Neural Networks 50 (2014) 124–141

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Pointwise probability reinforcements for robust statistical inference

Benoît Frénay *, Michel Verleysen

Machine Learning Group - ICTEAM, Université catholique de Louvain, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 27 March 2013
Received in revised form 14 October 2013
Accepted 14 November 2013

Keywords:
Maximum likelihood
Outliers
Robust inference
Filtering
Cleansing
Probability reinforcements

ABSTRACT

Statistical inference using machine learning techniques may be difficult with small datasets because of abnormally frequent data (AFDs). AFDs are observations that are much more frequent in the training sample that they should be, with respect to their theoretical probability, and include e.g. outliers. Estimates of parameters tend to be biased towards models which support such data. This paper proposes to introduce pointwise probability reinforcements (PPRs): the probability of each observation is reinforced by a PPR and a regularisation allows controlling the amount of reinforcement which compensates for AFDs. The proposed solution is very generic, since it can be used to robustify any statistical inference method which can be formulated as a likelihood maximisation. Experiments show that PPRs can be easily used to tackle regression, classification and projection: models are freed from the influence of outliers. Moreover, outliers can be filtered manually since an abnormality degree is obtained for each observation.

1. Introduction

In statistical inference and machine learning, the goal is often to learn a model from observed data in order to predict a given quantity. In a training sample $\mathbf{x} = (x_1, \ldots, x_n)$, the n observations $x_i \in \mathcal{X}$ are typically assumed to be i.i.d. drawn from the distribution $p(x)$ of the random variable $\mathbf{X}$, whereas the model belongs to a certain parametric family with parameters $\boldsymbol{\theta} \in \Theta$. In particular, many machine learning techniques can be cast as maximum likelihood methods. In this probabilistic framework, learning of the model parameters can be achieved by maximising the data log-likelihood

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

where $p(x_i|\boldsymbol{\theta})$ is the probability of the observation x_i under parameters $\boldsymbol{\theta}$. In order to penalise too complex models which could overfit training data, regularisation methods or Bayesian priors can also be used as a complement.

A common problem when the training sample size n is small is that some data may be much more frequent in the training sample that they should be, with respect to their theoretical probability of occurrence $p(x_i)$. These *abnormally frequent data* (AFDs) may pose a threat to statistical inference when maximum likelihood or similar methods are used. Indeed, maximising the log-likelihood corresponds to minimising the Kullback–Leibler divergence between the empirical distribution of observed data and the considered parametric distribution (Barber, 2012), in the hope that the empirical distribution is close to the real (unknown) distribution. Since the empirical probability of AFDs is much larger than their real probability, the parameter estimation is affected and biased towards parameter values which support the AFDs. For example, AFDs are well known to hurt Gaussian distribution fitting. In this paper, a method is proposed to deal with AFDs by considering that it is better to fit for instance 95% of the data well than to fit 100% of the data incorrectly. Notice that outliers are a subclass of AFDs. Indeed, outliers are observations which should theoretically never appear in a training sample, with respect to the parametric model being used (which reflect hypotheses being made about the data generating process). This includes e.g. data which are very far from the mean in Gaussian distribution fitting or data with incorrect labels in classification. Outliers are known to noticeably affect statistical inference. This paper addresses AFDs in general; experiments focus on the specific subclass of outliers.

In many applications, regularisation or Bayesian methods are used to deal with data which are not correctly described by the model, by penalising overly complex models and avoiding overfitting. However, these methods are only suited for the control of model complexity, not for the control of AFD effects. These two problems should be dealt with different methods. Hence, many approaches have been proposed to perform outlier detection (Barnett & Lewis, 1994; Beckman & Cook, 1983; Daszykowski, Kaczmarek, Heyden, & Walczak, 2007; Hawkins, 1980; Hodge & Austin, 2004) and anomaly detection (Chandola, Banerjee, & Kumar, 2009). It is well-known that many statistical inference methods are quite sensitive to outliers, like e.g. linear regression

* Corresponding author. Tel.: +32 10 47 81 33; fax: +32 10 47 25 98.
E-mail addresses: benoit.frenay@uclouvain.be (B. Frénay), michel.verleysen@uclouvain.be (M. Verleysen).

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.11.012

(Beckman & Cook, 1983; Cook, 1979; Hadi & Simonoff, 1993), logistic regression (Rousseeuw & Christmann, 2003) or principal component analysis (Archambeau, Delannay, & Verleysen, 2006; Daszykowski et al., 2007; Xu & Yuille, 1995). The approach proposed in this paper relies in part on weighted log-likelihood maximisation, which is often used in the literature to reduce the impact of some of the data (Hu & Zidek, 2002). For example, there exist such algorithms for kernel ridge regression (Jiyan, Guan, & Qun, 2011; Liu, Li, Xu, & Shi, 2011; Suykens, De Brabanter, Lukas, & Vandewalle, 2002; Wen, Hao, & Yang, 2010), logistic regression (Rousseeuw & Christmann, 2003) and principal component analysis (Fan, Liu, & Xu, 2011; Huber, 1981). The main problem with these approaches is that the weights are usually obtained through heuristics. Other methods for linear regression include e.g. M-estimators (Huber, 1964), the trimmed likelihood approach (Hadi & Luceo, 1997) and least trimmed squares (Rousseeuw, 1984; Ruppert & Carroll, 1980). One of the main advantages of the method proposed in this paper is that the observation weights are automatically computed.

AFDs have been widely studied in the classification literature, where labelling errors adversely impact the performances of induced classifiers (Zhu & Wu, 2004). For example, the information gain can be used to detect such AFDs (Guyon, Matic, & Vapnik, 1996). Similarly to the proposed approach, it has also been proposed in the classification literature to limit the influence of each observation during inference, in order to prevent the model parameters to be biased by only a few incorrectly labelled instances. However, each method relies on a different way to limit the contribution of observations which is specific to a given model. For example, instances with large dual weights can be identified as mislabelled for support vector machines (Ganapathiraju, Picone, & State, 2000), on-line learning of perceptrons can be robustified by preventing mislabelled instances to trigger updates too frequently (Kowalczyk, Smola, & Williamson, 2001) and boosting algorithms can impose an upper bound on instance weights (Domingo & Watanabe, 2000). It has also been proposed to associate each observation with a misclassification indicator variable which follows a Bernoulli model (Rekaya, Weigel, & Gianola, 2001), what is closer to the contribution of this paper; the indicators can be used to identify mislabelled observations (Hernandez-Lobato, Hernandez-Lobato, & Dupont, 2011; Zhang, Rekaya, & Bertrand, 2006). The approach proposed in this paper has the advantage of being simple to adapt to specific statistical models and not limited to classification problems.

This paper introduces pointwise probability reinforcements (PPRs), which allow the learner to deal with AFDs in a specific way. The probability of each observation is reinforced by a PPR and a regularisation allows one to control the amount of reinforcement which is awarded to compensate for AFDs. The proposed method is very generic, for it can be applied to any statistical inference method which is the solution of a maximum likelihood problem. Moreover, classical regularisation methods can still be used to further control the model complexity. Eventually, abnormality degrees are obtained, which can be e.g. used to manually screen outliers. In the literature, many outlier detection techniques exist; see e.g. Barnett and Lewis (1994), Beckman and Cook (1983), Hawkins (1980) and Hodge and Austin (2004) for a survey. However, the primary goal of the method proposed in this paper is not only to detect the outliers: the aim is rather to make maximum likelihood estimates less sensitive to observations which are abnormally frequent (including outliers) in the training sample, with respect to their theoretical probability. Consequently, common statistical inference methods like linear regression, kernel ridge regression (a.k.a. least squares support vector machines), logistic regression and principal component analysis are shown to be easily robustified using the proposed approach.

This paper is organised as follows. Section 2 introduces PPRs and motivates their formulation. Section 3 proposes a generic algorithm to compute PPRs and to use them during the statistical inference of model parameters. The proposed algorithm is adapted to several supervised and unsupervised problems in Section 4. It is shown that PPRs allow one to efficiently deal with outliers and Section 5 discusses how to choose the amount of reinforcement to use. The resulting methodology is assessed experimentally for kernel ridge regression in Section 6. Eventually, Section 7 concludes the paper.

2. Pointwise probability reinforcements: definition and concepts

As explained in Section 1, the problem with AFDs is that their empirical probability is much larger than their actual probability. As a consequence, the parameters of models inferred from data with AFDs are biased towards values which overestimate the probability of AFDs. For small training samples, this can have an important impact on the resulting model. For example, in linear regression, outliers can significantly bias the slope and the intercept of an estimated model.

In this paper, it is proposed to deal with AFDs by introducing *pointwise probability reinforcements* (PPRs) $r_i \in \Re^+$. The log-likelihood becomes

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

where 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 a *reinforced probability*. The above log-likelihood is called here the *reinforced log-likelihood*. The PPRs should remain small (or even zero), except for AFDs for which they will compensate for the difference between their large empirical probability and their small probability under a model with parameters $\boldsymbol{\theta}$. The spirit of the proposed method is similar to the one of M-estimators (Huber, 1964) and related approaches (Chen & Jain, 1994; Chuang, Su, & Hsiao, 2000; Liano, 1996). In regression, instead of minimising the sum of the squared residuals, the M-estimator approach consists in minimising another function of the residuals which is less sensitive to extreme residuals. Similarly, PPRs allow one to make maximum likelihood less sensitive to extremely small probabilities. However, there exist many different M-estimators and it is not necessarily easy to choose among them. Moreover, their use is limited to regression. On the contrary, PPRs can be used to robustify maximum likelihood methods for e.g. regression, classification or projection, as shown in Section 4. Moreover, Section 3 shows that PPRs can be easily controlled using regularisation, for example by introducing a notion of sparsity.

Eq. (2) can be motivated by considering methods which are used in the literature to deal with outliers. In classification, data consists of pairs $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ where x_i is a vector of observed feature values and y_i is the observed label. Label noise occurs when a few data have incorrect labels (e.g. false positives in medical diagnosis). In such a case, Lawrence and Schölkopf (2001) introduce a labelling error probability π_e which can be used to write

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \pi_e; \mathbf{x}, \mathbf{y}) &= \sum_{i=1}^{n} \log\left[(1-\pi_e)\, p(y_i|x_i, \boldsymbol{\theta}) \right. \\ &\quad \left. + \pi_e\,(1 - p(y_i|x_i, \boldsymbol{\theta}))\right] \\ &= \sum_{i=1}^{n} \log\left[p(y_i|x_i, \boldsymbol{\theta}) + \frac{\pi_e}{1-2\pi_e}\right] \\ &\quad + n \log\left[1 - 2\pi_e\right]. \end{aligned} \tag{3}$$