Log-linear models structure comparison

Jan Strappa and Facundo Bromberg

July 31, 2019

Abstract

Log-linear models have been largely used for representing the structure of probabilistic models containing context-specific interactions among variables. Automatic structure learning of these models from data is a popular Machine Learning problem, for which the comparison of models' structures is an important aid for the evaluation of algorithms. Nowadays, structure learning, weight learning and inference with this class of probabilistic models have a wide variety of applications in the literature of Machine Learning, such as computer vision and image analysis, language processing, computational biology and biomedicine, among many others. Existent measures such as Kullback-Leibler divergence do not allow the direct comparison of structural differences; instead, they compare the full distribution (structure and parameters of the models), which introduces several limitations. We propose a distance measure for comparing the dependence structures of log-linear models, similar to the Hamming distance commonly used for models represented by undirected graphs. Our method is proven to be a metric, and can be efficiently computed in terms of the number of variables of the domain. It can be used for the comparison of structure learning techniques and for obtaining insights from log-linear models.

1 Introduction

A *log-linear model* over a discrete domain $X = \{X_1, ..., X_N\}$ is defined as a set of feature functions $F = \{f_K(X_K)\}$, each one defining a numerical value for each assignment x_K to some subset $X_K \subset X$. Given the set F, the parameters of the log-linear model are the weights $\theta = \{\theta_K : f_K \in F\}$. The overall distribution is defined as $P(x) = \frac{1}{Z(\theta)} exp(\sum_{f_K \in F} \theta_K f_K(x_K))$, where $Z(\theta)$ is the partition function that ensures that the distribution is normalized (i.e., all entries sum to 1) [9, 8, 3, 5].

Unlike undirected graphical models, log-linear models are a more accurate representation when the distribution underlying the data contains context-specific independences (i.e., conditional independences that hold only for some values of the conditioning variables) [1, 6],

which makes them better suited for inference tasks in those cases. In addition, there are a number of graph-based representations that allow a qualitative interpretation of these models [2, 7].

Nevertheless, a straight-forward comparison of their independence structure would entail a superexponential number of comparisons. For this reason we have designed a method that takes advantage of the compactness of the feature representation of log-linear models in order to allow for an efficient computation of discrepancies between two structures, while also providing theoretical guarantees.

2 Approach

Our method consists in computing a confusion matrix, comprising the well-known measures of *true positives* (TP), *true negatives* (TN), *false positives* (FP), and *false negatives* (FN), obtained from the dependence structure of the models. To illustrate, for comparing undirected graphical models, the Hamming distance of the edges between two graphs is used as an ad-hoc measure, where the distance is equivalent to the discrepancies between a subset of independence assertions entailed by the structures (given by the *pairwise Markov property*). In our approach, the comparison is made between a subset of assertions of context-specific independence entailed by the dependence structure, and it is computed directly based on the features of the log-linear representation. Additionally, we provide proof that the proposed measure is a metric.

2.1 Example: Discriminating False Positives From False Negatives

We illustrate how our measure differs from KL-divergence [4] (the most commonly used measure for comparison of distributions) by generating synthetic graphs with different numbers of false positives and false negatives, and visualising the results. We use R for preprocessing of results and visualisation. We show an example of our results (for a synthetic model of 6 variables) in Figure 1.

References

[1] S. Della Pietra, V. Della Pietra, and L. J. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):390–393, 1997.

[2] S. Højsgaard. Statistical inference in context specific interaction models for contingency tables. *Scandinavian journal of statistics*, 31(1):143–158, 2004.

[3] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

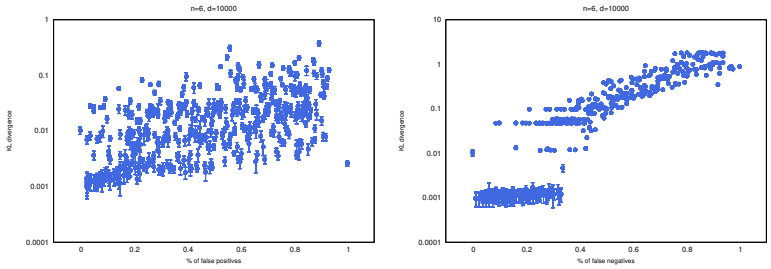

Figure 1: Comparison of errors for the proposed log-linear measure (x-axis) vs KL-divergence (y-axis) for synthetic datasets (6 variables, 10000 observations). Each point in the graph is one structure. X-axis shows % of FPs (left) and % of FNs (right) as computed by our measure.

[4] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[5] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[6] A. McCallum. Efficiently inducing feeatures of conditional random fields. *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2003.

[7] H. Nyman, J. Pensar, T. Koski, J. Corander, et al. Stratified graphical models-context-specific independence in graphical models. *Bayesian Analysis*, 9(4):883–908, 2014.

[8] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, 2nd edition, 1988.

[9] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning Series. MIT Press, 2nd edition, January 2000.