Online Mixed Model

CSC2506 - Project Proposal

Chris Cremer
ccremer@cs.toronto.edu

February 2016

1 Overview

Most learning algorithms assume that the training data are independent and identically distributed. However, this assumption is commonly violated in many real world problems where sub-groups of samples exhibit a high degree of correlation amongst both features and labels. Often there exists sampling biases that lead to some sub-groups being over represented in the training data. The types of models that account for these confounders are computationally expensive and thus limit the size of the training set. I propose to address the sampling bias problem in an online fashion. I introduce a model that accounts for the biased sampling by reducing the effect that similar samples have on the training of the model. The model can be trained via gradient descent and thus can accommodate very large datasets. I intend to demonstrate its effectiveness on protein structure prediction.

2 Sampling Bias

In most real world problems, samples are rarely collected without any bias, and thus are not independent and identically distributed (i.i.d.). Most learning algorithms assume that samples are i.i.d., consequently the biases in the sample space cause the model to learn a misrepresentation of the true distribution.

One option to address the problem of unbalanced sample space sampling could be to measure the similarity of each sample with all the others and then give higher weights to samples that are more dissimilar to the rest, in essence, balancing the sampling over the sample space. This method's runtime would scale quadratically with the size of the training set and would be impractical for very large datasets.

3 Previous Work

Sampling bias can be addressed by mixed models. Mixed models explicitly model the fixed and random effects of the data. The fixed effects are the effects that are dependent on the independent variables. In contrast, the random effects are the effects attributed to factors other than the independent variables.

Generalized linear mixed models (GLMMs) are computationally expensive relative to simpler models. In particular, the run time scales as the cube of the number of training samples. Work has been made to improve the efficiency of these models [1],[2].

Online learning is a learning paradigm in which a model is trained using data that becomes available in a sequential manner as opposed to batch learning techniques which trains a model on the entire training data set at once. Thus the benefits of online learning arise in large scale data settings where it is infeasible to train using the entire dataset at once.

Thus it would beneficial to develop a model that takes into account the fixed and random effects as well as being able to be trained online.

4 Proposed Model

The proposed model aims to account for the fixed and random effects using a foreground (F) and a background (B) model, respectively. The predictive distribution for the kth output $\hat{y}_k$ given $\hat{x}_k$ and all the previous samples D is defined as,

$$P(\hat{y}_k|\hat{x}_k, D, \pi) = \pi F(\hat{y}_k|\hat{x}_k, \theta) + (1-\pi)B(\hat{y}_k|\hat{x}_k, D) \tag{1}$$

where,

$$B(\hat{y}_k|\hat{x}_k, D) = \frac{\sum_{i=1}^{k-1} K(\hat{x}_k, x_i)y_i}{\sum_{i=1}^{k-1} K(\hat{x}_k, x_i)}. \tag{2}$$

The foreground model $F(\hat{y}_k|\hat{x}_k, \theta)$ can be any probabilistic prediction model. The variable θ represents the parameters of the foreground model. The variable D is the data we've seen up to sample k. The parameter π is the ratio of the prediction attributed to the foreground model versus the background model. The function $K(a, b)$ measures the similarity between samples a and b.
Let's examine the derivative of the log probability of sample k with respect to the parameters of the foreground model,

$$\frac{\partial ln(P(\hat{y}_k|\hat{x}_k, D))}{\partial \theta} = \frac{\pi}{P(\hat{y}_k|\hat{x}_k, D)}\Big(\frac{\partial F(\hat{y}_k|\hat{x}_k, \theta)}{\partial \theta}\Big) \tag{3}$$

$$= \frac{\pi}{\pi F(\hat{y}_k|\hat{x}_k, \theta) + (1-\pi)B(\hat{y}_k|\hat{x}_k, D)}\Big(\frac{\partial F(\hat{y}_k|\hat{x}_k, \theta)}{\partial \theta}\Big) \tag{4}$$

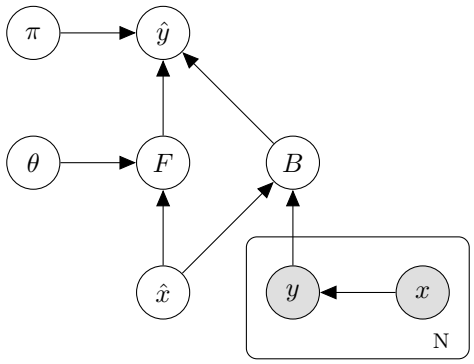

Figure 1: Graphical model of the proposed model

We see that the parameters of the foreground model change in proportion to the performance of both the foreground and background models. Consequently, if the background model is able to predict the sample, then there will be little change in the foreground model.

The current formulation for the background model iterates over all previous samples, which would be very computationally expensive. Accordingly, instead of iterating over all samples, we will only look at a fixed number of samples. These fixed samples will be iteratively updated based on the dissimilarity of the incoming samples compared to the current fixed samples.

This model addresses the sampling bias problem because if there exist samples that are very correlated, such that they don't provide any new information, then they will have little effect on the foreground model. Thus the foreground model will be trained on a more balanced distribution of the sample space. This model can be trained in an online manner via gradient descent.

5 Protein Structure Prediction

A motivating application of the proposed model is protein structure prediction from evolutionary couplings. For biological sequence data there are strong sampling biases due to phylogenetic relations between species, due to the sequencing of different strains of the same species, and due to a non-random selection of sequenced species. The sampling is therefore clustered in sequence space, thereby introducing spurious non-functional correlations, whereas other viable parts of sequence space are statistically underrepresented [3].

References

[1] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "Fast linear mixed models for genome-wide association studies," *Nature methods*, vol. 8, no. 10, pp. 833–835, 2011.

[2] M. Dundar, B. Krishnapuram, J. Bi, and R. B. Rao, "Learning classifiers when the training data is not iid.," in *IJCAI*, pp. 756–761, 2007.

[3] C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani, "Fast and accurate multivariate gaussian modeling of protein families: predicting residue contacts and protein-interaction partners," *PloS one*, vol. 9, no. 3, 2014.