Learning with noisy labels

Johan Jonasson and Rebecka Jörnsten
Mathematical Sciences
University of Gothenburg/Chalmers University of Technology

Overview

In the paradigms of supervised, semi-supervised or active learning with machine learning algorithms in general and deep neural nets in particular, the default models assume that the given annotations of training data are correct.

However, since well-annotated data sets can be expensive and time consuming to collect, some recent research has focused on using larger but noisy sets of training data. Such data sets are much cheaper to collect, e.g. via crowdsourcing. The working assumption is that collecting larger amounts of data that are labeled with a decent accuracy can compensate for the noise (inaccurate labels).

In the literature, there are three main approaches to dealing with noisy labels; (i) by incorporating the noise directly into the loss function of the training model [7, 6], either via a surrogate loss function that augments the standard classification loss with noise rate parameters or with a data reconstruction term (the unsupervised learning problem), (ii) methods from robust statistics, or (iii) using influence functions [4] to identify what labels that are most responsible for misclassifications of test data and, if possible, review those labels, correct them and then retrain the model.

There is a lack of statistical understanding of models for dealing with noisy data and attempts so far seem to have been mostly heuristic. In addition, there are also many ways to model the noise yet to be explored. The simplest noise is random classification noise, whereas the training problem is of course much more difficult if there is a structured noise in the labeling (i.e. a tendency to mislabel in a certain direction).

Project description

In short, this project will open the lid on the black box of deep neural nets. In particular, we set out to understand why it is that DNN:s are sometimes overwhelmingly successful and sometimes far from impressive and why they can be fooled by apparently negligible noise in terms of misclassification of training data and/or distortions of test examples. This will lead to new, more stable, DNN architectures that allow for training on noisily annotated training data. Tools from the field of Noisy sensitivity/stability of Boolean functions will be used to study DNN:s and, conversely, questions arising from the study of DNN:s will open new exciting avenues of investigation for the former subject.

A further and broader aim is to develop a better general understanding for the statistical principles that underpin learning in a noisy label setting. There are multiple goals with such research; 1) to gain deeper insight into the behaviour of even simple models under various noise distributions and through this propose improved models, and 2) investigate active learning settings where the models are trained to also re-label observations identified as likely mislabeled. This latter task is notoriously difficult; if we ask an oracle for a correct label of a training point which has a large influence on a classification algorithm, this will introduce a bias which could have large impact on the analysis.

On a more practical level, we aim to derive a scalable (stochastic) gradient method for robust estimation in deep learning. To incorporate large amounts of unannotated data we will investigate multi-task optimization pairing classification of labeled data with reconstruction error minimizing for all data. A related point of interest is how to do fast and accurate training on large sets of data. This is the classical problem of finding a global optimum in a fitness landscape of many potential local optima, a problem usually only dealt with heuristically by practitioners. However, using subsampling schemes, somewhat akin to simulated annealing, seems to be a way of achieving provably faster convergence.

Noise stability and neural networks

The works on handling noisy labels mentioned above, deserve deeper analysis and further development. The ideas are clearly related to the well-studied concepts of influence of individual random variables on n-variable Boolean functions, noise sensitivity and noise stability [1]. What are the precise relations and what possible avenues of cross-fertilization are there? There are several interesting lines of investigation to take. A Boolean function, like e.g. a class label, is said to be noise stable if there is some small $\epsilon > 0$, such that, asymptotically, randomly changing an ϵ-fraction of uniformly random input data with high probability does not change the value of the function. A function that is not stable for any ϵ is said to be noise sensitive. The influence of a single input is the probability that changing that input would change the value of the function.

Clearly it would be desirable to design a neural net classifier so that its output on a typical input is noise stable with respect to changing a small fraction of the labels or the features of the training data. Is this true for a typical neural net? Results in the literature suggests that this may/may not be the case, depending on the noise distribution, the type of neural net, the size of the correctly labeled data and the amount of corrupted data. While some results indicate that classification performance on correct labels can be noise stable (e.g. [9]), these results are limited in their use unless one can also confidently identify the anomalies. It is of interest to prove when a classifier is noise stable and when it is not, and to redesign the classifier to remove the sensitivity without significantly sacrificing performance.

Identifying the input data with the highest influence is valuable for detecting weaknesses in the network design and detecting misclassified training data (e.g. [4]). Classification performance can be improved by appropriate weighting of training data based on their importance or influence [5]. How sensitive are neural networks to such active learning (error correcting) strategies?

When annotations are obtained via crowdsourcing, this may involve thousands of labelers of varying skills. For each observation or annotation task, one or a few labelers provide independent labels and the set of labelers are in no way restricted or encouraged to overlap between tasks. Traditionally, the presence of multiple labels were dealt with by reducing them to a single label via e.g. majority voting. Statistical and Machine Learning research have since lead to the development of EM-based or Bayesian methods for jointly estimating the classification model and labeler skills. Recent research into deep neural nets for modeling crowdsourcing data augment these classifiers similarly [2, 8]. In this setting, the identification of influential observations and an analysis of noise sensitivity (e.g. distribution of skills) is clearly needed.

Techniques from the analysis of Boolean functions have also been used as a tool of model optimization. Assuming that a model depends on a large number of discrete hyperparameters, such as e.g. the number of layers of a neural net, it becomes of utmost importance to identify the variables of the highest influence on classification, since standard cross-validation for model choice quickly becomes infeasible when the number of such parameters grows [3].

Conversely, inspiration from problems in deep learning will drive the study of new concepts of stability/sensitivity of Boolean functions. Specifically, we will investigate the following questions.

- How many input variables would an adversary need to in order to change the value of a given Boolean function if the variables may be specified after the experiment? The same question can also be investigated for the case when adversary has to specify in advance what input variables to change.
- The analogous questions, but for a fixed correct value of the input data (such as a correct labeling of training data). This question is of course also interesting for usual non-adversarial noise.
- The concepts of influences and noise stability/sensitivity are most well studied for Boolean input variables. Applications in deep learning call out for a better understanding of more general settings.

Relation to AI

Machine learning in general and deep neural nets in particular are at the core of artificial intelligence. Applications are abundant, with image processing, image compression, natural language processing, stock market prediction, chess playing, to name a few prolific examples.

For supervised or semi-supervised learning with deep neural nets, one needs large sets of correctly annotated training examples. It is well known that these are hard to come by, whereas very large sets of data with noisy labels, i.e. data annotated with less accuracy, are easily available. In light of that, it becomes obvious that reliable learning with noisy labels is of utmost importance. Despite this, it is only recently that serious attempts have been made at this. One reason is the "black-box" phenomenon, i.e. that it is very challenging to understand exactly how a neural network achieves its predictive power. Therefore, even though it is known from experience that one can sometimes completely fool a neural net classifier by e.g. making tiny changes to a training example or misclassifying a (carefully chosen) example, it is hard to tell how to modify the net architecture to make it more noise stable.

The main thrust of this project is to take on this challenge. The starting point will be to investigate when the prediction of a typical test case given a certain training set and a certain architecture will be stable under small perturbations of the training data. Understanding this even for a simple ConvNet with only a few layers will provide valuable clues to why neural nets are spectacularly successful for some tasks and fail for others.

Project team

In addition to the methodological research described above, we will work with academic partners in probability theory (Jeff Steif), machine learning (Devdatt Dubhashi) and industry (Daniel Langkilde, Annotell) on an application of image classification and object identification.

The student will benefit from this interdisciplinary setup. Jonasson and Jörnsten currently co-supervise 2 master student teams at Annotell working on the crowdsourcing annotation problem in the context of image classification, with special attention to the allocation of labelers to different annotation tasks (a related research question to the above program).

In addition to this dynamic environment, the student's membership in the planned research school will also serve to provide a broader context of AI outside the planned applications specific to our team.

Supervisor: Professor Johan Jonasson. Jonasson has worked on various topics discrete probability, such as statistical mechanics, random graphs, mixing times of Markov chains and analysis of Boolean functions. In recent years, he has changed focus to probabilistic perspectives of models in natural language processing.
Co-supervisor: Professor Rebecka Jörnsten. Jörnsten works on machine learning and statistical modelling problems in the context of high-dimensional biological data. Recently, she has worked on deep learning methods for data integration and subtype detection in systems biology.
Research partners: Professor Devdatt Dubhashi, Professor Jeff Steif and MSc Daniel Langkilde. Dubhashi is a prolific researcher with extensive expertise in machine learning and AI. Steif is a world leading expert on noise sensitivity/stability and a member of the probability group at the Dept. of Mathematical Sciences. Langkilde is a cofounder of Annotell, a company that specializes in providing high quality annotated data for supervised machine learning. Master students teams working on related problems with co-supervisors.

References

[1] Christophe Garban and Jeffrey E Steif. *Noise sensitivity of Boolean functions and percolation*, volume 5. Cambridge University Press, 2014.

[2] Melody Y Guan, Varun Gulshan, Andrew M Dai, and Geoffrey E Hinton. Who said what: Modeling individual labelers improves classification. *arXiv preprint arXiv:1703.08774*, 2017.

[3] Elad Hazan, Adam Klivans, and Yang Yuan. Hyperparameter optimization: A spectral approach. *arXiv preprint arXiv:1706.00764*, 2017.

[4] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[5] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

[6] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1196–1204. Curran Associates, Inc., 2013.

[7] Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLR 2015*, 2015.

[8] Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. *arXiv preprint arXiv:1709.01779*, 2017.

[9] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.