FY2025/2025年度

Uncertainty Quantification Team
Futoshi Futami
不確実性定量化チーム　二見 太

RIKEN

AIP Center for Advanced Intelligence Project

Calibration of classification models

Basics

- **Calibration**: how well the model's predictive probability aligns with the actual frequency of the true label.
  - Binary classification $\mathcal{Y} = \{0, 1\}$
  - ML models returns the predicted probability of $f_w : \mathcal{X} \to [0, 1]$
  - Perfect calibration $\Pr(Y = 1 | f_w(X) = p) = p$ for all p a.s.

We evaluate this using **expected calibration error (ECE)**

$$\mathrm{ECE}(f_w) := \mathbb{E}|\Pr[Y = 1 | f_w(X)] - f_w(X)|$$

After standard training, models are often not well calibrated.

Q) How to estimate the ECE accurately using finite samples?

Q) How to improve the calibration by post-processing the hypothesis ?

Recalibration algorithm by PAC-Bayes

- A PAC-Bayesian generalization error analysis for the ECE.
- A recalibration algorithm based on Gaussian process derived by minimizing an upper bound on the PAC-Bayesian bound.

Generalization of ECE.

While the ECE is small on the training dataset, what can be said about
(1) an unseen test dataset?
(2) the population distribution?

PAC-Bayes bound (2)

$$\mathbb{E}_{V\sim\tilde{\rho}}|\mathrm{ECE}(\eta_V \circ f_w) - \widehat{\mathrm{ECE}}(\eta_V \circ f_w, S_{\mathrm{re}})| \leq \frac{1+L}{B} + \frac{\mathrm{KL}(\tilde{\rho}\|\tilde{\pi}) + B\log 2 + \log\frac{1}{\varepsilon} + \frac{2\lambda^2}{n_{\mathrm{re}}}}{\lambda}$$

Hypothesis **after recalibration**　Hypothesis **before recalibration**

Binning ECE

- We split the [0,1] into B intervals $\mathcal{I} = \{I_1, \ldots, I_B\}$
- Aggregate the values allocated into the same bins. (f and Y).

We use **a Gaussian process** as the recalibration model and optimize it in a manner analogous to **variational inference**.

M. Fujisawa* & **F. Futami***. PAC-Bayes Analysis for Recalibration in Classification. In International Conference on Machine learning, 2025. (ICML 2025)

Estimating ECE for PU setting

- Constructing an estimator ECE using only **PU data**.
- Analyzing the bias and generalization error of the estimator.
- Clarifying the effect of class-prior estimation bias in ECE.

$$S_{\mathrm{PU}} = \{X_m^{\mathrm{trU}}\}_{m=1}^{n_U} \sim P(X)^{n_U} \bigcup \{X_m^{\mathrm{trP}}\}_{m=1}^{n_P} \sim P(X|Y=1)^{n_P}$$

Proposed estimator: PU-ECE

$$\mathrm{ECE}_{\mathrm{PU}}(f, S_{\mathrm{PU}}) := \sum_{b=1}^{B}\left|\frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}}\sum_{m=1}^{n_{\mathrm{P}}}\mathbb{1}_{f(X_m^{\mathrm{P}})\in I_b} - \frac{1}{n_{\mathrm{U}}}\sum_{m=1}^{n_{\mathrm{U}}}\mathbb{1}_{f(X_m^{\mathrm{U}})\in I_b} f(X_m^{\mathrm{U}})\right|$$

$\pi_{\mathrm{P}} = P(Y=1)$: class prior, $S_{\mathrm{PU}} = \left(\left\{X_1^{\mathrm{U}}, \ldots, X_{n_{\mathrm{U}}}^{\mathrm{U}}\right\}, \left\{X_1^{\mathrm{P}}, \ldots, X_{n_{\mathrm{P}}}^{\mathrm{P}}\right\}\right)$: PU evaluation data

Rewrite supervised ECE as

$$\sum_{b=1}^{B} \boxed{P(f(X)\in I_b)}\left|\boxed{\mathbb{E}[f(X) \mid f(X)\in I_b]} - \underbrace{P(Y=1 \mid f(X)\in I_b)}_{=\frac{P(Y=1)\,P(f(X)\in I_b \mid Y=1)}{P(f(X)\in I_b)}}\right|$$

- (green) can be estimated from positive data
- (blue) can be estimated from unlabeled data
- (grey) is known or can be estimated

Under the same conditions as in PN learning

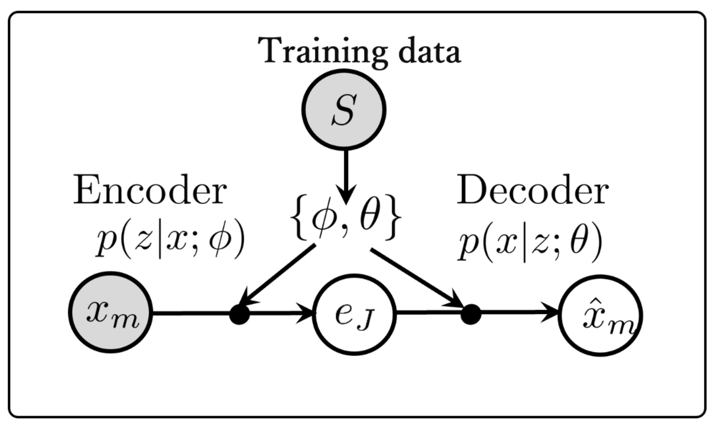
$$\mathrm{Bias}_{\mathrm{tot}}(f, S_{\mathrm{PU}}) := |\mathrm{ECE}_{\mathrm{PU}}(f, S_{\mathrm{PU}}) - \mathrm{TCE}(f)| = \mathcal{O}\left(\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}} + \frac{1}{n_{\mathrm{U}}}\right)^{\frac{1}{3}}\right)$$

R. Kiryo, **F. Futami**, M. Sugiyama, Estimating Expected Calibration Error for Positive-Unlabeled Learning. Transactions on Machine Learning Research, 2025.

Probabilistic models

Information theoretic analysis of VQ-VAE

- We propose a new **information-theoretic generalization analysis for VQ-VAEs** that elucidates the behavior of latent variables.
- The resulting **bound depends only on encoder overfitting and the complexity of the LVs, with no explicit dependence on the decoder.**
- We also analyze the generation performance.



- **Encoder-Decoder model：** Encoder network compresses the data into **latent variables (LVs), a.k.a representations.**
- Many learning algorithms use regularization over LVs.

$\min$ **Training error** + **Complexity of LVs**

Network parameters

Example
- Variational inference $\mathrm{KL}(p(Z|X;\phi)|p(Z))$
- Information bottleneck $I(Z;X)$

Q). Regularizing the LVs imply generalization ?

Information-theoretic generalization analysis

- We introduced a new data-dependent prior distribution over LVs,

Generalization of the reconstruction loss: the squared-loss discrepancy between the reconstructed data and the original data.

$$\mathrm{gen}(n, \mathcal{D}) \leq 3\Delta\sqrt{\frac{I(\tilde{\mathbf{J}};\mathbf{T}|\mathbf{e},\phi,\tilde{X}) + I(\mathbf{e},\phi;S)}{n}} + \frac{\Delta}{\sqrt{n}}.$$

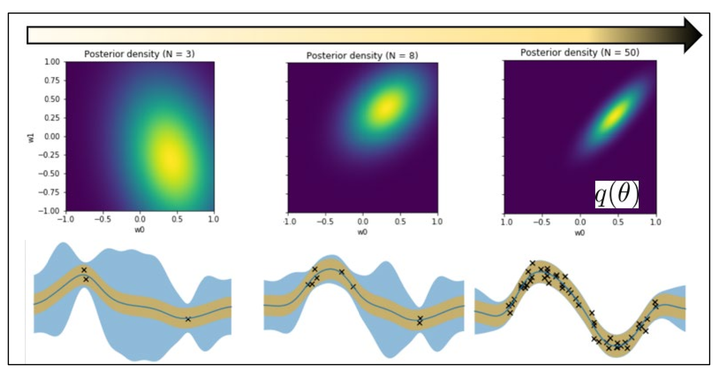
Complexity of **LVs**　Complexity of **Encoder network**

- Controlling the complexity of **both LVs and the encoder network** is crucial, whereas the complexity of the decoder has minimal effect.

F. Futami* & M. Fujisawa*. Information-theoretic Generalization Analysis for VQ-VAEs: A Role of Latent Variables. Neural Information Processing Systems, 2025. (NeurIPS 2025)

Evaluation of Epistemic uncertainty by PAC-Bayes

- We analyze common Bayesian uncertainty measures via PAC-Bayes.
- We show they lower-bound the excess risk and study their sample-size dependence.
- We propose a variational Bayesian method to control epistemic uncertainty.



Aleatoric uncertainty： Inherent noise in the data-generating process (fundamental difficulty of learning)

Epistemic uncertainty: Insufficient training data (when the model is well-specified)

$$\mathrm{Total\ risk} = \underbrace{\mathrm{Bayes\ risk}}_{\mathrm{Aleatoric}} + \underbrace{\mathrm{Excess\ risk}}_{\mathrm{Epistemic\ uncertainty}}$$

Widely used measures of predictive uncertainty lower bounds the excess risk

- **the predictive mean** ⇔ **squared loss**
- **the MI between the prediction and model** ⇔ **Log loss**

$$\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N) \leq 2\Big(\underbrace{\sigma^2\,\mathrm{KL}(\nu(Y|x)\,\|\,p(Y|x,\theta^*))}_{\text{Error due to model misspecification}} + \underbrace{\sigma^2\,\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*)}_{\text{Error due to insufficient data}}\Big)^{\frac{1}{2}}$$

MI optimistically evaluates the excess risk, assuming that our predictive model is correct.

F. Futami. Epistemic Uncertainty and Excess Risk in Variational Inference. In Artificial Intelligence and Statistics, 2025. (AISTATS 2025)

Convergence analysis of SVGD

- We present a method for analyzing convergence to the posterior in SVGD, an approximate Bayesian inference method.
- We propose a new approach for handling approximation error induced by projection.

M. Fujisawa* & **F. Futami***. On the Convergence of SVGD in KL divergence via Approximate gradient flow. Transactions on Machine Learning Research, 2025.