Assorted Topics (3)

CS772A: Probabilistic Machine Learning

Piyush Rai

Plan today

- A method for distribution-free uncertainty quantification
  - Conformal Prediction

- Getting uncertainty estimates that we can trust (e.g., if model is overconfident)
  - Model Calibration

Conformal Prediction

Conformal Prediction

- A simple technique to easily obtain confidence intervals
  - In classification, such an interval may refer to the set of highly likely classes for a test input

  - For more difficult test inputs, the set would typically be larger
- In a way, conformal prediction gives predictive uncertainty
  - However, unlike Bayesian ML, we don't get model uncertainty
  - Only one model is learned in the standard way and we construct the set of likely classes
  - It's like a black-box method; no change to training procedure for the model

Conformal Prediction

Assume it's a classification model which produces softmax scores

Conformal prediction can be used for regression problems too*

- Assume we already have a trained model $\hat{f}$ using some labelled data
- Suppose we get a test input X_{test} whose true (unknown) label is Y_{test}
- Use $\hat{f}$ and a calibration set of n examples to generate a prediction set $\mathcal{C}(X_{test})$ s.t.

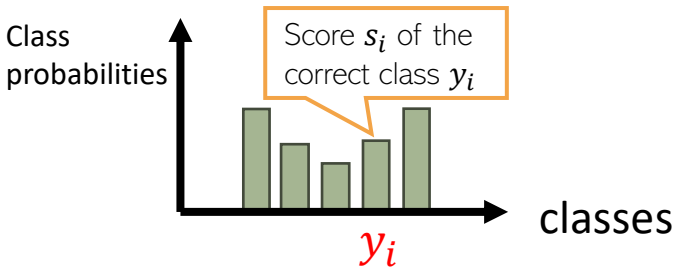
α is a user chosen error rate

$$1 - \alpha \leq p\left(Y_{test} \in \mathcal{C}(X_{test})\right) \leq 1 - \alpha + \frac{1}{n+1}$$

- To construct the set, we first compute, for each example in the calibration set

Probability/score of the correct class y_i of the input x_i

$$s_i = \hat{f}(x_i)_{y_i}$$

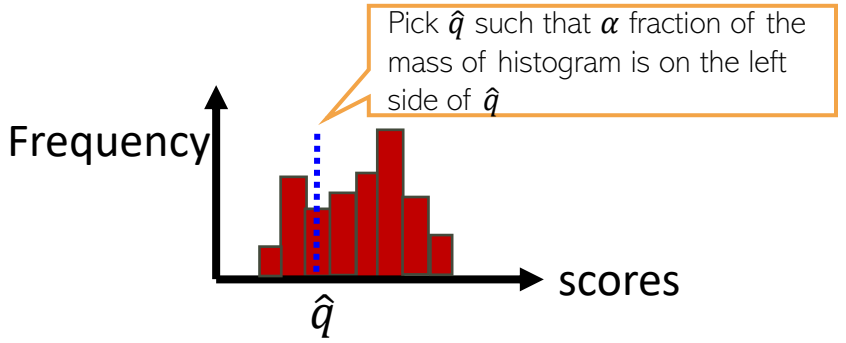


- Use the calibration set scores $s_1, s_2, \ldots, s_n$ to compute their α quantile

*A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification (Angelopoulos and Bates, 2022)

Conformal Prediction

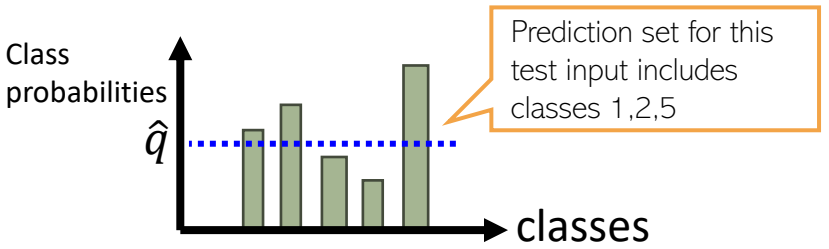
- Assume the α (say 0.1) quantile of the calibration set scores is equal to $\hat{q}$



- Assuming n is very large, roughly $(1-\alpha)$ fraction of inputs will have score higher than $\hat{q}$

Probability of true class

- Given a test input X_{test}, whose label is is unknown, we compute the class probabilities



- Define the prediction set for X_{test} as

Report all the classes whose probability is large enough (the "large enough" value is given by the α quantile $\hat{q}$)

$$\mathcal{C}(X_{test}) = \{y: \hat{f}(X_{test})_y \geq \hat{q}\}$$

Conformal Prediction

- A generic black-box method
- Can be easily applied to any already trained classifier
- Predicted set has some nice guarantees

$$1 - \alpha \leq p\left(Y_{test} \in \mathcal{C}(X_{test})\right) \leq 1 - \alpha + \frac{1}{n+1}$$

- Does not make any assumptions on the distribution of the data
  - Thus considered a "distribution-free" approach to uncertainty quantification
- Can also be applied to regression problems*

*A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification (Angelopoulos and Bates, 2022)

Model Calibration

Calibration

- A model called calibrated if predicted class probabilities match empirical frequencies
- Example: For binary classification, if for all test examples for which the model predicts $p(y = 1|x) = 0.8$, about 80% have true label = 1, then this model is well-calibrated
- Expected Calib. Error (ECE) is a popular measure of model calibration
- Suppose $f(x)_c = p(y = c|x)$, $\hat{y}_n = \text{argmax}_{c=\{1,2,\ldots,C\}} f(x_n)_c$, $\hat{p}_n = \max_{c=\{1,2,\ldots,C\}} f(x_n)_c$
- Suppose predicted probabilities are divided into B bins
- Assume $\mathcal{B}_b$ as set of samples whose predicted probabilities fall in $I_b = (\frac{b-1}{B}, \frac{b}{B}]$

$$\text{acc}(\mathcal{B}_b) = \frac{1}{|\mathcal{B}_b|} \sum_{n \in \mathcal{B}_b} \mathbb{I}(\hat{y}_n = y_n) \quad \text{conf}(\mathcal{B}_b) = \frac{1}{|\mathcal{B}_b|} \sum_{n \in \mathcal{B}_b} \hat{p}_n$$

$$\text{ECE}(f) = \sum_{b=1}^{B} \frac{|\mathcal{B}_b|}{B} |\text{acc}(\mathcal{B}_b) - \text{conf}(\mathcal{B}_b)|$$

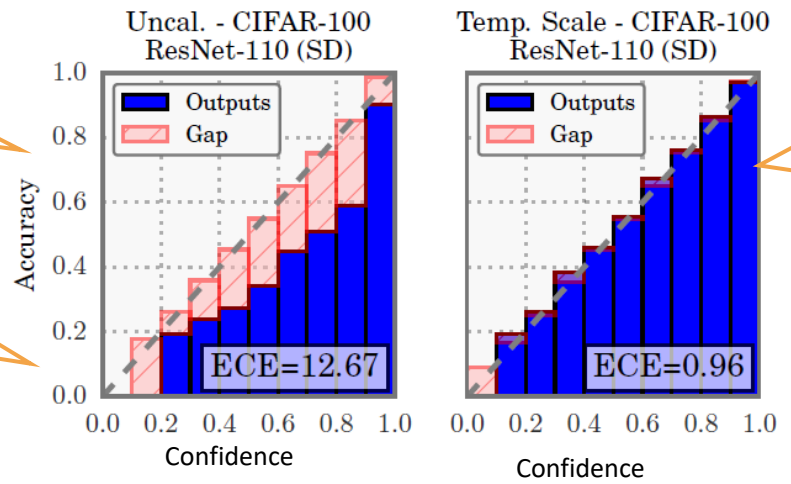
Difference between confidence and accuracy

Calibration

- A reliability diagram is often used as a visual indicator of calibration

ECE is the average "gap" area in the reliability diagram

Reliability diagram of an uncalibrated model



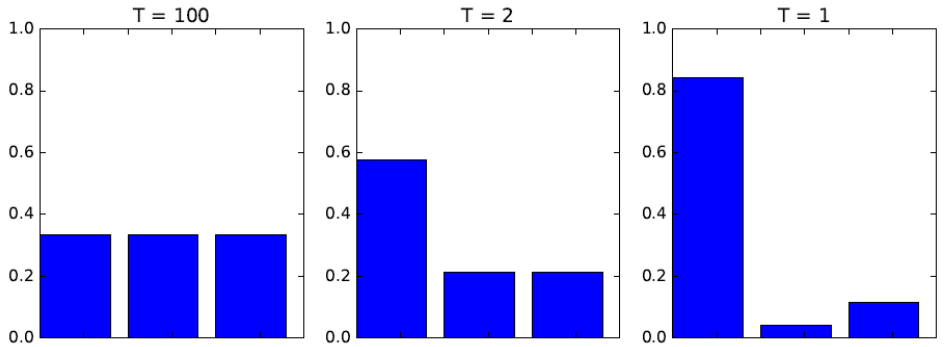
Reliability diagram of the same model after applying calibration post-processing via temperature scaling method

- Several approaches to improve a model's calibration
- A simple calibration approach for binary classifiers is Platt scaling
  - Rescale the logits z where $p = \sigma(z)$ as $az + b$
  - Learn a, b by doing MLE on a validation set
  - Calibrated probabilities will be $\hat{p} = \sigma(z)$
  - Can be extended to multi-class case as well

Calibration

- Temperature scaling is another simple and popular calibration method

"Temperature scaling" of softmax outputs as $softmax(a/T)$ is a popular and simple approach to reduce overconfidence (for figure on right, $a = [3,0,1]$)



- Histogram Binning, Label Smoothing, etc are also popular,
- All these methods can be applied as post-processing step to the outputs of Bayesian/non-Bayesian methods to improve calibration
- Bayesian methods are usually better calibrated but can still have poor calibration if test data is from a different distribution

Proper Scoring Rules and Calibration

- Assume a predictive distribution $p_\theta(y|x)$
- Define score of p_θ on an example $(x, y) \sim p^*(x, y) = p^*(x)p^*(y|x)$ as $s(p_\theta, (x, y))$
- The expected score of p_θ will be $s(p_\theta, p^*) = \int p^*(x)p^*(y|x)s\big(p_\theta, (x, y)\big)dydx$
- A scoring rule is said to be a "proper scoring rule" if $s(p_\theta, p^*) \leq s(p^*, p^*)$
- The log-likelihood $s\big(p_\theta, (x, y)\big) = \log p_\theta(y|x)$ is a proper scoring rule because

$$S(p_{\boldsymbol{\theta}}, p^*) = \mathbb{E}_{p^*(\boldsymbol{x})p^*(y|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})\right] \leq \mathbb{E}_{p^*(\boldsymbol{x})p^*(y|\boldsymbol{x})}\left[\log p^*(y|\boldsymbol{x})\right]$$

Holds because of Gibbs inequality – entropy less than or equal to cross-entropy

- Optimizing a proper scoring rule (e.g., NLL) should do the "right thing"
- Another proper scoring rule is the Brier score (lower is better)

$$S(p_{\boldsymbol{\theta}}, (y, \boldsymbol{x})) \triangleq \frac{1}{C}\sum_{c=1}^{C}(p_{\boldsymbol{\theta}}(y = c|\boldsymbol{x}) - \mathbb{I}\,(y = c))^2$$

If using such loss functions, the model will match true probabilities and be well-calibrated

But doesn't happen in practice due to optimization related issues, training set characteristics, etc

Can use Brier score (and also NLL) as a way to measure calibration of a model

Squared error of predictive distribution as compared to one-hot vector