Устойчивый к шуму метод обучения вариационного автокодировщика

Аспирант 2 года обучения Струминский К.А.
Научный руководитель Ветров Д.П.

Мотивация

- Методы обучения без учителя позволяют искать закономерности неразмеченных данных
- Сырые данные могут содержать выбросы, чистка данных может быть трудозатратной
- Мы предлагаем устойчивую к выбросам в данных модификацию вариационного автокодировщика

Доклад по статье М.В. Фигурнова, К.А. Струминского, Д.П. Ветрова “Устойчивый к шуму метод обучения вариационного автокодировщика”,

Вариационный автокодировщик

VAE: генеративная модель

Дано: наблюдения $\mathcal{D} = \{x_1, \ldots, x_N\}, x_i \in \{0, 1\}^D$

Цель: описать распределение данных

Решение: вероятностное моделирование $\mathcal{Z} = \mathbb{R}^d$

$$p(z) = \mathcal{N}(z|0, I)$$

$$p_\theta(x|z) = \prod_{i=1}^{D} \mathrm{Ber}(x_i|t_{\theta,i}(z))$$

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

Kingma D. P., Welling M. Auto-encoding variational bayes //arXiv preprint arXiv:1312.6114. – 2013.

VAE: наивная оценка правдоподобия

Для вычисления вероятности требуется интегрирование

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

Интеграл можно оценить с помощью методов Монте-Карло

$$\int_{\mathcal{Z}} p_\theta(x|z) p(z) dz \approx \frac{1}{M} \sum_{i=1}^{M} p_\theta(x|z_i), \quad z_i \sim p(z)$$

VAE: оценка с выборкой по значимости

Перепишем вероятность

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

При $q_\phi(z|x) = p_\theta(z|x)$ случайная величина имеет нулевую дисперсию

$$\frac{p_\theta(x|z)p(z)}{p_\theta(z|x)} = \frac{p_\theta(x,z)p_\theta(x)}{p_\theta(x,z)} = p_\theta(x)$$

Получили оценку $\int_{\mathcal{Z}} p_\theta(x|z)p(z)dz \approx \frac{1}{M}\sum_{i=1}^{M} \frac{p_\theta(x|z_i)p(z_i)}{q_\phi(z_i|x)}, \quad z_i \sim q_\phi(z|x)$

VAE: нижняя оценка правдоподобия

Оценим правдоподобие снизу

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

При $q_\phi(z|x) = p_\theta(z|x)$ оценка точна

Определение. Нижняя оценка на *обоснованность*

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

Вариационный автокодировщик при обучении максимизирует $\sum_{i=1}^N \mathcal{L}(x_i, \theta, \phi)$

для приближений вида $q_\phi(z|x) = \mathcal{N}(z|\mu_\phi(x), diag(\sigma_\phi(x))$

Оценки с выборкой по значимости

Мы определили

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

Определение. Нижняя оценка оценка с выборкой по значимости

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

Известно, что

$$\log p_\theta(x) \geq \mathcal{L}^K(x, \theta, \phi) \geq \mathcal{L}(x, \theta, \phi)$$

При $K \to \infty$ левое неравенство превращается в равенство

Burda Y., Grosse R., Salakhutdinov R. Importance weighted autoencoders //arXiv preprint arXiv:1509.00519. – 2015.

Устойчивое правдоподобие

Устойчивое к шуму правдоподобие: мотивация

Дано распределение на конечном носителе D меры $\mu(D)$

Для семейства распределений с вероятностью $p_\theta(x), \theta \in \Theta$ определим плотность

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

На качественном уровне определенная вероятность устойчива к выбросам

Устойчивое к шуму правдоподобие: определение

В $p_\theta^\omega(x) = \omega \frac{1}{\mu(\mathrm{D})} + (1-\omega)p_\theta(x)$ сделаем замену $\varepsilon = \frac{\omega}{(1-\omega)\mu(\mathrm{D})}$

Тогда $\sum_{i=1}^N \log p_\theta^\omega(x_i) = \sum_{i=1}^N \log\left[\varepsilon + p_\theta(x)\right] + N\log(1-\omega)$

Определение. Устойчивое к шуму правдоподобие для $\varepsilon > 0$

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

Всегда не меньше правдоподобия

Сравнение с ММП

Пусть наблюдения $x_1, \ldots, x_N$ приходят из конечного множества размера M

Обозначим эмпирические частоты $r_j = \frac{\sum_{i=1}^{N} [x_i = j]}{N}$

Для $p_1, \ldots, p_M$ на множестве $p_i > 0, \sum p_i = 1$

- Метод максимального правдоподобия: $\sum_{j=1}^{M} r_j \log p_j \to max$
- Метод максимального устойчивого правдоподобия:

$$\sum_{j=1}^{M} r_j \log(\varepsilon + p_j) \to max$$

Описание решений

Без ограничения общности частоты r_j упорядочены по возрастанию

Решение ММП: $r_j = p_j, \quad j = 1, \ldots, M$

Решение ММУП: 1) для $\delta(\varepsilon) = \frac{1}{M}\frac{\varepsilon}{1+\varepsilon}$ выполнено

$$r_j < \delta(\varepsilon) \Rightarrow p_j = 0$$

2) у решения сохраняется порядок

$$r_j \leq r_i \Rightarrow p_j \leq p_i$$

Нижняя оценка на устойчивое правдоподобие

УП для модели со скрытыми переменными

Снова модель со скрытыми переменными $p_\theta(x) = \int_{\mathcal{Z}} p_\theta(x|z)p(z)dz$

Выпишем два выражения для устойчивого правдоподобия $\sum_{i=1}^N \log\left[\varepsilon + p_\theta(x_i)\right]$

Вариант 1: $p_\theta^\omega(x, z) = \left[\frac{\omega}{\mu} + (1-\omega)p_\theta(x|z)\right]p(z)$

Вариант 2: $p_\theta^\omega(x, z) = \frac{\omega}{\mu} + (1-\omega)p_\theta(x|z)p(z)$

УП для модели со скрытыми переменными

Снова модель со скрытыми переменными $p_\theta(x) = \int_{\mathcal{Z}} p_\theta(x|z)p(z)dz$

Выпишем два выражения для устойчивого правдоподобия $\sum_{i=1}^N \log\left[\varepsilon + p_\theta(x_i)\right]$

Вариант 1: $p_\theta^\omega(x,z) = \left[\frac{\omega}{\mu} + (1-\omega)p_\theta(x|z)\right]p(z)$

$$L(D) = \sum_{i=1}^N \log\left[\mathbb{E}_{z\sim p(z)}(\varepsilon + p(x_i|z))\right]$$

Вариант 2: $p_\theta^\omega(x,z) = \frac{\omega}{\mu} + (1-\omega)p_\theta(x|z)p(z)$

$$L(D) = \sum_{i=1}^N \log\left[\varepsilon + \mathbb{E}_{z\sim p(z)}p(x_i|z)\right]$$

Нижние оценки на устойчивое правдоподобие

Повторяя вывод важностно-взвешенных оценок получаем две оценки на устойчивое правдоподобие

Определение. Нижняя взвешенная оценка на УП 1

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

Определение. Нижняя взвешенная оценка на УП 2

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

Соотношение между оценками

Теорема. Для K > 0 и ε > 0 выполнено

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

Более того, при $K \to \infty$ оценки $\mathcal{L}_q^K(x,\theta,\phi)$ и $\mathcal{L}_p^K(x,\theta,\phi)$ стремятся к $\log(\varepsilon + p_\theta(x))$

Оценки на УП для приближенного вывода

Утверждение. При $q_\phi(z|x) = p_\theta(z|x)$ оценка $\mathcal{L}_q^K(x,\theta,\phi)$ точна

Доказательство:

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

$$= \mathop{\mathbb{E}}_{z_1,\dots,z_K\sim p_\theta(z|x)} \log\left(\varepsilon + \frac{1}{K}\sum_{k=1}^{K} \frac{\cancel{p_\theta(z_k|x)}p_\theta(x)}{\cancel{p_\theta(z_k|x)}}\right) = \log\left(\varepsilon + p_\theta(x)\right)$$

Устойчивый к шуму вариационный автокодировщик

Выбор параметра ε

Применим оценку $\mathcal{L}_q^1(x,\theta,\phi) = \mathcal{L}_q(x,\theta,\phi)$ для обучения вариационного автокодировщика

Параметр ε меняем динамически согласно правилу

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

На качественном уровне пропорционально среднему значению правдоподобия на данных

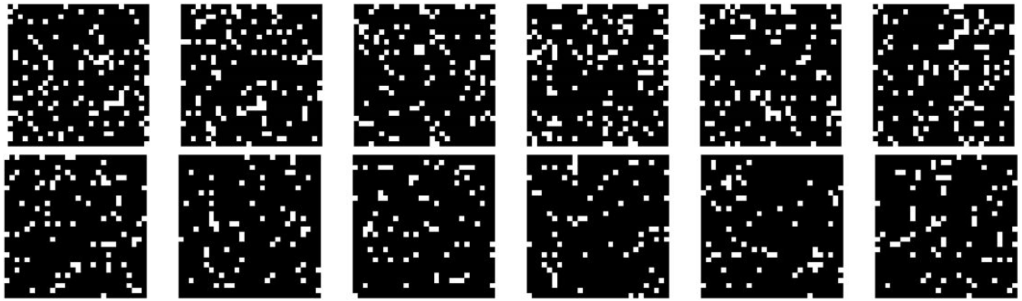
Выбор параметра на основе маргинального правдоподобия реализовать не удалось $\log(\varepsilon + p)$

Данные

Основа: MNIST, OMNIGLOT

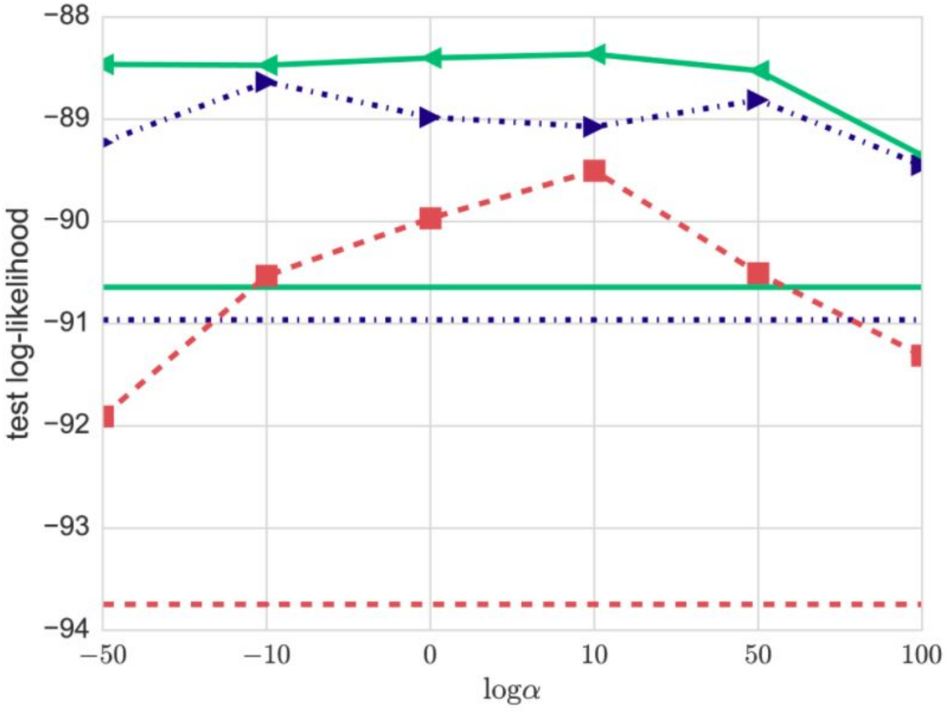
Шум: $x \sim \prod_{i=1}^{D} Ber(x_i|p)$

Доля шумных объектов: 2:1, 1:1, 1:2

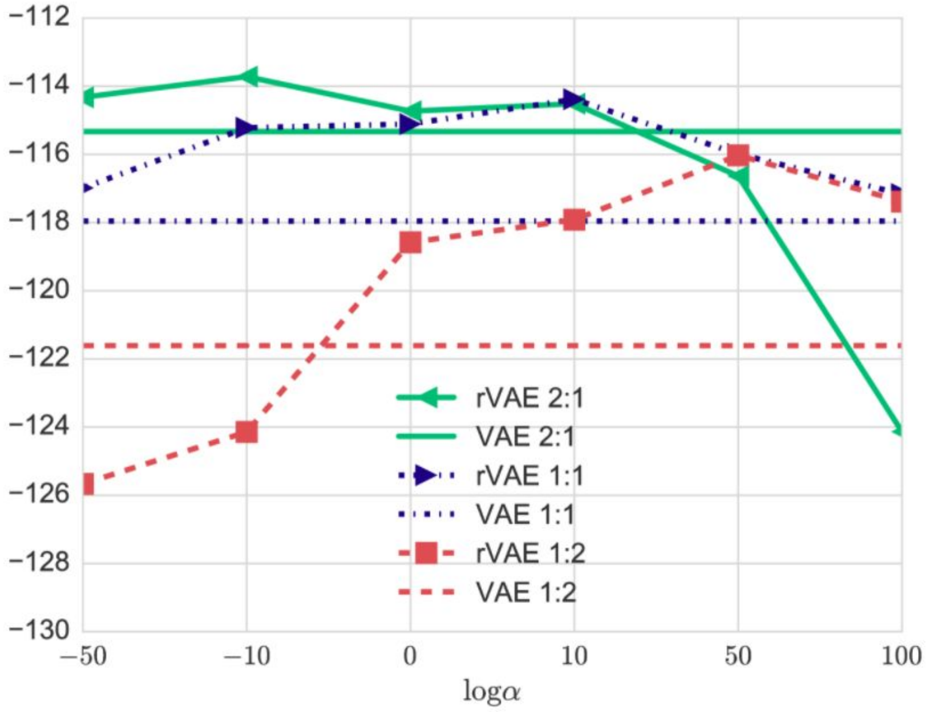


Эксперименты на данных с синтетическим шумом

MNIST



OMNIGLOT



Эксперименты на чистых данных