Устойчивый к шуму метод обучения вариационного автокодировщика *

М. В. Фигурнов, К. А. Струминский, Д. П. Ветров

Вариационный автокодировщик (ВАК) — вероятностный метод обучения без учителя, использующий глубинное обучение. В статье предлагается устойчивый к шуму метод обучения ВАК, основанный на модификации функции правдоподобия. Предлагаются и анализируются две нижние оценки в качестве целевых функций для ВАК. Эффективность метода продемонстрирована в экспериментах с искусственно добавленными шумовыми объектами.
Ключевые слова: обучение без учителя, генеративное моделирование, вариационный автокодировщик, важностно-взвешенный автокодировщик, робастность, устойчивость к шуму

1. Введение

Глубинное обучение (deep learning), класс техник машинного обучения, приобрёл большую популярность в последние годы [8]. Ключевая идея глубинного обучения — автоматическое извлечение иерархии признаковых представлений из высокоразмерных данных. К примеру, при обработке изображений на первых уровнях иерархии могут обучиться детекторы ориентированных градиентов, а на последующих — детекторы частей объектов [14].

На сегодняшний день наиболее успешны методы глубинного обучения с учителем, в особенности для задачи классификации. В этом случае

*Работа была выполнена при поддержке лаборатории тензорных сетей и глубинного обучения для интеллектуального анализа данных на базе Сколковского института науки и технологии, грант Правительства РФ № 14.756.31.0001. Работа также была поддержана проектом повышения конкурентоспособности российских университетов «5-100».

используют многослойные (глубинные) нейронные сети, которые принимают на вход неструктурированные данные высокой размерности, например, пиксели изображения или спектрограмму звука. Выходом сети является вероятностное распределение над дискретным множеством классов. Предположим, что имеется обучающая выборка, состоящая из объектов и правильных классов для них. Тогда параметры нейронной сети настраиваются стохастическим градиентным спуском, максимизирующим вероятность принадлежности объектов правильному классу. Для подсчёта градиента функционала качества по параметрам нейронной сети применяется метод обратного распространения ошибки (backpropagation).

К сожалению, методы обучения с учителем имеют ограниченную применимость. Во многих случаях получение правильных ответов для обучающих объектов трудозатратно, а иногда и вовсе невозможно. К примеру, в задаче обнаружения редких болезней по медицинским снимкам крайне трудно собрать репрезентативную выборку пациентов. В таких случаях необходимы методы обучения без учителя, то есть методы выявления закономерностей без известных правильных ответов для данных [1]. Методы обучения без учителя позволяют решить задачу детекции аномалий, а также существенно уменьшить требования методов обучения с учителем к размеру обучающей выборки.

Как правило, методы обучения без учителя не предполагают наличие шума в данных. В этой работе мы рассматриваем случай, когда выборка данных содержит большое число шумовых объектов, которые мы хотели бы игнорировать при обучении. В начале мы рассмотрим схему работы вариационного автокодировщика, распространённого метода обучения без учителя. Затем мы предложим новый функционал качества, устойчивое к шуму правдоподобие, предложим и проанализируем нижние оценки на этот функционал. В конце будут приведены результаты экспериментов на искусственно зашумлённых данных, демонстрирующие преимущество предложенного метода.

1.1. Вариационный автокодировщик

В данной работе мы предлагаем новый метод обучения вариационного автокодировщика [6]. Вариационный автокодировщик — метод обучения без учителя, который настраивает вероятностную модель порождения данных (генеративную модель). Опишем базовую схему работы вариационного автокодировщика.

Пусть имеется обучающая выборка $\mathcal{D} = \{x_1, \ldots, x_N\}$ из N неза-

висимых объектов, полученных из распределения данных. Признаковое описание объектов может быть как непрерывным, так и дискретным. Для определённости будем считать, что все признаки бинарны: $x_i \in \{0,1\}^D$, D — размерность описания. Также введём пространство скрытых переменных $\mathcal{Z} = \mathbb{R}^d$. Предположим следующий случайный процесс порождения данных. Сначала из многомерного нормального априорного распределения с нулевым средним и единичной матрицей ковариации $p(z) = \mathcal{N}(z|0, I)$ генерируется скрытая переменная $z \in \mathcal{Z}$. Затем объект x генерируется из условного параметрического распределения Бернулли на каждый признак $p_\theta(x|z) = \prod_{i=1}^{D} \text{Ber}(x_i|t_{\theta,i}(z))$. Здесь $t_{\theta,i}(z) \in [0,1]$ — вероятность события $x_i = 1$, функция от скрытых переменных, параметризуемая вектором θ. Заметим, что мы предположили независимость всех признаков при условии скрытой переменной, что не означает их безусловную независимость. Распределение данных, задаваемое этой моделью, имеет вид:

$$p_\theta(x) = \int_{\mathcal{Z}} p_\theta(x,z)dz = \int_{\mathcal{Z}} p_\theta(x|z)p(z)dz \tag{1}$$

Применим метод максимального правдоподобия к этой модели:

$$\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(x_i) = \frac{1}{N}\sum_{i=1}^{N} \log \int_{\mathcal{Z}} p_\theta(x_i|z)p(z)dz \to \max_\theta \tag{2}$$

Решение этой задачи максимизации сопряжено с двумя трудностями. Во-первых, в большинстве практически важных случаев интеграл (2) не вычисляется аналитически. Распространённое решение этой проблемы — методы вариационного вывода. Во-вторых, даже при известной аналитической форме интеграла максимизация по θ, как правило, не может быть проведена явно и требует применения стохастического градиентного спуска.

Для начала применим вариационный вывод и перейдём к задаче максимизации вариационной нижней оценки на логарифм маргинального правдоподобия. Для упрощения нотации будем рассматривать лишь один объект x; обобщение на случай нескольких объектов проводится взятием среднего по ним. Введём вспомогательное параметрическое распределение $q_\phi(z|x)$ на пространстве $\mathcal{Z}$, умножим и разделим подынтегральное выражение на него, чтобы получить математическое ожидание по этому распределению.

$$\log p_\theta(x) = \log \int_{\mathcal{Z}} q_\phi(z|x)\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}dz = \log \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}. \tag{3}$$

Применим неравенство Йенсена, чтобы получить вариационную нижнюю оценку на логарифм маргинального правдоподобия $\log p_\theta(x)$:

$$\log p_\theta(x) = \log \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} \geq \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}. \tag{4}$$

Определение 1. *Вариационной нижней оценкой на логарифм маргинального правдоподобия называется величина*

$$\mathcal{L}(x,\theta,\phi) = \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}. \tag{5}$$

Таким образом, задача обучения модели была сведена к задаче оптимизации

$$\mathcal{L}(x,\theta,\phi) \to \max_{\theta,\phi} \tag{6}$$

Заметим, что по правилу произведения вероятностей $p_\theta(x,z) = p_\theta(x|z)p(z) = p_\theta(x)p_\theta(z|x)$. Таким образом, последнее выражение может быть записано в следующем виде:

$$\begin{aligned}\mathcal{L}(x,\theta,\phi) = \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(x)p_\theta(z|x)}{q_\phi(z|x)} = \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \log p_\theta(x) - \\ - \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \log \frac{q_\phi(z|x)}{p_\theta(z|x)} = \log p_\theta(x) - \mathrm{KL}(q_\phi(z|x)||p_\theta(z|x)).\end{aligned} \tag{7}$$

Таким образом, максимизация вариационной нижней оценки по параметрам ϕ эквивалентна минимизации дивергенции Кульбака-Лейблера между $q_\phi(z|x)$ и $p_\theta(z|x)$. Поскольку дивергенция Кульбака-Лейблера неотрицательна и равна нулю тогда и только тогда, когда плотности распределений совпадают почти всюду, оптимум в задаче оптимизации (при достаточно широком множестве ϕ) достигается при совпадении $q_\phi(z|x)$ и $p_\theta(z|x)$ почти всюду. В этом случае вариационная нижняя оценка будет равна логарифму маргинального правдоподобия. Итак, $q_\phi(z|x)$ приближает истинное апостериорное распределение $p_\theta(z|x)$.

Мы хотели бы применить градиентный метод оптимизации в задаче (6). Для того, чтобы упростить взятие градиента по параметрам θ в случаях, когда невозможно выполнить аналитическое взятие мат. ожидания по z, требуется выполнить т.н. *трюк репараметризации* [6]. Опишем его для случая многомерного нормального распределения с независимыми компонентами. Предположим, что $q_\phi(z|x) = \prod_{j=1}^{d} \mathcal{N}(z_j|\mu_{\phi,j}(x), \sigma^2_{\phi,j}(x))$.

Тогда процедуру генерации точек из этого распределения можно переписать в следующем виде:

$$z \sim q_\phi(z|x) \leftrightarrow \xi \sim \mathcal{N}(\xi|0, I), z = f_\phi(\xi, x) = \mu_\phi(x) + \xi\sigma_\phi(x). \tag{8}$$

Другими словами, мы заменяем генерацию точек из параметрического распределения на генерацию точек из распределения без настраиваемых параметров. Затем эти точки преобразуются с помощью параметрической функции.

Используя трюк репараметризации, мы можем преобразовать выражение (5) следующим образом

$$\mathcal{L}(x, \theta, \phi) = \mathop{\mathbb{E}}_{\xi\sim\mathcal{N}(\xi|0,I)} \log \frac{p_\theta(x|f_\phi(\xi, x))p(f_\phi(\xi, x))}{q_\phi(f_\phi(\xi, x)|x)}. \tag{9}$$

Теперь градиент от вариационной нижней оценки по параметрам ϕ и θ можно занести внутрь математического ожидания. Мат. ожидание по распределению на ξ мы будем приближать с помощью одной точки по методу Монте-Карло. Наконец, для выборки объектов вместо взятия полного градиента по всем N объектам, мы будем оценивать градиент по мини-батчу из $n < N$ объектов.

В модели вариационного автокодировщика для параметризации распределений $q_\phi(z|x)$ и $p_\theta(x|z)$ используются многослойные нейронные сети, на выходном слое которых предсказываются параметры соответствующих распределений. Нейронные сети позволяет моделировать сложные нелинейные зависимости между наблюдаемыми переменными x и скрытыми переменными z. Модель получила название по аналогии с нейросетевым автокодировщиком, в котором вместо вместо условных распределений применяются «кодирующее» отображение $x \to z$ и «декодирующее» отображение $z \to x$.

Оценка логарифма правдоподобия данных. Оценка $\mathcal{L}(x, \theta, \phi)$ может быть существенно ниже истинного логарифма правдоподобия данных $\log p_\theta(x)$. Для более точной оценки логарифма правдоподобия используется метод выборки по значимости (importance sampling).

Определение 2. *Пусть $K > 1$. Важностно-взвешенной оценкой на логарифм правдоподобия с K точками называется выражение*

$$\mathcal{L}^K(x, \theta, \phi) = \mathop{\mathbb{E}}_{z_1,\ldots,z_K\sim q_\phi(z|x)} \log \frac{1}{K} \sum_{k=1}^{K} \frac{p_\theta(x|z_k)p(z_k)}{q_\phi(z_k|x)}. \tag{10}$$

Применяя неравенство Йенсена, нетрудно видеть, что это выражение действительно является нижней оценкой на логарифм правдоподобия. Кроме того, если $\forall z \in \mathcal{Z} \;\; q_\phi(z|x) > 0$, то $\lim_{K\to+\infty} \mathcal{L}^K(x,\theta,\phi) = \log p_\theta(x)$. Важностно-взешенные оценки были впервые использованы для обучения вариационных автокодировщиков в [2].

2. Обзор литературы

В последние годы появилось множество работ, предлагающих новые способы обучения вариационных автокодировщиков. Например, работа [2] предлагает оптимизировать оценку (10) на логарифм правдоподобия, а [10] — оптимизировать дивергенцию Реньи вместо дивергенции Кульбака-Лейблера. Другими возможными улучшениями является введение более выразительного распределения на скрытые переменные, к примеру, при помощи использования нормализационных потоков [11].

В работе [13] предлагается способ модификации вероятностных моделей для увеличения устойчивости к шуму в данных. Данный подход требует введения дополнительных распределений и усложняет вывод в модели. Кроме того, под вопросом остаётся возможность настройки порождающей модели данных при помощи подобных моделей.

Данная работа является существенно расширенной версией доклада [3].

3. Устойчивое к шуму правдоподобие

В статье мы предлагаем использовать для обучения вариационного автокодировщика модификацию функции правдоподобия - устойчивое к шуму правдоподобие. Ниже будет показано, что максимизация устойчивого к шуму правдоподобия позволяет присвоить нулевую вероятностную массу редким объектам (выбросам) в выборке, концентрируя распределение на характерных объектах выборки.

Зафиксируем семейство распределений $\{p_\theta(x)|\theta \in \Theta\}$ на конечном носителе D меры $\mu(\mathrm{D})$. Один из подходов робастной статистики заключается в моделировании данных с помощью семейств распределений с более тяжелыми хвостами [12]. Следуя этому подходу, мы предлагаем расширить данное семейство с помощью смеси с равномерным распределением на D:

$$p_\theta^\omega(x) = \omega\frac{1}{\mu(\mathrm{D})} + (1-\omega)p_\theta(x),\ \ \omega \in [0,1]. \tag{11}$$

Первое слагаемое выполняет вспомогательную роль и позволяет ослабить штраф на выбросы в данных в функции правдоподобия, а второе слагаемое позволяет описывать наблюдаемые данные, Мы предлагаем фиксировать параметр ω и выбирать параметры распределения $p_\theta(x)$ с помощью метода максимального правдоподобия для семейства $\{p_\theta^\omega(x)|\omega \in \Omega\}$. Предложенный подход эквивалентен максимизации функции с одним вещественным параметром $\varepsilon = \frac{\omega}{(1-\omega)\mu(\mathrm{D})}$, которую мы будем называть функцией устойчивого праводподобия.

Определение 3. *Функцией устойчивого правдоподобия модели $p_\theta(x)$ для параметра устойчивости $\varepsilon \geq 0$ и выборки $\mathcal{D} = \{x_1, \ldots, x_N\}$ называется функция*

$$L(\mathcal{D}) = \prod_{i=1}^{N}(\varepsilon + p_\theta(x_i)). \tag{12}$$

В отличие от функции правдоподобия, функция устойчивого правдоподобия не вырождается в ноль, если для небольшого числа объектов обучающей выборки модель дает $p_\theta(x_i) = 0$. Для более детального анализа функции сделаем в этом разделе несколько дополнительных допущений. Предположим, что мы моделируем дискретную случайную величину, принимающую конечное число значений $M < \infty$. Помимо этого допустим, что класс распределений $\{p_\theta|\theta \in \Theta\}$ охватывает все распределения на M исходах. В таком случае любое распределение представляется вещественным вектором $\mathbf{p} = (p_1, \ldots, p_M)$, а параметр θ можно опустить без ущерба изложению.

Для дальнейшего анализа перепишем функцию устойчивого правдоподобия в терминах эмпирических частот $r_j = \frac{\sum_{i=1}^{N}[x_i=j]}{N}$:

$$\frac{1}{N}\log L(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\log(\varepsilon + p_{x_i}) = \sum_{j=1}^{M} r_j \log(\varepsilon + p_j). \tag{13}$$

Утверждение 1. *Пусть распределение* $\mathbf{p} = (p_1, ..., p_M)$ *- решение задачи*

$$\begin{aligned} &\sum_{j=1}^{M} r_j \log(\varepsilon + p_j) \to \max_{\mathbf{p}} \\ &\sum_{j=1}^{M} p_j = 1 \\ &0 \le p_j \le 1, j = 1, \dots, M \end{aligned} \tag{14}$$

а частоты r_j *упорядочены по возрастанию. Тогда*

1) компоненты решения p_j *упорядочены по возрастанию;*

2) если $r_j \le \frac{1}{M}\frac{\varepsilon}{1+\varepsilon}$, *то* $p_i = 0$.

Доказательство. Целевая функция является вогнутой, а все ограничения аффинные. В данном случае условия Каруша-Куна-Такера будут достаточными условиями оптимальности. Более того, функция строго вогнутая, значит точка максимума существует и единственна. Лагранжиан задачи имеет вид

$$\Lambda(\mathbf{p}, \mu, \lambda) = \sum_{j=1}^{M} r_j \log[\varepsilon + p_j] + \sum_{j=1}^{M} \mu_j p_j - \lambda \left(\sum_{j=1}^{M} p_j - 1 \right). \tag{15}$$

Рассмотрим решение задачи оптимизации $\mathbf{p}$. Из условия стационарности $\frac{\partial \Lambda(p,\mu,\lambda)}{\partial p_j} = 0$ следует, что p_j выражается через r_j по формуле $p_j = \frac{r_j}{\lambda - \mu_j} - \varepsilon$.

При $p_j > 0$ условия дополняющей нежесткости $\mu_j p_j = 0$ позволяют упростить выражения до $p_j = \frac{r_j}{\lambda} - \varepsilon$. Отсюда следует, что неотрицательные p_j упорядочены по возрастанию.

При $p_j = 0$ из условия стационарности получается $\mu_j = \lambda - \frac{r_j}{\varepsilon}$. С учетом условия неотрицательности μ_j получаем $r_j \le \varepsilon\lambda$. Поскольку r_j упорядочены по возрастанию, это означает, что нулевые p_j могут встречаться только в начале вектора $\mathbf{p}$. Тем самым доказано, что компоненты решения упорядочены по возрастанию.

Выразим теперь λ из условия $\sum_{j=1}^{M} p_j = 1$:

$$\lambda = \frac{\sum_{j=1}^{M} [p_j > 0] r_j}{1 + \varepsilon \sum_j [p_j > 0]}. \tag{16}$$

Рассмотрим множество $P = \{j : p_j > 0\}$. Обозначим $\beta = \sum_{j \in P} r_j$. Среднее значение r_j на этом множестве равно $\frac{\beta}{|P|}$. Поскольку $\sum_{j=1}^{M} r_j = 1$, на дополнении к P среднее значение r_j равно $\frac{1-\beta}{M-|P|}$. Поскольку r_j и p_j упорядочены по возрастанию, среднее на P должно быть больше среднего на дополнении к P:

$$\frac{1-\beta}{M-|P|} < \frac{\beta}{|P|} \tag{17}$$

Отсюда $\beta > \frac{|P|}{M}$. Подставляя в выражение для λ, получаем

$$\lambda = \frac{\sum_{j=1}^{M}[p_j > 0]r_j}{1 + \varepsilon \sum_{j=1}^{M}[p_j > 0]} > \frac{\frac{|P|}{M}}{1 + |P|\varepsilon} \geq \frac{1}{M}\frac{1}{1+\varepsilon} \tag{18}$$

Пусть теперь для некоторого j выполнено $r_j \leq \frac{1}{M}\frac{\varepsilon}{1+\varepsilon}$, но $p_j > 0$. Оценим p_j сверху:

$$p_j = \frac{r_j}{\lambda} - \varepsilon \leq \frac{1}{\lambda M}\frac{\varepsilon}{(1+\varepsilon)} - \varepsilon < \varepsilon\frac{\lambda}{\lambda} - \varepsilon = 0, \tag{19}$$

получили противоречие. □

Устойчивое к шуму правдоподобие при $\varepsilon > 0$ обобщает функцию правдоподобия. Как видно из доказательства, решение задачи максимума устойчивого правдоподобия (14) для $p_j > 0$ определяется аффинным преобразованием эмпирических частот $p_j = \frac{r_j}{\lambda} - \varepsilon$. Для $\varepsilon = 0$ задача вырождается в задачу максимума правдоподобия с решением $p_j = r_j$. Но в общем случая для некоторых j величина $\frac{r_j}{\lambda} - \varepsilon$ оказывается отрицательна, поэтому в решении могут появляться нулевые компоненты $p_j = 0$ при $r_j > 0$.

4. Нижняя оценка на логарифм устойчивого правдоподобия

Важностно-взвешенные оценки на логарифм правдоподобия естественно обобщаются на случай устойчивого к шуму правдоподобия $\frac{1}{N}\sum_{i=1}^{N} \log[\varepsilon + p_\theta(x_i)]$. В этом разделе мы вводим две нижних оценки на логарифм устойчивого правдоподобия, а затем проводим их сравнение и анализ. Выводы в этом разделе без ограничения общности приводятся для одного объекта обучающей выборки $x \in \mathcal{D}$.

Определение 4. *Возъмем $K \in \mathbb{N}$, $\varepsilon > 0$ и вспомогательное параметрическое распределение $q_\phi(z|x)$. Нижней оценкой $\mathcal{L}_p^K(x, \theta, \phi)$ логарифма устойчивого к шуму правдоподобия $\log(\varepsilon + p_\theta(x))$ называется функция*

$$\mathcal{L}_p^K(x, \theta, \phi) = \mathop{\mathbb{E}}_{z_1,\dots,z_k \sim q_\phi(z|x)} \log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{[\varepsilon + p_\theta(x|z_k)]p(z_k)}{q_\phi(z_k|x)} \right) \tag{20}$$

Нижней оценкой $\mathcal{L}_q^K(x, \theta, \phi)$ логарифма устойчивого к шуму правдоподобия $\log(\varepsilon + p_\theta(x))$ называется функция

$$\mathcal{L}_q^K(x, \theta, \phi) = \mathop{\mathbb{E}}_{z_1,\dots,z_k \sim q_\phi(z|x)} \log \left(\varepsilon + \frac{1}{K} \sum_{k=1}^{K} \frac{p_\theta(x|z_k)p(z_k)}{q_\phi(z_k|x)} \right) \tag{21}$$

В случае $K = 1$ будем опускать у оценки верхний индекс и будем писать $\mathcal{L}_p(x, \theta, \phi)$ и $\mathcal{L}_q(x, \theta, \phi)$. Логику в названии оценок объясняет следующее наблюдение. Если с помощью неравенства Йенсена внести среднее под логарифм, сразу после применения неравенства для оценки $\mathcal{L}_p^K(x, \theta, \phi)$ скаляр ε будет представлен с помощью среднего по распределению $p(z)$. С другой стороны, для $\mathcal{L}_q^K(x, \theta, \phi)$ скаляр ε будет представлен с помощью среднего по распределению $q_\phi(z|x)$.

Теорема 1. *Для всех (ϕ, θ) и произвольного $K \in \mathbb{N}$ выполнены неравенства*

$$\mathcal{L}_p^K(x, \theta, \phi) \leq \mathcal{L}_q^K(x, \theta, \phi) \tag{22}$$

$$\mathcal{L}_p^K(x, \theta, \phi) \leq \mathcal{L}_p^{K+1}(x, \theta, \phi) \leq \log(\varepsilon + p_\theta(x)) \tag{23}$$

$$\mathcal{L}_q^K(x, \theta, \phi) \leq \mathcal{L}_q^{K+1}(x, \theta, \phi) \leq \log(\varepsilon + p_\theta(x)) \tag{24}$$

Более того, для всех наблюдений x и параметров (θ, ϕ), если $p_\theta(x|z)$ и $\frac{p(z)}{q_\phi(z|x)}$ непрерывны и ограничены относительно аргумента z, то при $K \to \infty$ оценки $\mathcal{L}_p^K(x, \theta, \phi)$ и $\mathcal{L}_q^K(x, \theta, \phi)$ сходятся к устойчивому правдоподобию $\log(\varepsilon + p_\theta(x))$.

Доказательство. Сначала оценим $\mathcal{L}_p^K(x, \theta, \phi) - \mathcal{L}_q^K(x, \theta, \phi)$ сверху. Для удобства введем функцию $s(x, z) = \frac{p(z)}{q_\phi(z|x)}$. Заметим, что $\mathbb{E}_{z \sim q_\phi(z|x)}\, s(x, z) =$

1.

$$\begin{aligned}
&\mathcal{L}_p^K(x,\theta,\phi) - \mathcal{L}_q^K(x,\theta,\phi) = \\
&= \mathop{\mathbb{E}}_{z_1,\ldots,z_K\sim q_\phi(z|x)} \log \frac{\sum_{k=1}^K (\varepsilon s(x,z_k) + p_\theta(x|z_k)s(x,z_k))}{\sum_{k=1}^K (\varepsilon + p_\theta(x|z_k)s(x,z_k))} = \\
&= \mathop{\mathbb{E}}_{z_1,\ldots,z_K\sim q_\phi(z|x)} \log \left[1 + \frac{\sum_{k=1}^K (\varepsilon s(x,z_k) - \varepsilon)}{\sum_{k=1}^K (\varepsilon + p_\theta(x|z_k)s(x,z_k))}\right] \leq \\
&\leq \{\log(1+x) \leq x\} \leq \mathop{\mathbb{E}}_{z_1,\ldots,z_K\sim q_\phi(z|x)} \frac{\sum_{k=1}^K (\varepsilon s(x,z_k) - \varepsilon)}{\sum_{k=1}^K (\varepsilon + p_\theta(x|z_k)s(x,z_k))} \leq \\
&\leq \mathop{\mathbb{E}}_{z_1,\ldots,z_K\sim q_\phi(z|x)} \frac{\sum_{k=1}^K (\varepsilon s(x,z_k) - \varepsilon)}{\sum_{k=1}^K \varepsilon} = \sum_{k=1}^K \frac{\varepsilon - \varepsilon}{K} = 0.
\end{aligned} \tag{25}$$

Следовательно, $\mathcal{L}_p^K(x,\theta,\phi) \leq \mathcal{L}_q^k(x,\theta,\phi)$.

В оставшейся части доказательства мы переносим рассуждения [2] на случай нижних оценок на устойчивое правдоподобие.

Докажем неравенство $\mathcal{L}_q^K(X,\theta,\phi) \leq \mathcal{L}_q^{K+1}(X,\theta,\phi)$. Неравенство $\mathcal{L}_p^K(X,\theta,\phi) \leq \mathcal{L}_p^{K+1}(X,\theta,\phi)$ получается аналогично. Рассмотрим равномерно распределенное случайное подмножество $I \subset \{1,\ldots,K+1\},\ |I| = K$. Для него и произвольного набора чисел $a_1,\ldots,a_{K+1}$ выполнено $\mathbb{E}_{I=\{i_1,\ldots,i_K\}} \left[\frac{a_{i_1}+\cdots+a_{i_K}}{K}\right] = \frac{a_{i_1}+\cdots+a_{i_{K+1}}}{K+1}$. Соотношение вместе с неравенством Йенсена дает

$$\begin{aligned}
\mathcal{L}_q^{K+1} &= \mathop{\mathbb{E}}_{z_1,\ldots,z_{K+1}\sim q_\phi(z|x)} \left[\log\left(\varepsilon + \frac{1}{K+1}\sum_{k=1}^{K+1} \frac{p_\theta(x|z_k)p(z_k)}{q_\phi(z_k|x)}\right)\right] \\
&= \mathop{\mathbb{E}}_{z_1,\ldots,z_{K+1}\sim q_\phi(z|x)} \left[\log \mathop{\mathbb{E}}_{I=\{i_1,\ldots,i_k\}} \left(\varepsilon + \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(x|z_{i_k})p(z_{i_k})}{q_\phi(z_{i_k}|x)}\right)\right] \\
&\geq \mathop{\mathbb{E}}_{z_1,\ldots,z_{K+1}\sim q_\phi(z|x)} \mathop{\mathbb{E}}_{I=\{i_1,\ldots,i_k\}} \left[\log\left(\varepsilon + \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(x|z_{i_k})p(z_{i_k})}{q_\phi(z_{i_k}|x)}\right)\right] \\
&= \mathop{\mathbb{E}}_{z_1,\ldots,z_K\sim q_\phi(z|x)} \left[\log\left(\varepsilon + \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(x|z_k)p(z_k)}{q_\phi(z_k|x)}\right)\right] = \mathcal{L}_q^K
\end{aligned} \tag{26}$$

Неравенство $\mathcal{L}_q^{K+1}(x,\theta,\phi) \leq \log(\varepsilon + p_\theta(x))$ следует из неравенства Йенсена:

$$\mathcal{L}_q^{K+1}(x,\theta,\phi) = \mathop{\mathbb{E}}_{z_1,\dots z_{K+1}\sim q_\phi(z|x)} \log\left(\varepsilon + \frac{1}{K+1}\sum_{k=1}^{K+1}\frac{p_\theta(x|z_k)p(z_k)}{q_\phi(z_k|x)}\right) \leq$$

$$\leq \log \mathop{\mathbb{E}}_{z_1,\dots,z_{K+1}\sim q_\phi(z|x)} \left(\varepsilon + \frac{1}{K+1}\sum_{k=1}^{K+1}\frac{p_\theta(x|z_k)p(z_k)}{q_\phi(z_k|x)}\right) = \log\left(\varepsilon + p_\theta(x)\right) \tag{27}$$

Поскольку $\mathcal{L}_p^K(x,\theta,\phi) \leq \mathcal{L}_q^K(x,\theta,\phi)$, из последнего также следует, что $\mathcal{L}_p^{K+1}(x,\theta,\phi) \leq \log(\varepsilon + p_\theta(x))$.

Докажем сходимость. Поскольку случайные величины $z_1,\dots,z_{K+1}$ независимы и одинаково распределены, а из непрерывности отображения $\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}$ следует его измеримость, случайные величины $\frac{p_\theta(x|z_k)p(z_k)}{q_\phi(z_k)}$, $1 \leq k \leq K+1$ независимы и одинаково распределены. У них также определено математическое ожидание, потому что отношение $\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}$ ограничено. По усиленному закону больших чисел слагаемое $\frac{1}{K}\sum_{k=1}^{K}\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}$ в оценке $\mathcal{L}^K(x,\theta,\phi)$ сходится почти наверное к $\mathbb{E}_{q_\phi(z|x)}\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} = p_\theta(x)$. Из этого следует, что $\mathcal{L}_q^K(x,\theta,\phi) \to \log(\varepsilon + p_\theta(x))$. По тем же соображениям, $\frac{1}{K}\sum_{i=1}^{K}\varepsilon\frac{p(z)}{q_\phi(z|x)} \to \mathbb{E}_{q_\phi(z|x)}\,\varepsilon\frac{p(z)}{q_\phi(z|x)} = \varepsilon$. Следовательно $\mathcal{L}_p^K(x,\theta,\phi) \to \log(\varepsilon + p_\theta(x))$. □

Оценка $\mathcal{L}_q^K(x,\theta,\phi)$ не только точнее $\mathcal{L}_p^K(x,\theta,\phi)$, но также может быть использована для приближенного вариационного вывода. Это свойство нижней вариационной оценки, которое сохраняется при введении параметра ε.

Утверждение 2. *Допустим, что для параметра ϕ^* распределение $q_{\phi^*}(z|x)$ совпадает с апостериорным распределение $p_\theta(z|x)$. В таком случае, для любого $K \in \mathbb{N}$ нижняя оценка $\mathcal{L}_q^K(x,\theta,\phi^*)$ достигает максимального значения $\log(\varepsilon + p_\theta(x))$ при фиксированном θ.*

Доказательство. По построению, нижняя оценка $\mathcal{L}_q^K(x,\theta,\phi^*)$ не превосходит $\log(\varepsilon + p_\theta(x))$. Заменим $q_{\phi^*}(z|x)$ на апостериорное распределение $p_\theta(z|x)$:

$$\mathcal{L}_q^K(x,\theta,\phi^*) = \mathop{\mathbb{E}}_{z_1,\ldots,z_K\sim p_\theta(z|x)} \log\left(\varepsilon + \frac{1}{K}\sum_{k=1}^{K}\frac{p_\theta(x|z_k)p(z_k)}{p_\theta(z_k|x)}\right) =$$
$$= \mathop{\mathbb{E}}_{z_1,\ldots,z_K\sim p_\theta(z|x)} \log\left(\varepsilon + \frac{1}{K}\sum_{k=1}^{K}\frac{\cancel{p_\theta(z_k|x)}p_\theta(x)}{\cancel{p_\theta(z_k|x)}}\right) = \log\left(\varepsilon + p_\theta(x)\right). \tag{28}$$

□

5. Устойчивый к шуму вариационный автокодировщик

Мы предлагаем метод обучения вариационного автокодировщика с помощью нижних оценок на логарифм устойчивого правдоподобия. Полученную таким образом модель мы будем называть устойчивым вариационным автокодировщиком.

Доказанное выше утверждение 1 описывает решения метода максимального устойчивого к шуму правдоподобия, найденного среди всех возможных дискретных распределений. Обученный с помощью нижней оценки на устойчивое правдоподобие вариационный автокодировщик решает ослабленную задачу в менее богатом классе распределений. Тем не менее, выделенные в утверждении свойства метода сохраняются и в случае устойчивого к шуму вариационного автокодировщика. Для упрощения анализа рассмотрим случай $K = 1$ и оценку $\mathcal{L}_q(x,\theta,\phi)$.

Как говорилось раньше, вариационные автокодировщики обучаются с помощью градиентных методов оптимизации. Сравним градиенты нижней оценки на правдоподобие и нижней оценки на устойчивое правдоподобие. Далее с помощью символа ∇ обозначается градиент относительно параметров θ и ϕ. Используя сокращенное обозначение для репараметризации $z_m = f_\phi(\xi_m, x), \xi_m \sim \mathcal{N}(\xi|0,I)$, определим Монте-Карло оценку $\mathcal{L}_q(x,\theta,\phi)$:

$$\mathcal{L}_q(x,\theta,\phi) \approx \frac{1}{M}\sum_{m=1}^{M}\log\left[\varepsilon + \frac{p_\theta(x|z_m)p(z_m)}{q_\phi(z_m|x)}\right] = \bar{\mathcal{L}}_q(x,\theta,\phi), \tag{29}$$

а также оценку $\mathcal{L}(x,\theta,\phi)$.

$$\mathcal{L}(x,\theta,\phi) \approx \frac{1}{M}\sum_{m=1}^{M}\log\left[\frac{p_\theta(x|z_m)p(z_m)}{q_\phi(z_m|x)}\right] = \bar{\mathcal{L}}(x,\theta,\phi) \tag{30}$$

.

Благодаря репараметризации, градиент оценки $\bar{\mathcal{L}}_q(x, \theta, \phi)$ будет несмещенной оценкой на градиент $\mathcal{L}_q(x, \theta, \phi)$. Введем обозначение: $r(x, z) = \log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}$.

Утверждение 3. *Оценка $\nabla\bar{\mathcal{L}}_q(x, \theta, \phi)$ получается из оценки $\nabla\bar{\mathcal{L}}(x, \theta, \phi)$ умножением слагаемых на веса $\sigma(r(x, z_k) - \log\varepsilon)$:*

$$\nabla\bar{\mathcal{L}}_q(x, \theta, \phi) = \frac{1}{M}\sum_{m=1}^{M} \sigma\left(r(x, z_m) - \log\varepsilon\right)\nabla r(x, z_m), \tag{31}$$

где $\sigma(x) = \frac{1}{1+\exp(-x)}$.

В результате при $r(x, z_m) \ll \log\varepsilon, m = 1, \ldots, M$ оценка градиента умножается на веса близкие к нулю. В частности, при $M = 1$ градиенты $\nabla\bar{\mathcal{L}}_q(x, \theta, \phi)$ и $\nabla\bar{\mathcal{L}}(x, \theta, \phi)$ сонаправлены, но их длина отличается в $\sigma(r(x, z_1) - \log\varepsilon)$ раз. Отсюда следует, что в процессе обучения правдоподобие будет быстрее увеличиваться у тех объектов, у которых оно уже достаточно высоко по сравнению с $\log\varepsilon$. Это качественно соответствует свойствам решений из утверждения 1. Аналогичные выводы не удается провести для градиента нижней оценки $\nabla\bar{\mathcal{L}}_p(x, \theta, \phi)$.

Доказательство. Соотношение является простым следствием определения оценки:

$$\begin{aligned}
\nabla\bar{\mathcal{L}}_q(x, \theta, \phi) &= \frac{1}{M}\sum_{m=1}^{M} \nabla \log\left[\varepsilon + \exp(r(x, z_m))\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} \frac{\exp(r(x, z_m))}{\varepsilon + \exp(r(x, z_m))}\nabla r(x, z_m) \\
&= \frac{1}{M}\sum_{m=1}^{M} \frac{1}{\exp(\log\varepsilon - r(x, z_m)) + 1}\nabla r(x, z_m) \\
&= \frac{1}{M}\sum_{m=1}^{M} \sigma\left(r(x, z_m) - \log\varepsilon\right)\nabla r(x, z_m)
\end{aligned} \tag{32}$$

□

Из утверждения 3 также следует, что на ранних этапах обучения высокие значения ε будут препятствовать обучению модели. Действительно, при $r(x, z_m) \ll \log \varepsilon$ веса перед градиентом будут принимать значения близкие к нулю, а шаги градиентного спуска не будут оказывать влияния на параметры модели. С другой стороны, при фиксированном низком значении ε устойчивый к шуму автокодировщик будет обучаться, но не будет качественно отличаться от обычного вариационного автокодировщика.

Для решения этой проблемы мы предлагаем повышать значение ε на этапе обучения модели после каждого градиентного шага. Положим величину ε пропорциональной средней оценке на правдоподобие данных, где α — гиперпараметр метода.

$$\varepsilon = \alpha \exp\left(\frac{\sum_{x \in \mathcal{D}} \mathcal{L}(x, \theta, \phi)}{|\mathcal{D}|}\right). \tag{33}$$

Поскольку вычисление $\sum_{x \in \mathcal{D}} \mathcal{L}(x, \theta, \phi)$ по правилу (33) требовало бы проход по всей обучающей выборке для каждого обновления параметра, в экспериментах эта величина приближалась экспоненциальным сглаживанием несмещенных оценок величины $\sum_{x \in \mathcal{D}} \mathcal{L}(x, \theta, \phi)$.

6. Эксперименты

Для экспериментов была выбрана архитектура вариационного автокодировщика с $d = 50$ скрытыми переменными. Параметры распределений $p_\theta(x|z)$ и $q_\phi(z|x)$ задавались двуслойными полносвязными нейронными сетями с 200 нейронами в каждом слое. Использовалась функция активации нейронов $\mathrm{PReLU}(x) = \max(\delta x, x)$ [4], где параметр δ настраивался совместно с другими параметрами нейронной сети.

Параметры сети настраивались с помощью стохастического градиентного метода оптимизации ADAM [5] с параметрами $\beta_1 = 0.99, \beta_2 = 0.999, \epsilon = 10^{-4}$. Каждый автокодировщик обучался 1000 эпох с коэффициентом скорости обучения 10^{-3}, $n = 200$ элементами в минибатче. Коэффициент устойчивости ε выбирался при помощи экспоненциального сглаживания $\log \varepsilon_t = 0.99 \log \varepsilon_{t-1} + 0.01 \log \varepsilon$ для посчитанного по формуле (33) на t шаге градиентного спуска значения $\log \varepsilon$. Для обучения устойчивого вариационного автокодировщика использовалась нижняя оценка $\mathcal{L}_q$.

Для экспериментального анализа устойчивого к шуму вариационного автокодировщика мы использовали два набора синтетически сгенериро-

ванных данных. Взяв за основу наборы данных MNIST [9] и OMNIGLOT [7], изображения разрешения 28×28 с рукописными символами, мы добавили к данным серые квадраты с интенсивностью, равной средней интенсивности картинок исходных данных. Эксперименты проводились с разным количеством шумовых объектов; их доля в выборке менялась от 1:2 до 2:1.

К изображениям применялась динамическая бинаризация [2]. Каждый пиксель изображения из обучающей выборки принимает значение один с вероятностью, равной интенсивности выбранной точки на исходной картинке. По сравнению со статической (однократной) бинаризацией, этот подход расширяет обучающие данных и повышает обобщающей способности модели. Примеры цифр и шумовых объектов после динамической бинаризации показаны на рисунке 1.

Для оценки качества обученных моделей в качестве приближения правдоподобия вычислялась средняя по тестовой выборке оценка $\mathcal{L}^{200}$, подсчитанная по формуле (10). В тестовую выборку шумовые объекты не добавлялись.

Результаты представлены в таблице 1. Несмотря на шум в данных, устойчивому вариационному автокодировщику удавалось успешно обучиться. В то же время тестовое правдоподобие обычного вариационного автокодировщика значительно ухудшалось с увеличением доли шума в обучающих данных. Из таблицы видно, что оптимальное значение определенного по формуле 33 параметра α существенно зависит от исходного набора обучающих данных и их зашумленности на этапе обучения. Например, в экспериментах на данных OMNIGLOT оптимальный α растет по мере увеличения доли шума. В экспериментах на наборе данных MNIST подобной зависимости на наблюдается.

Мы также сравнили устойчивый вариационный автокодировщик с вариационным автокодировщиком на данных без синтетического шума. Результаты этой серии экспериментов представлены в таблице 2. В этом случае лучшие значения правдоподобия моделей были достигнуты при минимальных значениях α. Тем не менее, устойчивому вариационному автокодировщику удалось достичь небольшого улучшения правдоподобия в сравнении с вариационным автокодировщиком. Это позволяет предположить, что устойчивый вариационный автокодировщик может быть использован для регуляризации генеративных моделей даже при обучении на данных без шума.

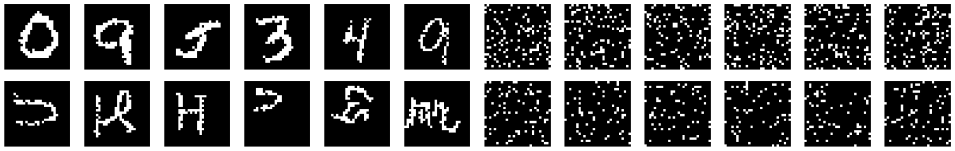

Рис. 1: Примеры данных из обучающей выборки после динамической бинаризации: **сверху** примеры чистых и шумовых объектов для данных MNIST, **снизу** примеры чистых и шумовых объектов для данных OMNIGLOT.

Данные	Ч:Ш	Устойчивый ВАК, $\log \alpha$						ВАК
		-50	-10	0	10	50	100	-
MNIST	2:1	-88.46	-88.47	-88.40	**-88.36**	-88.52	-89.35	-90.64
	1:1	-89.23	**-88.63**	-88.98	-89.07	-88.81	-89.45	-90.96
	1:2	-91.91	-90.53	-89.97	**-89.50**	-90.51	-91.31	-93.75
OMNIGLOT	2:1	-114.32	**-113.72**	-114.73	-114.51	-116.63	-124.13	-115.33
	1:1	-116.98	-115.22	-115.11	**-114.38**	-115.99	-117.16	-117.96
	1:2	-125.67	-124.14	-118.58	-117.92	**-116.03**	-117.38	-121.61

Таблица 1: Логарифм правдоподобия тестовой выборки (больше — лучше) без шумовых объектов. Обучение производилось на данных с синтетическом шумом, отношение ε к среднему правдоподобию данных поддерживалось на уровне α согласно формуле 33. Отношение числа (чистые:шумовые) объектов обучающей выборки приведено во втором столбце. Устойчивый к шуму вариационный автокодировщик успешно справляется с задачей обучения в присутствии шумовых объектов, в то время как правдоподобие вариационного автокодировщика (ВАК) существенно убывает по мере увеличения зашумленности данных.

Данные	Устойчивый ВАК, $\log \alpha$							ВАК
	-100	-50	-10	0	10	50	100	-
MNIST	**-88.04**	-88.44	-92.87	-95.76	-101.61	-127.05	-165.03	-88.74
OMNIGLOT	-113.52	-115.60	-123.78	-126.86	-131.05	-150.59	-183.53	**-112.94**

Таблица 2: Логарифм правдоподобия тестовой выборки (больше — лучше) для устойчивого автокодировщика и вариационного автокодировщика, обученных на данных без синтетического шума. Достаточно малые значения параметра α приводят к эффекту регуляризации и позволяют улучшить результаты на MNIST.

7. Заключение

В данной работе предложено и проанализировано устойчивое к шуму правдоподобие, новый функционал качества для обучения моделей машинного обучения. Представлен способ максимизации устойчивого правдоподобия для модели вариационного автокодировщика при помощи двух вариационных нижних оценок. Доказано, что одна из этих оценок превосходит другую. Экспериментально показано, что устойчивые нижние оценки уменьшают чувствительность вариационного автокодировщика к синтетическому шуму в обучающих данных. В дальнейшем планируется применение предложенного метода к другим моделям машинного обучения.

Авторы выражают признательность анонимному рецензенту за ценные замечания к приведенным в статье рассуждениям, а также за найденные неточности в выкладках.

Список литературы

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

[2] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In International Conference on Learning Representations, 2015.

[3] Michael Figurnov, Kirill Struminsky, and Dmitry Vetrov. Robust variational inference. In Advances in Approximate Bayesian Inference (NIPS Workshop), 2016.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pages 1026–1034, 2015.

[5] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In International Conference on Learning Representations, 2014.

[7] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. Science, 350(6266):1332–1338, 2015.

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436–444, 2015.

[9] Yann LeCun, Corinna Cortes, and Christopher JC Burges. The mnist database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`, 1998.

[10] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In Advances in Neural Information Processing Systems, pages 1073–1081, 2016.

[11] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Proceedings of The 32nd International Conference on Machine Learning, pages 1530–1538, 2015.

[12] Chong Wang and David M Blei. A general method for robust bayesian modeling. arXiv preprint arXiv:1510.05078, 2015.

[13] Yixin Wang, Alp Kucukelbir, and David M Blei. Reweighted data for robust probabilistic models. arXiv preprint arXiv:1606.03860, 2016.

[14] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In European conference on computer vision, pages 818–833. Springer, 2014.