Notation

The ground truth target for input x_i is denoted by t_i. In classification, t_i take the value $+1$ or -1, and in regression a value in R.

Basic concepts

1. Predictive uncertainty refers to the information about the uncertainty in prediction

2. One approach to represent uncertainty in prediction is to treat the variable being predicted as a random variable, and to make predictions in the form of probability distributions, also known as predictive distributions.

3. A simple baseline is to always output the same predictive uncertainty, regardless of the input, based on the properties of the dataset.

Loss functions

4. There are loss functions that takes into account predictive uncertainties and there are loss functions that do not.

5. For classification, the average classification error as shown below does not take into account predictive uncertainties. To see this, consider a single input x_i with $t_1 = +1$, the loss is 0 as long as $p(y_i = +1|x_i) \geq 0.5$. However, if we care about predictive uncertainty, the loss function should be a function of $p(y_i = +1|x_i)$. When the prediction is correct, the loss should be lower when the model is more certain about its accurate prediction.

$$L = \frac{1}{n}\left[\sum_{\{i|t_i=+1\}} 1_{p(y_i=+1|x_i)<0.5} + \sum_{\{i|t_i=-1\}} 1_{p(y_i=+1|x_i)\geq 0.5}\right]$$

6. For classification, the negative log probability (NLP) loss takes into account predictive uncertainties.

$$L = -\frac{1}{n}\left[\sum_{\{i|t_i=+1\}} \log p(y_i = +1|x_i) + \sum_{\{i|t_i=-1\}} \log[1 - p(y_i = +1|x_i)]\right]$$

7. The NLP's minimum is zero and is achieved only when the model predicts correctly with 100% confidence.

8. The NLP penalizes under-confident accurate predictions and over-confident inaccurate predictions. Over-confident inaccurate predictions can be infinitely penalized.

9. For regression, the average normalized mean squared error (nMSE) does not take into account predictive uncertainties.

$$L = \frac{1}{n}\sum_{i=1}^{n} \frac{(t_i - m_i)^2}{vat(t)}, \quad m_i \text{ is the mean of the predictive distribution } p(y_i|x_i)$$

10. The MSE is normalized w.r.t the variance of the true targets so that predicting the empirical mean of the training targets, independently of the test input, leads to a normalized MSE of approximately 1. $var(t)$ can be approximated by the sample variance of the test targets.

11. For regression, the average negative log predictive density (NLPD) takes into account the predictive uncertainties.

$$L = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i = t_i|x_i)$$

12. Consider the case of scalar regression and Gaussian predictive distribution, then given an input x_i with ground truth target t_i, $L_i = \frac{1}{2}\left[\log v_i + \frac{(t_i - m_i)^2}{v_i}\right] + c$ where m_i, v_i are the mean and variance of the predictive distribution and c is a constant. Given m_i, the optimal value for v_i is $(t_i - m_i)^2$.

13. The NLPD penalizes both over $(v_i > (t_i - m_i)^2)$ and under-confident $(v_i < (t_i - m_i)^2)$ prediction.

14. All else equal, the NLPD favors conservative models, that is models that tend to be under-confident (large v_i) rather than over-confident model (small v_i).

15. "Consider the binary classification problem, if n data points are observed, it might seem ambitious to have predictive uncertainties smaller than $1/n$; one has just not observed enough data to be more confident than that. So one obvious technique to avoid infinite penalties in classification would be to replace those predictive probabilities smaller than $1/n$ by $1/n$ and those larger than $1 - 1/n$ by $1 - 1/n$". Why is the reasoning here true?

16. In regression with finite discrete set as output, using the NLPD can be degenerate. One obvious strategy to minimize the NLPD is to distribute the available probability mass equally on tiny intervals around each discrete output value. Since the NLPD only cares about density, the NLPD can be made arbitrarily small by decreasing the width of the interval.