Vатематическое разложение оценки неопределенности для нейронных сетей.

Насыров Р.Р., Зайцев А.А.

МФТИ

21 декабря 2024 г.

Содержание

Введение

Актуальность

- В чувствительных (здравоохранение, финансы) системах нужно оценивать надежность работы системы
- Нейросеть — один из компонентов системы
- Нужно оценивать неопределенность ответа нейросети

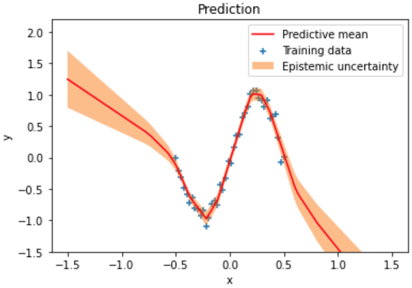

Рис.: Неопределенность.

http://krasserm.github.io/2020/09/25/reliable-uncertainty-estimates/

Модель классификации $h : \mathcal{X} \to \mathcal{Y} \times [0,1]$ называется *скалиброванной*, если предсказываемые вероятности равны реальным вероятностям:

$$\mathbb{P}_{(X,Y)\sim\pi}(\widehat{Y} = Y | \widehat{P} = p) = p \, \forall p \in [0,1]$$

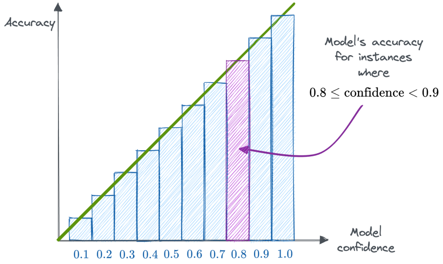

Рис.: Точность и уверенность.

https://www.dailydoseofds.com/a-crash-course-of-model-calibration-classification-models/

Введение

Неопределенность

- *Алеаторная* неопределенность — связана с шумом в данных
- *Эпистемическая* неопределенность — связана с ограничениями модели и доступных данных

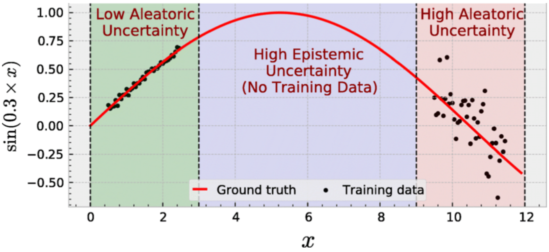

Рис.: 2 типа неопределенности.

https://www.researchgate.net/figure/Illustration-of-epistemic-and-aleatoric-uncertainty$_f ig3_3$58723173

Введение

Новизна

Цель исследования: предложить метод обучения нейросетей, обеспечивающий высокую калибрацию полученной модели без снижения качества классификации.
Задача: используя выведенное разложение полного риска модели, обеспечить лучшую скалиброванность моделей, чем при тренировке с обычным риском.

$$R_{Tot} = R_{Tot}^{OHE} + \langle G'(\widehat{\eta}), ohe(y) - \eta \rangle$$

Постановка задачи

- Решается задача классификации.
- Обозначим

$$D = \{(X_i, Y_i)\}_{i=1}^N; \quad X_i \in \mathbb{R}^d, \;\; Y_i \in \mathcal{Y} = \{1, \dots, K\}$$

— i.i.d. данные из истинного распределения

$$P(x, y) = P(x)P(y|x)$$

.

- Обозначим истинное условное распределение и модельное как

$$\eta(x) = \mathbb{P}(Y|X = x); \quad \widehat{\eta}(x)_\theta = \widehat{\eta}(x) = \mathbb{P}(Y|X = x, \theta)$$

соответственно.

- Обозначим лосс функцию

$$l : \Delta^K \times \mathcal{Y} \to \mathbb{R}$$

Решение задачи. Обозначения

- Proper scoring rule — такое семейство лосс функций $l(P, y) : \mathcal{P}_K \times Y \to \mathbb{R}$, для которых выполнено:

$$\int l(P, y) dQ(y) = l(P, Q) \geq l(Q, Q) \;\; \forall P, Q \in \Delta^K$$

- Под некоторыми условиями на l, любой такой лосс представим в виде

$$l(\eta, y) = \langle G'(\eta), \eta \rangle - G'_y(\eta) - G(\eta)$$

для какой-то выпуклой функции

$$G : \mathcal{P}_K \to \mathbb{R}$$

- Введем обозначения полного (total) риска и частного (OHE) риска:

$$R_{Tot} = \langle l(\widehat{\eta}), \eta \rangle$$

$$R_{Tot}^{OHE} = l(\widehat{\eta}, y) = \langle l(\widehat{\eta}), ohe(y) \rangle$$

Решение задачи. Вывод формулы

Распишем формулу OHE риска для proper scoring rule:

$$R_{Tot}^{OHE} = R_{Tot} + \langle I(\widehat{\eta}), ohe(y) - \eta \rangle$$

$$\langle I(\widehat{\eta}), ohe(y) - \eta \rangle = \langle (\langle G'(\widehat{\eta}), \widehat{\eta} \rangle - G(\widehat{\eta})) \cdot \mathbb{1} - G'(\widehat{\eta}), ohe(y) - \eta \rangle =$$

$$= (\langle G'(\widehat{\eta}), \widehat{\eta} \rangle - G(\widehat{\eta})) \cdot \langle \mathbb{1}, ohe(y) - \eta \rangle - \langle G'(\widehat{\eta}), ohe(y) - \eta \rangle =$$

$$-\langle G'(\widehat{\eta}), ohe(y) - \eta \rangle$$

Итого, получим:

$$R_{Tot} = R_{Tot}^{OHE} + \langle G'(\widehat{\eta}), ohe(y) - \eta \rangle$$

Решение задачи. Аппроксимация риска

Имеем:

$$R_{Tot} = R_{Tot}^{OHE} + \langle G'(\widehat{\eta}), ohe(y) - \eta \rangle$$

$$s = ohe(y) - \eta$$

Изучим свойства s:

1. $\langle \mathbb{1}, s \rangle = 0$.
2. $\exists! i : s_i \geq 0 \wedge \forall j \neq i : s_j \leq 0$. $s_i \in [-1, 0], s_j \in [0, 1]$.
3. Нет градиентов s по $\widehat{\eta}$.
4. $\mathbb{E}_{y \sim \eta(y|x)} s = 0$, т.к. $\mathbb{E}_{y \sim \eta(y|x)} ohe(y) = \eta$.

Решение задачи. Аппроксимация риска

3 варианта выбора s, удовлетворяющих свойствам выше:

1. (Равномерное)

$$s_k = \begin{cases} \varepsilon(1 - \frac{1}{K}), & \textit{if } \ y = k \\ -\frac{\varepsilon}{K}, & \text{иначе} \end{cases}$$

2. (Априорное)

$$s_k = \begin{cases} \varepsilon(1 - \frac{N_k}{N}), & \textit{if } \ y = k \\ -\varepsilon\frac{N_k}{N}, & \text{иначе} \end{cases}$$

3. (Предсказанное)

$$s_k = \begin{cases} \varepsilon(1 - SG(\widehat{\eta}_k)), & \textit{if } \ y = k \\ -\varepsilon SG(\widehat{\eta}_k), & \text{иначе} \end{cases}$$

Гипотезы

Гипотезы к проверке:

1. Addition term помогает построить лучше откалиброванную модель. Таким образом, с этим членом ECE должен быть меньше.
2. Accuracy не сильно снижается при введении addition term. Таким образом, модель остается точной.
3. ECE модели является выпуклой функцией эпохи/точности: сначала она уменьшается, затем увеличивается. Это соответствует следующему утверждению: в начале обучения OHE risk направлен на калибровку, а к концу он подталкивают модель к тому, чтобы быть менее калиброванной и более точной.

Результаты экспериментов. CE loss

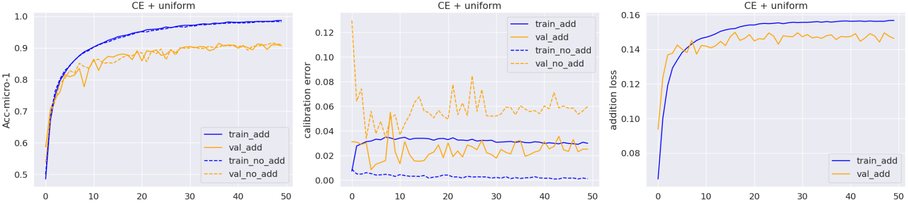

Рис.: Cross Entropy Loss and Uniform addition

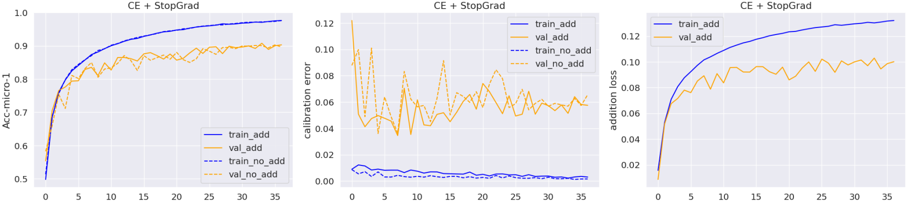

Рис.: Cross Entropy Loss and StopGrad addition

Результаты экспериментов. Brier loss

Рис.: Brier Loss and Uniform addition

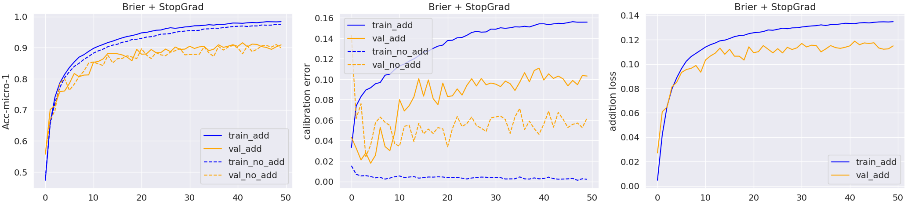

Рис.: Brier Loss and StopGrad addition

Выводы

- Гипотеза №2 (про Accuracy) подтвердилась
- Гипотеза №3 (про выпуклость ECE) подтвердилась частично, больше экспериментов нужно провести.
- Гипотеза №1 (про калибрационные свойства addition term) под вопросом.

Следующие шаги

- Провести модельный эксперимент для экспериментального подтверждения гипотез
- Сформулировать и доказать теорему о выпуклости ECE
- Подтвердить гипотезы с помощтю построения доверительных интервалов на картиночных данных.

Литература

1. Nikita Durasov et al. "Zigzag: Universal sampling-free uncertainty estimation through two-step inference". In: arXiv preprint arXiv:2211.11435
2. Chuan Guo et al. "On calibration of modern neural networks". In: International conference on machine learning. PMLR. 2017, pp. 1321–1330.
3. Nikita Kotelevskii and Maxim Panov. "Predictive Uncertainty Quantification via Risk Decompositions for Strictly Proper Scoring Rules". In: arXiv preprint arXiv:2402.10727
4. Jeremy Nixon et al. "Measuring Calibration in Deep Learning." In: CVPR workshops. Vol. 2. 7. 2019.
5. Cheng Wang. "Calibration in deep learning: A survey of the state-of-the-art". In: arXiv preprint arXiv:2308.01222 (2023).