Work in Progress

Quantification vs. Reduction: On Evaluating Regression Uncertainty

Domokos M. Kelen[1], Adám Jung[1], András A. Benczúr[1]

[1]HUN-REN SZTAKI

MOTIVATION & SUMMARY

- Traditionally, the quality of uncertainty distributions is evaluated by **Negative Log-Likelihood**
- We argue that this is problematic: NLL measures RMSE and calibration at the same time, **conflating the two**
- We demonstrate the effect through **analytical examples**

TRADITIONAL EVALUATION

- Negative Log-Likelihood (or other scoring rule):

$$-\mathbb{E}\Big[\log\, \tilde{q}_\theta(Y \mid X)\Big] \approx -\frac{1}{n}\sum_{i=1}^{n} \log\, \tilde{q}_\theta(y_i \mid x_i)$$

  Strictly Proper Scoring rule theory: minimum value exactly when predicted distribution matches target

- RMSE

DETERMINISTIC EXAMPLE

UNCERTAINTY = MODEL ERROR

- Assume **no aleatoric uncertainty**: $Y = \mathbb{E}[Y \mid X]$
- Model has to learn a **function** $\tilde{q}_\theta(\mu(x) \mid X = x)$
- The distribution **changes** with the accuracy of the model: *Quantifying* uncertainty just means the model estimating **its own error**.
- However, this **target distribution is not fixed**, therefore scoring rule theory doesn't apply
- In general, the minimum possible NLL is the **entropy**, and $$H(\alpha X) = H(X) + \log|\alpha|$$

GAUSSIAN EXAMPLE

- Further assume a **Gaussian** distribution:

$$\tilde{q}_\theta(y \mid X = x) \;=\; \mathcal{N}\Big(y \,\big|\, \tilde{\mu}(x;\theta),\; \tilde{\sigma}^2(x;\theta)\Big)$$

- Denote the **residual** as

$$\varepsilon(x) \;=\; Y - \tilde{\mu}(x;\theta)$$

- Then the NLL can be expressed as

$$-\mathbb{E}\big[\log \tilde{q}_\theta(Y \mid X)\big] \;=\; -\mathbb{E}\Big[\log\Big(\mathcal{N}\big(\varepsilon(X)\,\big|\,0,\; \tilde{\sigma}^2(X;\theta)\big)\Big)\Big]$$

- However, the residual itself is is **defined by the model's chosen prediction function**. As the residual changes, so does the uncertainty distribution that we compare aga-inst.

GENERAL CASE

- In general, we assume a **parametric** aleatoric distribution:

  $Y \mid X$ is characterized by some parameter vector $\mathbf{r}_x = (r_x^{(1)}, \ldots, r_x^{(n)})$

- We need predictive **distributions for the parameters**:

$$p(y \mid X = x, \theta) = \int_{\mathbb{R}^n} p(y \mid r)\, p(r \mid \theta, X = x)\, \mathrm{d}r$$

- The target $\mathbf{r}$ can still be assumed **deterministic** given $\mathbf{X}$
- **Larger error** in $\mathbf{r}$ still manifests as **uncertainty**
- However the overall effect is **unclear**, as certain parameters values themselves affect the resulting uncertainty

STUDENT-T EXAMPLE

- In an analytically tractable example, we show that overall variance **only depends on the expected value of the variance parameter** (and not its uncertainty)

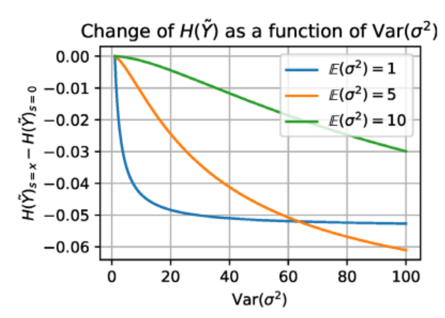
Proposition A. *Assume a normal aleatoric uncertainty distribution with $\mathbf{r}_x = (\mu_x, \sigma_x^2)$ and suppose we have the true mean function $\mu(x) = \mu_x$ at our disposal. Further assume that our prediction for the variance σ_x^2 follows an inverse-gamma distribution $\Gamma^{-1}(\alpha, \beta)$. Then denoting the marginal distribution of the predicted variable $\tilde{Y}_x$,*

$$\mathrm{Var}(\tilde{Y}_x) = \mathbb{E}\left(\Gamma^{-1}(\alpha,\beta)\right) = \frac{\beta}{\alpha - 1}. \qquad (10)$$

- Essentially the **law of total variance** from the model's perspective: $\mathrm{Var}(\tilde{Y}) = \mathbb{E}\Big[\mathrm{Var}\Big(\tilde{Y} \mid X\Big)\Big] + \mathrm{Var}\Big[\mathbb{E}\Big(\tilde{Y} \mid X\Big)\Big]$
- However, the uncertainty of the variance does change the shape of the distribution: **increasing uncertainty** results in **decreased entropy** in the normal-inverse gamma case

Proposition C. *Assume a normal aleatoric uncertainty distribution with $\mathbf{r}_x = (\mu_x, \sigma_x^2)$ and suppose we have the true mean function $\mu(x) = \mu_x$ at our disposal. Further assume that our prediction for the variance σ_x^2 follows an inverse-gamma distribution $\Gamma^{-1}(\alpha, \beta)$. Then denoting the marginal distribution of the predicted variable $\tilde{Y}_x$, the differential entropy $H(\tilde{Y}_x)$ decreases monotonically as the variance of the distribution $\Gamma^{-1}(\alpha, \beta)$ increases.*

- This is highly **counter-intuitive**, but true:



- Essentially: *If we have an* ***unbiased variance estimate****, the* ***uncertainty*** *of the variance only changes the* ***shape****.*
- However, with only one sample y per prediction, the distributions can always be considered **unbiased**
- Implies that the **marginal variance** is a good estimate

NUMERIC EVALUATION

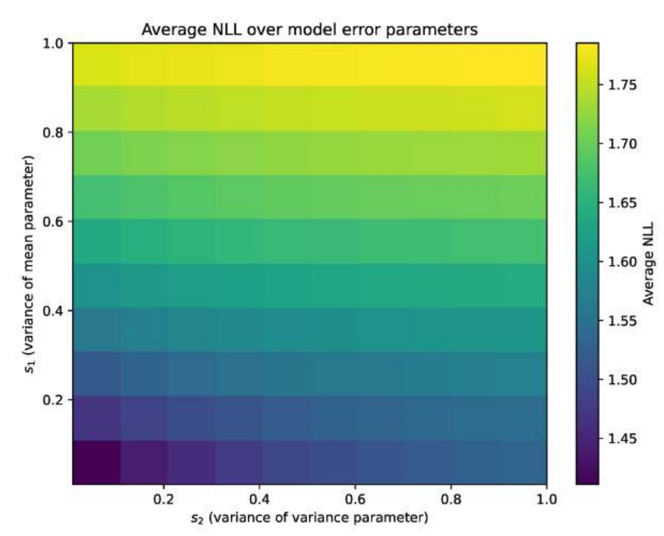


Figure 3: Visualizing NLL as a function of s_1, s_2.

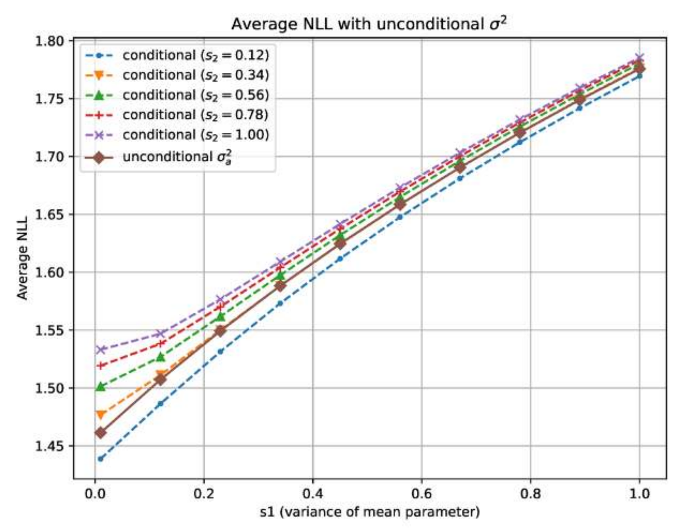


Figure 4: Visualizing NLL with an unconditional estimate for σ^2.

SZTAKI

kdomokos@info.ilab.sztaki.hu