Robustifying Maximum Likelihood Inference

Benoît Frénay BENOIT.FRENAY@UNAMUR.BE
PReCISE, Faculty of Computer Science, Université de Namur, rue Grandgagnage 21, 5000 Namur, Belgium

Michel Verleysen MICHEL.VERLEYSEN@UCLOUVAIN.BE
MLG, ICTEAM, Université catholique de Louvain, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

Keywords: maximum likelihood, robust inference, outliers, pointwise probability reinforcements, regularisation

Probabilistic modelling is popular in machine learning, since it allows a sound formulation of diverse models. For example, ordinary least squares can be seen as performing a linear regression under the hypothesis that the target noise is Gaussian. If prior knowledge indicates that the target noise is distributed according to another distribution, probabilistic modelling can easily use this information to adapt inference. Maximum likelihood is one of the possible ways to infer the parameters of a probabilistic model, but it is well-known to be sensitive to abnormally frequent data, i.e. outliers. Given the probability $p(x_i|\boldsymbol{\theta})$ of the observation x_i under parameters $\boldsymbol{\theta}$, maximising the data log-likelihood

$$\mathcal{L}(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta}) \tag{1}$$

corresponds to minimising the Kullback-Leibler divergence between the empirical distribution of observed data and the chosen parametric distribution. Since outliers have a small probability of occurrence, parameter estimation may be significantly affected and biased towards parameter values which support them.

This abstract presents a work published in Neural Networks (Frénay & Verleysen, 2014), which introduces pointwise probability reinforcements (PPRs) to robustify maximum likelihood inference. Each observation x_i is given a PPR r_i which acts as a reinforcement to the probability $p(x_i|\boldsymbol{\theta})$, resulting in the new objective

$$\mathcal{L}(\boldsymbol{\theta};\mathbf{x},\mathbf{r}) = \sum_{i=1}^{n} \log\left[p(x_i|\boldsymbol{\theta}) + r_i\right] \tag{2}$$

which is called the *reinforced log-likelihood*. For non-outliers, PPRs should remain small or even zero. For outliers, the purpose of the PPR r_i is to compensate between the large empirical probability and the small probability under the parametric model, so as to reduce the impact on the inference of parameters $\boldsymbol{\theta}$.

The reinforced log-likelihood (2) cannot be directly maximised without contraints on the PPRS. Frénay and Verleysen (2014) show that L_1 regularisation results in sparse PPRs, with only a few of them being non-zero, whereas L_2 regularisation provides smoother solutions. Since no closed-form expression exists to optimise the penalised reinforced log-likelihood, Frénay and Verleysen (2014) proposes an iterative algorithm which computes and maximises the lower bound

$$\mathcal{L}_{\mathbf{w}}(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n} w_i \log p(\mathbf{x}_i|\boldsymbol{\theta}) + \text{const.} \tag{3}$$

where $w_i = p\left(\mathbf{x}_i|\boldsymbol{\theta}^{\text{old}}\right)/p\left(\mathbf{x}_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i$ is an instance weight and $\boldsymbol{\theta}^{\text{old}}$ is the current estimate of the model parameters. Close-form expressions exist for PPRs and only instances that are inconsistent with the model get a large PPR r_i and a small weight w_i. The algorithm converges in a few iterations towards parameters $\boldsymbol{\theta}$ which are less biased by outliers, whose weight w_i becomes close to zero as their PPR r_i gets larger.

The main advantage of the PPR methodology is that it can easily be used to robustify any technique that can be obtained as a maximum likelihood solution. Contrarily to other approaches, it is not necessary to modify the probabilistic model itself (what may be complex in some cases). Only the parameter inference step needs to be modified, what is easily achieved. For example, Frénay and Verleysen (2014) show how to robustify linear regression, kernel ridge regression (a.k.a. least squares support vector machines), logistic regression and principal component analysis. Experimental results show that PPRs are effective in these cases.

References

Frénay, B., & Verleysen, M. (2014). Pointwise probability reinforcements for robust statistical inference. *Neural Networks*, *50*, 124 – 141.