol. 26. — Pp. 3385–97.
73. *Shieh G.* On power and sample size calculations for Wald tests in generalized linear models // *Journal of Statistical Planning and Inference*. — 2005. — Vol. 128. — Pp. 43–59.
74. *Motrenko A., Strijov V., Weber G.* Sample Size Determination for Logistic Regression // *J. Comput. Appl. Math.* — 2014. — Vol. 255, no. C. — Pp. 743–752.
75. *Qumsiyeh M.* Using the bootstrap for estimation the sample size in statistical experiments // *Journal of modern applied statistical methods*. — 2013. — Vol. 8. — Pp. 305–321.

76. Lawrence J., Wolfson D., Berger R. Sample Size Calculations for Binomial Proportions Via Highest Posterior Density Intervals // *Statistician*. — 1995. — Vol. 44. — Pp. 143–154.
77. Joseph L., Berger R., Bélisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination // *Statistician*. — 1997. — Vol. 16, no. 7. — Pp. 769–781.
78. Lindley D. The choice of sample size // *Statistician*. — 1997. — Vol. 46. — Pp. 129–138.
79. Kloek T. Note on a large-sample result in specification analysis // *Econometrica*. — 1975. — Vol. 43. — Pp. 933–936.
80. Rubin D., Stern H. Sample size determination using posterior predictive distributions // *Sankhya: The Indian Journal of Statistics Special Issue on Bayesian Analysis*. — 1998. — Vol. 60. — Pp. 161–175.
81. Wang F., Gelfand A. A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models // *Statistical Science*. — 2002. — Vol. 17. — Pp. 193–208.
82. Quinlan J. Learning With Continuous Classes. — World Scientific, 1992. — Pp. 343–348.
83. Kwapisz J., Weiss G., Moore S. Activity recognition using cell phone accelerometers // *ACM SigKDD Explorations Newsletter*. — 2011. — Vol. 12, no. 2. — Pp. 74–82.
84. Revisiting Optimal Delaunay Triangulation for 3D Graded Mesh Generation / Z. Chen, W. Wang, B. Lévy et al. // *SIAM Journal on Scientific Computing*. — 2014. — Pp. 930–954.
85. Ignatov A., Strijov V. Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer // *Multim. Tools Appl.* — 2016. — Vol. 75, no. 12. — Pp. 7257–7270.
86. Detection of (In)activity Periods in Human Body Motion Using Inertial Sensors: A Comparative Study / A. Olivares, J. Ramírez, J. Górriz et al. // *Sensors*. — 2012. — Vol. 12, no. 5. — Pp. 5791–5814.
87. Period-aware content attention RNNs for time series forecasting with missing values / Y. Cinar, H. Mirisaee, P. Goswami et al. // *Neurocomputing*. — 2018. — Vol. 312. — Pp. 177–186.
88. Motrenko A., Strijov V. Extracting Fundamental Periods to Segment Biomedical Signals // *IEEE journal of biomedical and health informatics*. — 2015. — Vol. 20.
89. Данилова Д. Л., Жигловский А. А. Главные компоненты временных рядов: метод “Гусеница”. — Санкт-Петербург: Санкт-Петербургский университет, 1997. — 218 pp.
90. Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. — Москва: Финансы и статистика, 2003. — 416 pp.

91. Ивкин И. П., Кузнецов М. П. Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию // *Машинное обучение и анализ данных*. — 2015. — Vol. 1, no. 11. — Pp. 1471–1483.
92. Katrutsa A., Strijov V. Stress test procedure for feature selection algorithms // *Chemometrics and Intelligent Laboratory Systems*. — 2015. — Vol. 142. — Pp. 172–183.
93. Borg I., Groenen P. Modern Multidimensional Scaling: Theory and Applications. — Springer, 2005. — 380 pp.
94. Grabovoy A., Strijov V. Quasi-Periodic Time Series Clustering for Human Activity Recognition // *Lobachevskii Journal of Mathematics*. — 2020. — Vol. 41, no. 3. — Pp. 333–339.
95. Grabovoy A., Strijov V. Prior Distribution Selection for a Mixture of Experts // *Computational Mathematics and Mathematical Physics*. — 2021. — Vol. 61, no. 7. — Pp. 1140–1152.
96. Akhtar N., Mian A. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey // *IEEE Access*. — 2018. — Vol. 6. — Pp. 14410–14430.
97. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review / H. Xu, Y. Ma, H. Liu et al. // *International Journal of Automation and Computing*. — 2020. — Vol. 17, no. 2. — Pp. 151–178.
98. Ribeiro M., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — New York, NY, USA: Association for Computing Machinery, 2016. — Pp. 1135–1144.
99. TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing / Z. Yang, Y. Cui, Z. Chen et al. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. — Association for Computational Linguistics, 2020. — Pp. 9–16.