Deriving the calibration scaling term for a deep ensemble

Stephen Malina

December 10, 2020

Assume that we've trained m regression models, denoted f_i for $i \in [m]$, with different random seeds. Let $\mu(x_i)$ denote the ensemble mean,

$$\frac{1}{m} \sum_{i=1}^{m} f_i(x_i),$$

and $\sigma^2(x_i)$ the predictive variance,

$$\frac{1}{m} \sum_{i=1}^{m} (f_i(x_i) - \mu(x_i))^2,$$

for a single data point.

As is often done for ensembles, suppose that

$$y_i \sim \mathcal{N}(\mu(x_i), \lambda \cdot \sigma^2(x_i)).$$

We'll now derive a formula for λ that maximizes the log-likelihood of **y**.

The log-likelihood is

$$\mathcal{L} = \sum_{i=1}^{n} \log p(y_i \mid \mu(x_i), \sigma^2(x_i)) \tag{1}$$

$$= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \lambda - \sum_{i=1}^{n} \log(\sigma(x_i)) - \sum_{i=1}^{n} \left[\frac{1}{2 * \lambda * \sigma^2(x_i)} (x_i - \mu(x_i))^2 \right]. \tag{2}$$

Its derivative with respect to λ is

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -\frac{n}{2\lambda} + \sum_{i=1}^{n} \frac{1}{2 * \lambda^2 * \sigma^2(x_i)} (x_i - \mu(x_i))^2. \tag{3}$$

Setting this to 0 and solving gives us

$$-\frac{n}{2\lambda} + \sum_{i=1}^{n} \frac{1}{2 * \lambda^2 * \sigma^2(x_i)} (x_i - \mu(x_i))^2 = 0 \tag{4}$$

$$\sum_{i=1}^{n} \frac{1}{2 * \lambda^2 * \sigma^2(x_i)} (x_i - \mu(x_i))^2 = \frac{n}{2\lambda} \tag{5}$$

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sigma^2(x_i)} (x_i - \mu(x_i))^2 = \lambda. \tag{6}$$

Intuitively, this means that λ maximizes the log-likelihood when it's set to the average variance-weighted mean-squared error.