Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Бахтеев Олег Юрьевич

Последовательное порождение моделей глубокого обучения оптимальной сложности

010990 — Математические и информационные технологии

Выпускная квалификационная работа магистра

Научный руководитель:
д. ф.-м. н. Стрижов Вадим Викторович

Москва
2016

Содержание

Аннотация

В работе рассматривается проблема выбора оптимальной модели глубокого обучения и оптимизации ее параметров. Принимается статистическая гипотеза о распределении зависимой переменной и параметров модели. Каждая рассматриваемая модель декомпозируется на порождающую и разделяющую. На основании статистических предположений принимается оптимизируемая функция ошибки для каждой подмодели. Предлагаются методы оптимизации полученных функций ошибок. В качестве рассматриваемой порождающей модели используется вариационный автокодировщик. В качестве рассматриваемой разделяющей модели используется нейронная сеть с одним скрытым слоем. Для выбора разделяющей модели и контроля переобучения без использования скользящего контроля предлагается алгоритм, основанный на стохастическом градиентном спуске. По предложенному алгоритму выбора модели глубокого обучения был проведен эксперимент на выборке изображений рукописных цифр MNIST.

Ключевые слова: выбор оптимальной модели; глубокое обучение; байесовский вывод; вариационный вывод.

Введение

Актуальность темы. Задача выбора модели является одной из ключевых задач машинного обучения. Проблема выбора моделей глубокого обучения является вычислительно сложной в силу большого количества параметров, и как следствие, большой вычислительной стоимости их оптимизации.

Достаточно популярной эвристикой для выбора модели является послойный выбор и жадное предобучение каждого слоя модели с дальнейшим дообучением [1]. Одной из проблем при подобной стратегии является избыточность параметров полученной модели, влияющая на устойчивость модели в целом [2, 3].

В данной работе также предлагается послойное построение модели, при этом на каждом слое происходит выбор подмодели, доставляющей максимальное значение правдоподобия модели, которое характеризует сложность сети. На этапе дообучения предлагается использовать метод стохастического градиентного спуска с контролем энтропии, позволяющий определить начало переобучения сети глубокого обучения. Данный метод позволяет отказаться от использования скользящего контроля в случае высокой вычислительной сложности оптимизации.

Цель работы. Целью данной работы является получение метода автоматического построения моделей глубокого обучения субоптимальной сложности, т.е. моделей, имеющих приемлемое качество при небольшом количестве избыточных параметров модели.

Методы исследования. Для достижения поставленной цели предлагается декомпозировать модель на порождающую и разделяющую. В качестве функционалов качества для каждой из них выступает вариационная нижняя оценка интеграла правдоподобия модели [4, 5].

Основные положения, выносимые на защиту.

1. Критерий субоптимальной сложности модели классификации.

2. Исследование зависимости правдоподобия модели от устойчивости модели и возможности переобучения.

3. Алгоритм выбора субоптимальной модели классификации без использования кросс-валидации.

4. Теорема об энтропии распределения под действием градиентного спуска.

Научная новизна. Разработан метод получения моделей субоптимальной сложности для моделей глубокого обучения. Доказана теорема, позволяющая использовать стохастический градиентный спуск для получения вариационной нижней оценки правдоподобия модели.

Практическая значимость. Предложенный в работе алгоритм позволяет получать модели глубокого обучения, имеющие низкую структурную сложность при приемлемом качестве классификации без использования кросс-валидации.

Степень достоверности и апробация работы. Достоверность результатов подтверждена экспериментальной проверкой полученных методов.

Публикации по теме дипломной работы.

1. Бахтеев О.Ю., Попова М.С., Стрижов В.В. Системы и средства глубокого обучения в задачах классификации // Системы и средства информатики, 2016, 2.

Обзор литературы. Одна из проблем построения моделей глубокого обучения — большое количество параметров моделей [6, 7]. Поэтому задача выбора моделей глубокого обучения включает в себя выбор стратегии построения модели, эффективной по вычислительным ресурсам. В работе [8] приводятся некоторые теоретические оценки построения нейросетей с использованием жадных стратегий. В работе [9] предлагается жадная стратегия выбора модели нейросети с использованием релевантных априпорных распределений. Данный метод был также применен к задаче построения модели метода релевантных векторов [10]. Альтернативой данным алгоритмам построения моделей являются методы, основанные на прореживании сетей глубокого обучения [3, 11]. Популярным методом построения моделей глубокого обучения является жадное послойное построение модели с отдельным критерием оптимизации для каждого слоя [1, 12]. В ряде работ [13–15] предлагается декомпозиция

модели на порождающую и разделяющую, оптимизируемых последовательно.

В качестве порождающих моделей в сетях глубокого обучения могут выступать ограниченные машины Больцмана [6] и автокодировщики [16]. В работе [17] рассматриваются некоторые типы регуляризации автокодировщиков, позволяющие формально рассматривать данные модели как порождающие модели с использованием байесового вывода. В работе [18] также рассматриваются регуляризованные автокодировщики и свойства оценок их правдоподобия. В работе [19] предлагается обобщение автокодировщика с использованием вариационного байесовского вывода [5]. В работе [20] рассматриваются модификации вариационного автокодировщика и ступенчатых сетей (англ. ladder network) [21] для случая построения многослойных порождающих моделей.

В качестве критерия выбора модели в ряде работ [4, 5, 22–25] выступает правдоподобие модели. В работах [22–25] рассматривается проблема выбора модели и оценки гиперапараметров в задачах регрессии. Альтернативным критерием выбора модели является минимальная длина описания [26], являющаяся показателем статистической сложности модели и заданной выборки. В работе [27] рассматривается перечень критериев сложности моделей глубокого обучения и их взаимосвязь. В работе [28] в качестве критерия сложности модели выступает показатель нелинейности, характеризуемый степенью полинома Чебышева, аппроксимирующего функцию. В работе [2] анализируется показатель избыточности параметров сети. Утверждается, что по небольшому набору параметров в глубокой сети с большим количеством избыточных параметров можно спрогнозировать значения остальных. В работе [29] рассматривается показатель робастности моделей, а также его взаимосвязь с топологией выборки и классами функций, в частности рассматривается влияние функции ошибки и ее липшицевой константы на робастность моделей. Схожие идеи были рассмотрены в работе [30], в которой исследуется устойчивость классификации модели под действием шума.

Одним из методов получения приближенного значения интеграла правдоподобия является вариационный метод получения нижней оценки интеграла [5]. В работе [31] рассматривается стохастическая версия вариационного метода. В работе [32] рассматривается алгоритм получения вариационной нижней оценки правдоподобия

для оптимизации гиперпараметров моделей глубокого обучения. В работе [33] рассматривается взаимосвязь градиентных методов получения вариационной нижней оценки интеграла с методом Монте-Карло. В работе [34] рассматривается стохастический градиентный спуск в качестве оператора, порождающего распределение, аппроксимирующее апостериорное распределение параметров модели. В данной работе отмечается, что стохастический градиентный спуск не оптимизирует вариационную оценку интеграла правдоподобия и приближает ее только до некоторого количества итераций оптимизации. Схожий подход рассматривается в работе [35], где также рассматривается стохастический градиентный спуск в качестве оператора, порождающего апостериорное распределение параметров. В работе [36] предлагается модификация стохастического градиентного спуска, аппроксимирующая апостериорное распределение.

Альтернативным методом выбора модели является выбор модели на основе скользящего контроля [22,37]. Проблемой такого подхода является возможная высокая вычислительная сложность [38,39]. В работах [40,41] рассматривается проблема смещения оценок качества модели и гиперпараметров, получаемых при использовании k-fold метода скользящего контроля, при котором выборка делится на k-частей с обучением на $k-1$ части и валидацией результата на оставшейся части выборки.

Задачей, связанной с проблемой выбора модели, является задача оптимизации гиперпараметров [4, 5]. В работе [22] рассматривается оптимизация гиперпарамтров с использованием метода скользящего контроля и методов оптимизации интеграла правдодобия моделей, отмечается низкая скорость сходимости гиперпараметров при использовании метода скользящего контроля. В ряде работ [42,43] рассматриваются градиентные методы оптимизации гиперпараметров, позволяющие оптимизировать большое количество гиперпараметров одновременно. В работе [42] предлагается метод оптимизации гиперпараметров с использованием градиентного спуска с моментом, в качестве оптимизируемого функционала рассматривается ошибка на валидационной части выборки. В работе [35] рассматривается задача оптимизации параметров градиентного спуска с использованием нижней вариационной оценки интеграла правдоподобия. В работе [34] отмечается возможность использовать градиентный метод для оптимизации гиперпараметров с использованием вариационной нижней

оценки интеграла правдоподобия в качестве оптимизируемого функционала.

1 Постановка задачи

Пусть задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, N, \tag{1}$$

состоящая из множества пар «объект - класс», $\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^n, y_i \in \mathbf{Y}$.

Каждый объект $\mathbf{x}_i$ принадлежит одному из Z классов с меткой y_i.

Сетью глубокого обучения $\mathbf{f}$ назовем суперпозицию функций [44]

$$\mathbf{f}(\mathbf{w}, \mathbf{x}) = \mathbf{f}_1(\mathbf{f}_2(\ldots \boldsymbol{f}_K(\mathbf{x}))) : \mathbb{R}^n \to [0,1]^Z, \tag{2}$$

где $\mathbf{f}_k$, — модели, параметрическое семейство вектор-функций, $k \in \{1, \ldots, K\}$; $\mathbf{w}$ — вектор параметров моделей;

c-я компоненту вектора $\mathbf{f}(\mathbf{x}, \mathbf{w})$ — вероятность отнесения объекта $\mathbf{x}_i$ к классу с меткой c.

Множество всех рассматриваемых моделей обозначим за $\mathfrak{F}$. Будем полагать, что для каждой модели $\mathbf{f} \in \mathfrak{F}$ задано априорное распределение параметров $p(\mathbf{w}|\mathbf{f})$.

Определение Модель классификации $\mathbf{f}$ назовем оптимальной среди моделей $\mathfrak{F}$, если достигается максимум интеграла [4]:

$$p(\mathfrak{D}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w}) p(\mathbf{w}|\mathbf{f}) d\mathbf{w}. \tag{3}$$

Требуется найти оптимальную модель $\mathbf{f}$ среди заданного множества моделей $\mathfrak{F}$, а также значения ее параметров $\mathbf{w}$, доставляющие максимум апостериорной вероятности:

$$p(\mathbf{w}|\mathfrak{D}, \mathbf{f}) \sim p(\mathfrak{D}|\mathbf{w}, \mathbf{f}) p(\mathbf{w}|\mathbf{f}).$$

1.1 Поиск субоптимальной модели

В качестве функционала качества, приближающего логарифм интеграла (3), будем рассматривать его вариационную нижнюю границу, полученную при помощи неравенства Йенсена [5]:

$$\log p(\mathfrak{D}|\mathbf{f}) \geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} = \tag{4}$$

$$= -\mathrm{D}_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) + \int_{\mathbf{w}} q(\mathbf{w})\log\, p(\mathfrak{D}|\mathbf{w},\mathbf{f})d\mathbf{w},$$

где $\mathrm{D}_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) = -\int_{\mathbf{w}} q(\mathbf{w})\log\, \frac{p(\mathbf{w}|\mathbf{f})}{q(\mathbf{w})}d\mathbf{w}$.

Декомпозируем модель $\mathbf{f}$ на порождающую $\mathbf{f}_G$ и разделяющую $\mathbf{f}_D$ [13–15]. Будем полагать, что выборка $\mathbf{X}$ была порождена некоторой случайной величиной $\mathbf{z}$ [19, 45]. Положим решающее правило классификации:

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}_D(\arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x})),$$

где p — условная вероятность реализации случайной величины $\mathbf{z}$ при условии наблюдения $\mathbf{x}$ и порождающей функции f_G, f_D — разделяющая функция.

Поиск наилучшей модели будем осуществлять последовательно на некотором подмножестве моделей.

Определение Порождающую модель $\mathbf{f}_G$ назовем субоптимальной на множестве порождающих моделей $\mathfrak{F}_G$ по семейству распределений Q, если модель доставляет максимум нижней вариационной оценке интеграла [13]:

$$\max_{q\in Q} \int_{\mathbf{w}} q(\mathbf{w})\log\, \frac{p(\mathbf{X},\mathbf{w}|\mathbf{f}_G)}{q(\mathbf{w})}d\mathbf{w}. \tag{5}$$

В качестве множества моделей $\mathfrak{F}_G$ будем рассматривать модели вариационных автокодировщиков [19].

Определение Разделяющую модель $\mathbf{f}_D$ назовем субоптимальной для модели $\mathbf{f}_G$ на множестве разделяющих моделей $\mathfrak{F}_D$ по семейству распределений Q, если модель доставляет максимум нижней вариационной оценке интеграла:

$$\max_{q\in Q} \int_{\mathbf{w}} q(\mathbf{w})\log\, \frac{p(\mathbf{Y},\mathbf{w}|\mathbf{f}_D,\hat{\mathbf{Z}})}{q(\mathbf{w})}d\mathbf{w}, \tag{6}$$

где $\hat{\mathbf{Z}} = \mathrm{argmax}_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X})$, $\mathbf{Z}$ — скрытое представление выборки $\mathbf{X}$, порожденное моделью $\mathbf{f}_G$.

Определение Модель классификации $\mathbf{f}$ назовем субоптимальной, если она получена композицией субоптимальных моделей порождения и разделения.

Задача нахождения субоптимальной модели классификации $\mathbf{f}$ сводится к следующим подзадачам:

- нахождение субоптимальной модели порождения $\mathbf{f}_G$;
- нахождение оптимальных параметров $\mathbf{w}_G$ модели порождения $\mathbf{f}_G$;
- нахождение субоптимальной модели разделения $\mathbf{f}_D$;
- нахождение оптимальных параметров $\mathbf{w}_D$ модели порождения $\mathbf{f}_D$.
- дообучение всех параметров модели с единым критерием оптимизации [1]:

$$p(\mathbf{Y}, \mathbf{w}|\mathbf{X}, \mathbf{f}) \to \max$$

.

Заметим, что декомпозиция модели $\mathbf{f}$ на порождающую $\mathbf{f}_G$ и разделяющую $\mathbf{f}_D$ является обоснованной только в случае, когда распределение $p(\mathbf{x})$ содержит достаточно информации о принадлежности объектов $\mathbf{x} \in \mathbf{X}$ к соответствующим классам $\mathbf{Y}$ [1].

2 Рассматриваемые модели

2.1 Однослойная softmax-сеть

Однослойная сеть представляет собой логистическую вектор-функцию:

$$\mathbf{a}(\mathbf{x}) = \mathbf{w}_2^{\mathsf{T}} \boldsymbol{\sigma}(\mathbf{w}_1^{\mathsf{T}} \mathbf{x}), \tag{7}$$

$$\mathbf{f}_{\mathrm{SM}}(\mathbf{x}) = \frac{\mathbf{exp}\big(\mathbf{a}(\mathbf{x})\big)}{\sum_{j=1}^{Z} \exp\big(a_j(\mathbf{x})\big)},$$

где c-я компонента вектора $\mathbf{f}_{\mathrm{SM}}(\mathbf{x})$ интерпретируется как вероятность принадлежности объекта $\mathbf{x}$ классу c, $\boldsymbol{\sigma}$ — нелинейная функция. Итоговая функция классификации (2) ставит в соответствие объекту $\mathbf{x}$ метку класса y, где y — класс, к которому принадлежит $\mathbf{x}$ с наибольшей вероятностью:

$$f(\mathbf{w}, \mathbf{x})(c) = \begin{cases} 1, \text{ если } c = \arg\max_{c'} f_{\mathrm{SM}}(\boldsymbol{f}_1(\boldsymbol{f}_2(\ldots \boldsymbol{f}_K(\mathbf{x}))))(c'), \\ 0 \text{ иначе.} \end{cases}$$

где $f_{\mathrm{SM}}(\mathbf{x})(c)$ — c-я компонента вектора $\mathbf{f}_{\mathrm{SM}}$.

Итоговая задача оптимизации однослойной нейросети выглядит следующим образом:

$$\hat{\mathbf{w}} = \arg\min \sum_{\mathbf{x},y \in \mathfrak{D}} \sum_{c=1}^{Z} [y = c] \log(f_{\mathrm{SM}}(\mathbf{x})(c)),$$

где $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_2^{\mathsf{T}}, \hat{\mathbf{w}}_1^{\mathsf{T}}]$ — параметры сети (7).

2.2 Вариационный автокодировщик

Автокодировщик Автокодировщик — модель, предназначенная для снижения размерности исходного пространства признаков. Автокодировщик состоит из кодирующего блока

$$\mathbf{z} = \boldsymbol{\sigma}(\mathbf{w}_e \mathbf{x} + \mathbf{b}_e).$$

и декодирующего блока

$$r(\mathbf{x}) = \boldsymbol{\sigma}(\mathbf{w}_r^{\mathrm{T}} \mathbf{z} + \mathbf{b}_r).$$

Оптимизацию параметров модели $\mathbf{w} = [\mathbf{w}_e, \mathbf{w}_r, \mathbf{b}_e, \mathbf{b}_r]$ проводят таким образом, чтобы по образу вектора $\mathbf{x}$, получаемому с помощью кодирующего блока, можно было получить вектор $r(\mathbf{x})$, близкий к исходному входному $\mathbf{x}$:

$$||r(\mathbf{x}) - \mathbf{x}||_2^2 \to \min.$$

В работе [19] была предложена модификация автокодировщика, имеющая байесовскую интерпретацию. Будем полагать, что объекты $\mathbf{x} \in \mathbf{X}$ порождены при условии случайной величины $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{z}, \mathbf{w}).$$

Аппроксимируем неизвестное распределение $p(\mathbf{z}|\mathbf{x}, \mathbf{w})$ распределением $q_\phi(\mathbf{z}|\mathbf{x})$. Для нахождения правдоподобия данных при условии параметров $\mathbf{w}$ применим вариационную оценку (4). Решая задачу оптимизации найдем q_ϕ и $p(\mathbf{x}|\mathbf{z}, \mathbf{w})$:

$$\log p(\mathbf{x}|\mathbf{w}) \geq \mathsf{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log\, p(\mathbf{x}|\mathbf{z}, \mathbf{w}) - \mathrm{D}_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \to \max.$$

Параметризация распределений q_ϕ и $p(\mathbf{x}|\mathbf{z},\mathbf{w})$. В работе [19] в качестве используемой параметризации для распределений q_ϕ и $p(\mathbf{x}|\mathbf{z},\mathbf{w})$ предлагается использовать выходы многослойной нейросети:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x})),$$

$$p(\mathbf{x}|\mathbf{z},\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}(\mathbf{z}), \boldsymbol{\sigma}^2_\mathbf{w}(\mathbf{z})).$$

Оптимизацию нижней оценки будем проводить с использованием метода Монте-Карло. Нижняя оценка правдоподобия принимает вид:

$$\log \hat{p}(\mathbf{X}|\mathbf{w}) \simeq \frac{1}{2}\sum_{i=1}^{m}\Big(\sum_{j=1}^{n}\big(1 + \log(\sigma_{\phi,j}(\mathbf{x}_i)^2) - \mu_{\phi,j}(\mathbf{x}_i)^2 - \sigma_{\phi,j}(\mathbf{x}_i)^2\big) + \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(\mathbf{x}_i|\mathbf{z}^{i,l})\Big), \tag{8}$$

где L — число реализаций случайной величины $\mathbf{z}$ для каждого объекта выборки $\mathbf{x} \in \mathbf{X}$, $\mathbf{z}^{i,l} = \boldsymbol{\mu}(\mathbf{x}_i) + \boldsymbol{\sigma}(\mathbf{x}_i) \odot \mathbf{e}$, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\sigma_{\phi,j}$ — j-я компонента вектора $\boldsymbol{\sigma}_\phi$.

3 Вариационная нижняя оценка правдоподобия модели

3.1 Оценка для вариационного автокодировщика

Для вычисления вариационной оценки правдоподобия модели $\mathbf{f}_G$ применим метод Монте-Карло, порождая реализации значений параметров $\mathbf{w}$ из распределения $q_\mathbf{w}$. Вариационная оценка правдоподобия модели для вариационного кодировщика получается следующим образом [19]:

$$\log p(\mathbf{x}|\mathbf{f}) = \log \int_\mathbf{w}\int_\mathbf{z} p(\mathbf{x}|\mathbf{z},\mathbf{w},\mathbf{f})p(\mathbf{w}|\mathbf{f})p(\mathbf{z})d\mathbf{w}d\mathbf{z} \geq$$

$$\int_\mathbf{w} q_\mathbf{w}(\mathbf{w})\big(\log p(\mathbf{x}|\mathbf{w},\mathbf{f}) + \log p(\mathbf{w}|\mathbf{f}) - \log q_\mathbf{w}(\mathbf{w})\big)d\mathbf{w}. \tag{9}$$

Правдоподобие модели с учетом всей выборки вычисляется как сумма интегралов (9) по всем объектам выборки $\mathbf{x} \in \mathbf{X}$:

$$\log \hat{p}(\mathbf{X}|\mathbf{f}) \simeq \log \hat{p}(\mathbf{X}|\hat{\mathbf{w}}) - \mathrm{D}_{KL}(q(\mathbf{w})||p(\mathbf{w})),$$

$\mathbf{x} \in \mathbf{X}, \hat{\mathbf{w}}$ — реализация случайной величины из распределения $q_\mathbf{w}$, $\log \hat{p}$ — вариационная оценка правдоподобия (8).

Заметим, что полученная оценка является вариационным приближением как по распределению параметров модели $\mathbf{w}$, так и по неизвестному распределению $p(\mathbf{z}|\mathbf{x})$. В случае, если распределения $q_{\mathbf{w}}, q_\phi$ — гауссовые, слагаемые $\mathrm{D}_{KL}(q(\mathbf{w})||p(\mathbf{w}))$, $\mathrm{D}_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$ вычисляются аналитически [19].

3.2 Оценка с использованием градиентного спуска

Представим интеграл (6) в виде:

$$\log p(\mathbf{Y}|\hat{\mathbf{Z}},\mathbf{f}) \geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{Y},\mathbf{w}|\hat{\mathbf{Z}},\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} = \mathsf{E}_{q(\mathbf{w})}[\log p(\mathbf{Y},\mathbf{w}|\hat{\mathbf{Z}},\mathbf{f})] - \mathsf{S}(q(\mathbf{w})),$$

где S — энтропия распределения q. Рассмотрим параметрическое множество распределений, получаемых из некоторого начального распределения q^0 под действием оператора T:

$$q^\tau = T(q^{\tau-1}).$$

В качестве оператора T можно использовать градиентный спуск [34]:

$$\Delta\mathbf{w} = \alpha \nabla L(\hat{\mathbf{Z}}),$$

где α — шаг градиентного спуска, L — дифференцируемая функция потерь.

Будем проводить оптимизацию в режиме мультистарта [46], т.е. запускать несколько оптимизаций из разных начальных приближений. Значения параметров из каждого приближения будем интерпретировать как реализацию случайной величины из распределения q^τ. Докажем и обобщим утверждение, представленное в работе [34].

Теорема. Пусть L — функция потерь, градиент которой — непрерывно-дифференцируемая функция с константой Липшица C. Пусть $\mathbf{w}^1, \ldots, \mathbf{w}^r$ — начальные приближения оптимизации модели. Пусть α — шаг градиентного спуска, такой что:

- $\alpha < \frac{1}{C}$,
- $\alpha^{(-1)} > \max_{\gamma \in \{1,\ldots,r\}} \lambda_{\max}(\mathbf{H}(\mathbf{w}^\gamma))$,

где $\lambda_{\max}$ — наибольшее по модулю собственное значение гессиана функции потерь $\mathbf{H}$.

Тогда

$$\mathcal{S}(q^{\tau}(\mathbf{w})) - \mathsf{S}(q^{\tau-1}(\mathbf{w})) \sim \frac{1}{r}\sum_{\gamma=1}^{r}\big(-\alpha Tr[\mathbf{H}(\mathbf{w}^{\gamma})] - \alpha^2 Tr[\mathbf{H}(\mathbf{w}^{\gamma})\mathbf{H}(\mathbf{w}^{\gamma})]\big) + o_{\alpha\to 0}(1), \quad (10)$$

где $\mathbf{H}$ — гессиан функции потерь L.

Доказательство. Рассмотрим оптимизацию на некотором шаге τ:

$$\Delta\mathbf{w} = \alpha\nabla L(\mathbf{X}), \quad \mathbf{w} \sim q^{\tau-1}.$$

Если $\alpha < \frac{1}{C}$, то оператор градиентного спуска T является биекцией [47]. Тогда справедлива следующая формула для энтропии [48]:

$$\mathsf{S}(q^{\tau}(\mathbf{w})) - \mathsf{S}(q^{\tau-1}(\mathbf{w})) = \int_{\mathbf{w}} q(\mathbf{w})\log|\frac{\partial T(\mathbf{w})}{\partial \mathbf{w}}|d\mathbf{w}.$$

По усиленному закону больших чисел

$$\mathcal{S}(q^{\tau}(\mathbf{w})) - \mathsf{S}(q^{\tau-1}(\mathbf{w})) \sim \frac{1}{r}\sum_{\gamma=1}^{r}\log|\frac{\partial T(\mathbf{w}^{\gamma})}{\partial \mathbf{w}}|.$$

Логарифм якобиана $\log|\frac{\partial T(\mathbf{w}^{\gamma})}{\partial \mathbf{w}}|$ оператора T можно записать как [34]:

$$\log|\frac{\partial T(\mathbf{w}^{\gamma})}{\partial \mathbf{w}}| = \log|\mathbf{I} - \alpha\mathbf{H}| = \sum_{i=1}^{|\mathbf{w}|}\log(1-\alpha\lambda_i),$$

где λ_i — i-е собственное значение гессиана $\mathbf{H}$.

Т.к. $|\alpha\lambda_i| < \alpha\lambda_{\max} < 1$, то полагая значения λ_i фиксированными последнее выражение можно разложить в ряд Тейлора:

$$\sum_{t=1}^{|\mathbf{w}|}\log(1-\alpha\lambda_i) = -\alpha Tr[\mathbf{H}] - \alpha^2 Tr[\mathbf{H}(\mathbf{w}^{\gamma})\mathbf{H}(\mathbf{w}^{\gamma})] + o_{\alpha\to 0}(1).$$

Т.к. точек мультистарта r конечное число, вынесем $o_{\alpha\to 0}(1)$ за скобки. Теорема доказана. □

Заметим, что в качестве оператора T также можно использовать псевдослучайный стохастический градиентный спуск:

$$\Delta\mathbf{w} = \alpha\nabla L(\hat{\mathbf{X}}),$$

где $\hat{\mathbf{X}}$ — случайная подвыборка выборки $\mathbf{X}$, одинаковая для всех точек мультистарта.

Основной проблемой данного метода оценки интеграла (6) является недостаточная аппроксимация исходного распределения $p(\mathbf{w}|\mathbf{f},\mathfrak{D})$. Градиентный спуск не минимизирует дивергенцию $D_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f},\mathfrak{D}))$, поэтому оценка интеграла (4) может быть существенно заниженной. При этом, при приближении к точке экстремума снижается вариационная оценка интегральной функции правдоподобия, что интерпретируется как возможное начало переобучения [34].

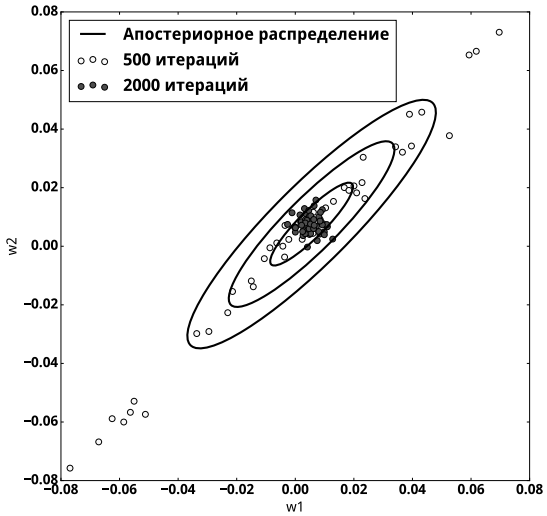
Приведем иллюстративный пример аппроксимации апостериорного распределения с помощью градиентного спуска. На рис. 1 изображена аппроксимация апостериорного распределения $p(\mathbf{w}|\mathbf{X},\mathbf{f})$ по выборке $\mathbf{X}$ порожденной моделью:

$$\mathbf{X} = \mathcal{N}(\mathbf{w}, \mathbf{A}), \quad \mathbf{w} = \mathbf{0}, \tag{11}$$

$$\mathbf{A} = \begin{pmatrix} 2 & 1.8 \\ 1.8 & 2 \end{pmatrix}.$$

В качестве априорного распределения параметров $\mathbf{w}$ было выбрано стандартное распределение: $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Как видно из графика, градиентный спуск сходится к моде распределения, однако при небольшом количестве итераций может аппроксимировать апостериорное распределение [34, 35].

Рис. 1: Аппроксимация распределения стохастическим градиентным спуском



3.3 Оценка с использованием стохастической динамики Ланжевина

Для более точной вариационной оценки интеграла (6) будем использовать стохастическую динамику Ланжевина [36]. Стохастическая динамика Ланжевина представляет собой вариант градиентного спуска с добавлением гауссового шума [36]:

$$\Delta \mathbf{w} = \alpha \nabla (\log p(\mathbf{w}|\mathbf{f}) + \frac{m}{\hat{m}} \log p(\hat{\mathfrak{D}}|\mathbf{w}, \mathbf{f})) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \frac{\alpha}{2}\mathbf{I}), \tag{12}$$

где $\hat{\mathfrak{D}}$ — псевдослучайная подвыборка, $\hat{m}$ — размер подвыборки. Шаг оптимизации α изменяется с количеством итераций:

$$\sum_{\tau=1}^{\infty} \alpha_\tau = \infty, \quad \sum_{\tau=1}^{\infty} \alpha_\tau^2 < \infty.$$

В отличие от стохастического градиентного спуска, стохастическая динамика Ланжевина сходится к апостериорному распределению параметров $p(\mathbf{w}|\mathfrak{D}, \mathbf{f})$ [35, 36, 49]. В качестве примера аппроксимации рассмотрим выборку (11). График аппроксимации апостериорного распределения с использованием динамики Ланжевина и стохастического градиентного спуска представлен на рис. 2. При одинаковом количестве итераций динамика Ланжевина продолжает аппроксимировать апостериорное распределение, в то время как градиентный спуск сходится к моде распределения.

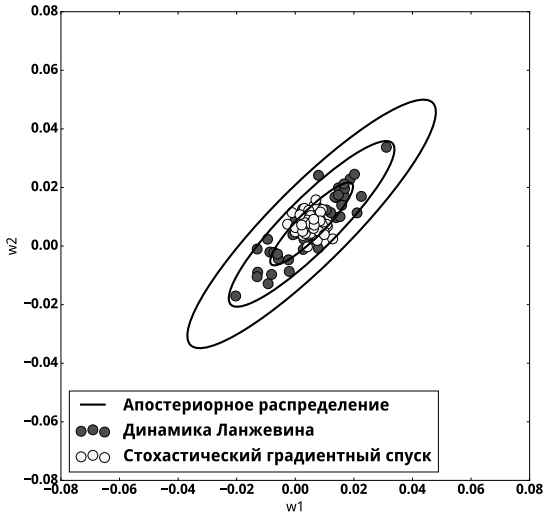
Таким образом, с использованием стохастического градиентного спуска и его модификации (12) можно получить как вариационную оценку интегральной функции правдоподобия (4), так и оценку на количество итераций оптимизации, после которого начнется переобучение модели [34].

Заметим, что при добавлении шума оценка энтропии изменятся [50, 51]:

$$\hat{\mathsf{S}}\big(q^\tau(\mathbf{w})\big) \geq \frac{1}{2}|\mathbf{w}| \log\Big(\exp\big(\frac{2\mathsf{S}(q^\tau(\mathbf{w}))}{|\mathbf{w}|}\big) + \exp\big(\frac{2\mathsf{S}(\boldsymbol{\epsilon})}{|\mathbf{w}|}\big)\Big).$$

где $|\mathbf{W}|$ — мощность множества параметров, $\mathsf{S}(\mathcal{N}(0, \frac{\alpha}{2}))$ — энтропия нормального распределения, $\hat{\mathsf{S}}(q^\tau(\mathbf{w}))$ — энтропия распределения q^τ с учетом добавленного шума $\boldsymbol{\epsilon}$.

Рис. 2: Аппроксимация распределения стохастической динамикой Ланжевина



4 Вычислительный эксперимент

Для проверки работоспособности предложенного критерия субоптимальности модели, а также методов получения нижних оценок интегральной функции правдоподобия был проведен ряд экспериментов на выборке изображений рукописных цифр MNIST [52]. Выборка представляет собой множество изображений размером 28×28 пикселей. В силу высокой вычислительной сложности часть экспериментов была проведена на модифицированной выборке MNIST (далее — подбвыборка MNIST), полученной из исходной домножением на матрицу 784×50. Таким образом было получено представление выборки меньшей размерности.

Было проведено три эксперимента. В первом эксперименте проводился выбор порождающей модели автокодировщика и анализ устойчивости полученных моделей на подбвыборке MNIST. Во втором эксперименте был проведен выбор разделяющей модели f_D субпотимальной сложности без использования порождающей модели на подбвыборке MNIST, т.е. в случае, когда $\mathfrak{F}_G = \{\mathbf{f}_G(\mathbf{x}) = \mathbf{x}\}$. Третий эксперимент заключался в выборе субоптимальной модели для полной выборки MNIST, анализе полученной модели. Размер подвыборок для стохастического градиентного спуска был выбран на уровне $|\hat{\mathfrak{D}}| = 100$.

Выбор порождающей модели. Выбор порождающей модели $\mathbf{f}_G$ проводился среди моделей вариационных автокодировщиков с $20k, k \in \{1, \ldots, 5\}$ нейронами на первом слое и втором слое. В качестве нелинейной функции активации для вариационного автокодировщика использовалась кусочно-линейная функция relu:

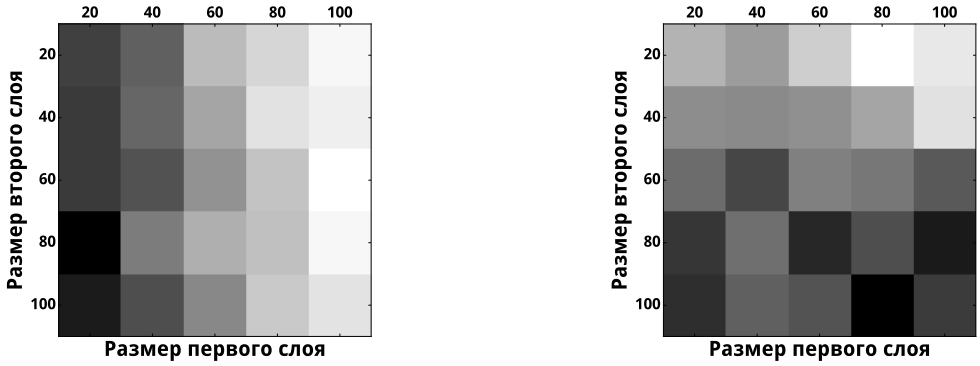
$$\text{relu}(\mathbf{x}) = \max(\mathbf{0}, \mathbf{x}) \sim \ln(1 + \mathbf{e}^{(\mathbf{x})}). \tag{13}$$

Априорное распределение для вариационного автокодировщика было задано как $p(\mathbf{w}|\mathbf{f}_G) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Рис. 3

(a) Оценки моделей вариационных автокодировщиков методом максимального апостериорного правдоподобия

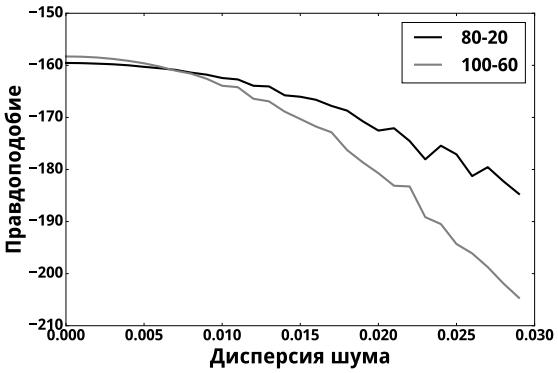
(b) Оценки моделей вариационных автокодировщиков по правдоподобию моделей



В ходе эксперимента были получены оценки правдоподобия моделей (4) и максимальной апостериорной вероятности. График нормированных оценок приведен на рис. 3, наибольшие оценки соответствуют белому цвету. Как видно из графиков, оценки и отношения между моделями существенно отличаются. Рассмотрим, для примера, модели, имеющие наибольшие значения соответствующих оценок: модель с 80 нейронами на первом слое и 20 нейронами на втором слое, и модель со 100 нейронами на первом слое и 60 нейронами на втором слое. График зависимости качества модели от возмущения параметров приведен на Рис. 4. Можно заметить, что при небольшом возмущении параметров модель, имеющая наибольшую оценку максимальной апостериорной вероятности оказывается менее устойчивой. Модель, дающая меньшее качество при фиксированных параметрах, но более устойчивая, яв-

ляется более предпочтительной, т.к. при дообучении модели, т.е. оптимизации всех параметров по единому критерию, значения параметров модели могут сильно измениться.

Рис. 4: Оценки моделей вариационных автокодировщиков при возмущении параметров модели

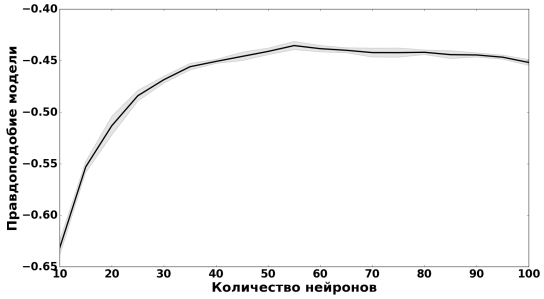


Выбор разделяющей модели. Во втором эксперименте был проведен выбор разделяющей модели среди моделей, состоящих из однослойных сетей с $10k$ нейронов, $k \in \{1, \ldots, 10\}$. В качестве априорного распределения $p(\mathbf{w}|\mathbf{f}_D)$ использовалось начальное распределение параметров [34]. Оценка строилась по пяти точкам мультистарта: $r = 5$. В качестве нелинейной функции активации использовался гиперболический тангенс. Результаты эксперимента приведены на рис. 5. На графиках видно, что оценка субоптимальной сложности (55 нейронов) соответствует моделям с небольшим количеством параметров и приемлемым качеством классификации. Заметим, что при дальнейшем усложнении модели качество классификации меняется незначительно.

Выбор субоптимальной модели. Выборка MNIST. Выбор порождающей модели $\mathbf{f}_G$ проводился среди моделей вариационных автокодировщиков с $100k$, $k \in \{1, \ldots, 7\}$ нейронами на первом слое, соответствующем отображениям $\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\sigma}(\mathbf{x})$ и $25k, k \in \{1, \ldots, 5\}$ нейронами на втором слое, соответствующем отображению q_ϕ. Количество нейронов на втором слое выбиралось меньшим, чем на первом исходя из вычислительных экспериментов в работах по вариационным автокодировщикам [19, 20].

Рис. 5

(a) Интегральная оценка правдоподобия для однослойной нейросети



(b) Зависимость качества модели от количества параметров однослойной нейросети

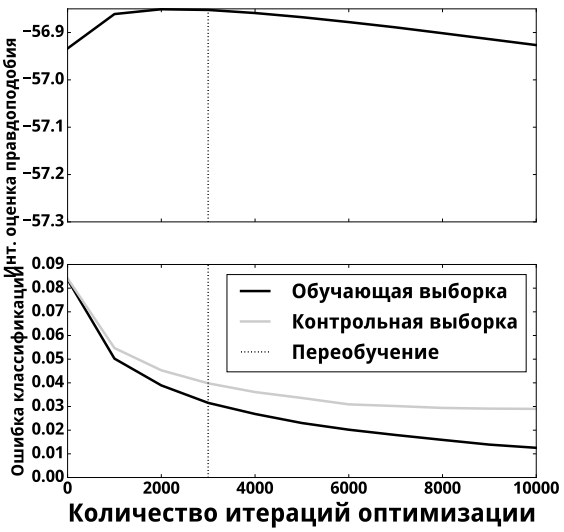
Выбор разделяющей модели $\mathbf{f}_D$ проводился среди нейросетей с одним скрытым слоем с $50k, k \in \{1, \ldots, 14\}$ нейронами. В обеих подмоделях $\mathbf{f}_G, \mathbf{f}_D$ использовалась кусочно-линейная функция активации (13). Априорное распределение для порождающей модели было выбрано как $p(\mathbf{w}|\mathbf{f}_G) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Априорное распределение для разделяющей модели и при дообучении было выбрано как $p(\mathbf{w}|\mathbf{f}_D) \sim \mathcal{N}(\mathbf{0}, 1000 \cdot \mathbf{I})$. Подобное априорное распределение соответствует l_2 регуляризации с параметром $\lambda_{l_2} = 10^{-3}$ и определяет большую величину неопределенности параметров [53]. Оценка разделяющей модели $\mathbf{f}_D$ строилась по восьми точкам мультистарта: $r = 8$.

График качества классификации субоптимальной модели, а также дообученной модели представлен на рис 6. Была получена субоптимальная модель, имеющая следующую структуру: 500 нейронов на первом слое, 75 нейронов на втором слое и 50 нейронов на третьем слое. Субоптимальная модель после дообучения показывает приемлемое качество при небольшом количестве параметров. В ходе эксперимента также было проверено качество модели при ранней остановке, критерием которой являлось снижение вариационной оценки (4) при использовании стохастического градиентного спуска. Можно заметить, что полученная при ранней остановке модель имеет заниженное качество по сравнению с моделью, прошедшей полное дообучение. Как видно из графика, максимуму вариационной нижней оценки соответствует шаг оптимизации, после которого разница между качеством на контрольной и обучающей выборке начинает существенно увеличиваться. Таким образом, можно сделать вывод о применимости критерия ранней остановки в случае высокой вычислительной

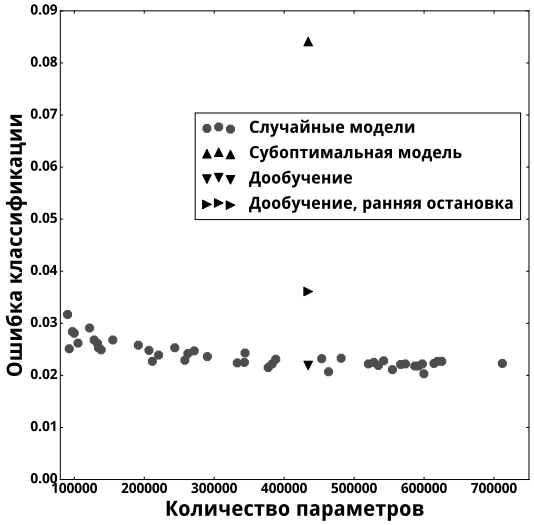
стоимости оптимизации или при нехватке размеченных данных [13,38,39].

Рис. 6

(a) Зависимость интегральной оценки правдоподобия и качества классификации от числа итераций оптимизации



(b) Зависимость качества модели от количества параметров на выборке MNIST



Заключение

В работе предложен критерий оптимальности моделей глубокого обучения. Предложен метод декомпозиции модели, а также критерий субоптимальности для каждой из подмоделей. Предложен алгоритм поиска субоптимальной модели, основанный на получении вариационной нижней оценки интеграла правдоподобия модели. Для разделяющей подмодели предложен метод получения оценки, основанный на стохастическом градиентном спуске, позволяющий проводить выбор модели и оптимизацию модели единообразно. Исследованы свойства стохастического градиентного спуска, а также оценок правдоподобия, полученных с его использованием.

Работа представленного алгоритма проиллюстрирована на выборке изображений рукописных цифр. Произведен анализ полученных моделей, а также качества соответствующих подмоделей.

Список литературы

[1] Greedy Layer-Wise Training of Deep Networks / Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle // Advances in Neural Information Processing Systems 19 / Ed. by B. Schölkopf, J. C. Platt, T. Hoffman. — MIT Press, 2007. — Pp. 153–160. http://papers.nips.cc/paper/3048-greedy-layer-wise-training-of-deep-networks.pdf.

[2] Predicting Parameters in Deep Learning / Misha Denil, Babak Shakibi, Laurent Dinh et al. // Advances in Neural Information Processing Systems 26 / Ed. by C.j.c. Burges, L. Bottou, M. Welling et al. — 2013. — Pp. 2148–2156. http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips26/1053.pdf.

[3] *Попова М. С., Стрижов В. В.* Выбор оптимальной модели классификации физической активности по измерениям акселерометра // *Информатика и ее применения.* — 2015. — Т. 9(1). — С. 79–89. http://strijov.com/papers/Popova2014OptimalModelSelection.pdf.

[4] *MacKay David J. C.* Information Theory, Inference & Learning Algorithms. — New York, NY, USA: Cambridge University Press, 2002.

[5] *Bishop Christopher M.* Pattern Recognition and Machine Learning (Information Science and Statistics). — Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[6] *Salakhutdinov Ruslan, Hinton Geoffrey E.* Learning a Nonlinear Embedding by Preserving Class Neighbourhood Structure // Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07) / Ed. by Marina Meila, Xiaotong Shen. — Vol. 2. — Journal of Machine Learning Research - Proceedings Track, 2007. — Pp. 412–419. http://jmlr.csail.mit.edu/proceedings/papers/v2/salakhutdinov07a/salakhutdinov07a.pdf.

[7] On the importance of initialization and momentum in deep learning / Ilya Sutskever, James Martens, George E. Dahl, Geoffrey E. Hinton // Proceedings of the 30th International Conference on Machine Learning (ICML-13) / Ed. by Sanjoy Dasgupta, David Mcallester. — Vol. 28. — JMLR Workshop and Conference Proceedings, 2013. — Май. — Pp. 1139–1147. http://jmlr.org/proceedings/papers/v28/sutskever13.pdf.

[8] Approximation and learning by greedy algorithms / Andrew R. Barron, Albert Cohen, Wolfgang Dahmen, Ronald A. DeVore // *Ann. Statist.* — 2008. — 02. — Vol. 36, no. 1. — Pp. 64–94. http://dx.doi.org/10.1214/009053607000000631.

[9] *Tzikas Dimitris, Likas Aristidis.* An Incremental Bayesian Approach for Training Multilayer Perceptrons // Artificial Neural Networks – ICANN 2010: 20th International Conference, Thessaloniki, Greece, September 15-18, 2010, Proceedings, Part I / Ed. by Konstantinos Diamantaras, Wlodek Duch, Lazaros S. Iliadis. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. — Pp. 87–96. http://dx.doi.org/10.1007/978-3-642-15819-3_12.

[10] *Tipping Michael E.* Sparse Bayesian Learning and the Relevance Vector Machine // *J. Mach. Learn. Res.* — 2001. — Сентябрь. — Vol. 1. — Pp. 211–244. http://dx.doi.org/10.1162/15324430152748236.

[11] *Cun Yann Le, Denker John S., Solla Sara A.* Optimal Brain Damage // Advances in Neural Information Processing Systems. — Morgan Kaufmann, 1990. — Pp. 598–605.

[12] *Hinton Geoffrey E., Osindero Simon, Teh Yee-Whye.* A Fast Learning Algorithm for Deep Belief Nets // *Neural Comput.* — 2006. — Июль. — Vol. 18, no. 7. — Pp. 1527–1554. http://dx.doi.org/10.1162/neco.2006.18.7.1527.

[13] Semi-supervised Learning with Deep Generative Models / Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, Max Welling // Advances in Neural Information Processing Systems 27 / Ed. by Z. Ghahramani, M. Welling, C. Cortes et al. — Curran Associates, Inc., 2014. — Pp. 3581–3589. http://papers.nips.cc/paper/5352-semi-supervised-learning-with-deep-generative-models.pdf.

[14] *Li Yi, Shapiro L. O., Bilmes J. A.* A generative/discriminative learning algorithm for image classification // Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. — Vol. 2. — 2005. — Oct. — Pp. 1605–1612 Vol. 2.

[15] *J. Lasserre.* Hybrid of generative and discriminative methods for machine learning: Ph.D. thesis / University of Cambridge. — 2008.

[16] *Cho Kyunghyun.* Foundations and Advances in Deep Learning: G5 Artikkeliväitöskirja. — Aalto University; Aalto-yliopisto, 2014. — P. 277. http://urn.fi/URN:ISBN:978-952-60-5575-6.

[17] *Alain Guillaume, Bengio Yoshua.* What regularized auto-encoders learn from the data-generating distribution // *Journal of Machine Learning Research.* — 2014. — Vol. 15, no. 1. — Pp. 3563–3593. http://dl.acm.org/citation.cfm?id=2750359.

[18] *Kamyshanska Hanna, Memisevic Roland.* On autoencoder scoring // Proceedings of the 30th International Conference on Machine Learning (ICML-13) / Ed. by Sanjoy Dasgupta, David Mcallester. — Vol. 28. — JMLR Workshop and Conference Proceedings, 2013. — Май. — Pp. 720–728. http://jmlr.org/proceedings/papers/v28/kamyshanska13.pdf.

[19] *D. Kingma M. Welling.* Auto-Encoding Variational Bayes // Proceedings of the International Conference on Learning Representations (ICLR). — 2014.

[20] How to Train Deep Variational Autoencoders and Probabilistic Ladder Networks. / Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe et al. // *CoRR.* — 2016. — Vol. abs/1602.02282. http://dblp.uni-trier.de/db/journals/corr/corr1602.html#SonderbyRMSW16.

[21] Semi-Supervised Learning with Ladder Network. / Antti Rasmus, Harri Valpola, Mikko Honkala et al. // *CoRR.* — 2015. — Vol. abs/1507.02672. http://dblp.uni-trier.de/db/journals/corr/corr1507.html#RasmusVHBR15.

[22] *Токмакова А. А., Стрижов В. В.* Оценивание гиперпараметров линейных и регрессионных моделей при отборе шумовых и коррелирующих признаков // *Информатика и её применения.* — 2012. — Т. 6(4). — С. 66–75. http://strijov.com/papers/Tokmakova2011HyperParJournal_Preprint.pdf.

[23] *Зайцев А. А., Стрижов В. В., Токмакова А. А.* Оценка гиперпараметров регрессионных моделей методом максимального правдоподобия // *Информационные технологии.* — 2013. — Vol. 2. — Pp. 11–15. http://strijov.com/papers/ZaytsevStrijovTokmakova2012Likelihood_Preprint.pdf.

[24] *Strijov V., Weber Gerhard-Wilhelm.* NONLINEAR REGRESSION MODEL GENERATION USING HYPERPARAMETERS OPTIMIZATION: Preprint 2009-21. — Middle East Technical University, 06800 Ankara, Turkey: Institute of Applied Mathematics, 2009. — Октябрь. — Preprint No. 149.

[25] *Стрижов В. В.* Порождение и выбор моделей в задачах регрессии и классификации: Ph.D. thesis / Вычислительный центр РАН. — 2014. http://strijov.com/papers/Strijov2015ModelSelectionRu.pdf.

[26] *Grünwald Peter.* A Tutorial Introduction to the Minimum Description Length Principle // Advances in Minimum Description Length: Theory and Applications. — MIT Press, 2005.

[27] *Перекрестенко Д.О.* Анализ структурной и статистической сложности суперпозиции нейронных сетей. — 2014. http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Perekrestenko2014ComplexityAnalysis/

[28] *Vladislavleva E.* Other publications TiSEM: : Tilburg University, School of Economics and Management, 2008. http://EconPapers.repec.org/RePEc:tiu:tiutis:65a72d10-6b09-443f-8cb9-88f3bb3bc31b.

[29] *Xu Huan, Mannor Shie.* Robustness and generalization // *Machine Learning.* — 2012. — Vol. 86, no. 3. — Pp. 391–423. http://dx.doi.org/10.1007/s10994-011-5268-1.

[30] Intriguing properties of neural networks. / Christian Szegedy, Wojciech Zaremba, Ilya Sutskever et al. // *CoRR.* — 2013. — Vol. abs/1312.6199. http://dblp.uni-trier.de/db/journals/corr/corr1312.html#SzegedyZSBEGF13.

[31] Stochastic Variational Inference / Matthew D. Hoffman, David M. Blei, Chong Wang, John Paisley // *J. Mach. Learn. Res.* — 2013. — Май. — Vol. 14, no. 1. — Pp. 1303–1347. http://dl.acm.org/citation.cfm?id=2502581.2502622.

[32] *Graves Alex.* Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett et al. — Curran Associates, Inc., 2011. — Pp. 2348–2356. http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf.

[33] *Salimans Tim, Kingma Diederik P., Welling Max.* Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. // ICML / Ed. by Francis R. Bach, David M. Blei. — Vol. 37 of *JMLR Proceedings*. — JMLR.org, 2015. — Pp. 1218–1226. http://dblp.uni-trier.de/db/conf/icml/icml2015.html#SalimansKW15.

[34] *Maclaurin Dougal, Duvenaud David K., Adams Ryan P.* Early Stopping is Nonparametric Variational Inference // *CoRR*. — 2015. — Vol. abs/1504.01344. http://arxiv.org/abs/1504.01344.

[35] *Mandt Stephan, Hoffman Matthew D, Blei David M.* Continuous-Time Limit of Stochastic Gradient Descent Revisited.

[36] *Welling Max, Teh Yee Whye.* Bayesian Learning via Stochastic Gradient Langevin Dynamics // Proceedings of the 28th International Conference on Machine Learning (ICML-11) / Ed. by Lise Getoor, Tobias Scheffer. — ICML '11. — New York, NY, USA: ACM, 2011. — June. — Pp. 681–688.

[37] *Arlot Sylvain, Celisse Alain.* A survey of cross-validation procedures for model selection // *Statist. Surv.* — 2010. — Vol. 4. — Pp. 40–79. http://dx.doi.org/10.1214/09-SS054.

[38] Fast and Accurate Support Vector Machines on Large Scale Systems / Abhinav Vishnu, Jeyanthi Narasimhan, Lawrence Holder et al. // 2015 IEEE International Conference on Cluster Computing, CLUSTER 2015, Chicago, IL, USA, September 8-11, 2015. — 2015. — Pp. 110–119. http://dx.doi.org/10.1109/CLUSTER.2015.26.

[39] Cross-validation pitfalls when selecting and assessing regression and classification models / Damjan Krstajic, Ljubomir J. Buturovic, David E. Leahy, Simon Thomas // *Journal of Cheminformatics*. — 2014. — Vol. 6, no. 1. — Pp. 1–15. http://dx.doi.org/10.1186/1758-2946-6-10.

[40] *Hornung Roman, Bernau Christoph, Truntzer Caroline et al.* Full versus incomplete cross-validation: measuring the impact of imperfect separation between training and test sets in prediction error estimation. — 2014. http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-20682-6.

[41] *Bengio Yoshua, Grandvalet Yves.* No Unbiased Estimator of the Variance of K-Fold Cross-Validation // *J. Mach. Learn. Res.* — 2004. — Декабрь. — Vol. 5. — Pp. 1089–1105. http://dl.acm.org/citation.cfm?id=1005332.1044695.

[42] *Maclaurin Dougal, Duvenaud David, Adams Ryan.* Gradient-based Hyperparameter Optimization through Reversible Learning // Proceedings of the 32nd International Conference on Machine Learning (ICML-15) / Ed. by David Blei, Francis Bach. — JMLR Workshop and Conference Proceedings, 2015. — Pp. 2113–2122. http://jmlr.org/proceedings/papers/v37/maclaurin15.pdf.

[43] *Domke Justin.* Generic Methods for Optimization-Based Modeling. // AISTATS / Ed. by Neil D. Lawrence, Mark A. Girolami. — Vol. 22 of *JMLR Proceedings*. — JMLR.org, 2012. — Pp. 318–326. http://dblp.uni-trier.de/db/journals/jmlr/jmlrp22.html#Domke12.

[44] *Попова М.С. Стрижов В.В.* Выбор оптимальной модели классификации физической активности по измерениям акселерометра // *Информатика и ее применения.* — 2015. — Т. 9(1). — С. 79–89.

[45] Supervised Probabilistic Principal Component Analysis / Shipeng Yu, Kai Yu, Volker Tresp et al. // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '06. — New York, NY, USA: ACM, 2006. — Pp. 464–473. http://doi.acm.org/10.1145/1150402.1150454.

[46] *Shang Yi, Wah B. W.* Global optimization for neural network training // *Computer*. — 1996. — Mar. — Vol. 29, no. 3. — Pp. 45–54.

[47] Gradient descent converges to minimizers / Jason D Lee, Max Simchowitz, Michael I Jordan, Benjamin Recht // *University of California, Berkeley.* — 2016. — Vol. 1050. — P. 16.

[48] *Geiger Bernhard.* Information Loss in Deterministic Systems: Ph.D. thesis. — Graz, 2014. — June.

[49] *Sato Issei, Nakagawa Hiroshi.* Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process // Proceedings of the 31st International Conference on Machine Learning (ICML-14). — 2014. — Pp. 982–990.

[50] *Dembo Amir, Cover Thomas M, Thomas Joy A.* Information theoretic inequalities // *Information Theory, IEEE Transactions on.* — 1991. — Vol. 37, no. 6. — Pp. 1501–1518.

[51] *Nicholas Altieri, D. Duvenaud.* Variational Inference with Gradient Flows. URL: http://approximateinference.org/accepted/AltieriDuvenaud2015.pdf.

[52] *LeCun Yann, Cortes Corinna.* MNIST handwritten digit database. — 2010. http://yann.lecun.com/exdb/mnist/.

[53] Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks / Chunyuan Li, Changyou Chen, David E. Carlson, Lawrence Carin // Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA. — 2016. — Pp. 1788–1794. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11835.