Adaptation of a mixture of multivariate Bernoulli distributions

Ankur U. Kamthe, Miguel Á. Carreira-Perpiñán and Alberto E. Cerpa
Electrical Engineering and Computer Science
University of California, Merced, U.S.A.

Introduction

- Mixture of multivariate Bernoulli distribution (MMB) is a widely-used statistical model for high-dimensional binary data.
- Recently, the MMB has been used to model packet loss patterns in wireless links for use in network simulators. The binary sequences of receptions and losses is split into windows of W-bits e.g. 00111000, 11110000, ...
- Learning an accurate model for a new link requires extensive data collection, a costly process in practice.
- Contribution: A new algorithm that adapts a preexisting MMB trained with extensive data (reference MMB) to a new link from which very limited data is available.

Mixture of multivariate Bernoulli (MMB) distributions

- Given a data vector $\mathbf{x} \in \{0,1\}^W$ with W binary variables, the MMB density with M components is

$$p(\mathbf{x}) = \sum_{m=1}^{M} \pi_m p(\mathbf{x}|m) \quad \text{where} \quad p(\mathbf{x}|m) = \prod_{w=1}^{W} p_{mw}^{x_w}(1-p_{mw})^{1-x_w}$$

$$\text{and} \quad \pi_m > 0; \quad \sum_{m=1}^{M} \pi_m = 1; \quad \mathbf{p}_m \in [0,1]^W$$

- Given a training set, an MMB is usually trained with an EM algorithm.
- But, training with little data leads to estimates that overtrain and generalize poorly to future data.
- For example, some dimensions in the data may consist mostly (or only) of 0s or 1s. The corresponding p_{mw} value will clamp to (close to) 0 or 1.

Adapting the MMB

- Rather than training an MMB from scratch, we use the little data we have for our target link to adapt a pre-existing MMB that was trained with lots of data for a different link (reference MMB).
- Our MMB adaptation algorithm is based on the idea of tying the MMB parameters together through a transformation of the reference parameters.
- Given a *reference* model and an *adaptation dataset*, we want to learn a new MMB model, with parameters $\{\widetilde{\pi}_m, \widetilde{\mathbf{p}}_m\}_{m=1}^M$, for the *target* distribution.

$$\widetilde{p}_{mw} = \sigma(p_{mw}; a_m, b_m) = \frac{1}{1+e^{-(a_m p_{mw} + b_m)}} \quad \text{where} \quad w = 1, \ldots, W.$$

- The transformation has few parameters $\{a_m, b_m\}_{m=1}^M$, that can be learned from the small adaptation dataset, yet their effect propagates to all the MMB parameters.
- The transformation is nonlinear because the p_{mw} values must be in $[0,1]$. With a linear transformation, the reference values p_{mw} close to either 0 or 1 would saturate and prevent the remaining values from adapting.
- For the mixing proportions, there is only one per component, we consider them as free during adaptation.
- Thus, our algorithm needs to maximize the likelihood of the adaptation data over a total of $3M-1$ free parameters (mixing proportions and sigmoid parameters), which with our high-dimensional data is far less than $(W+1)M-1$ parameters for retraining.

A generalized EM algorithm for adaptation

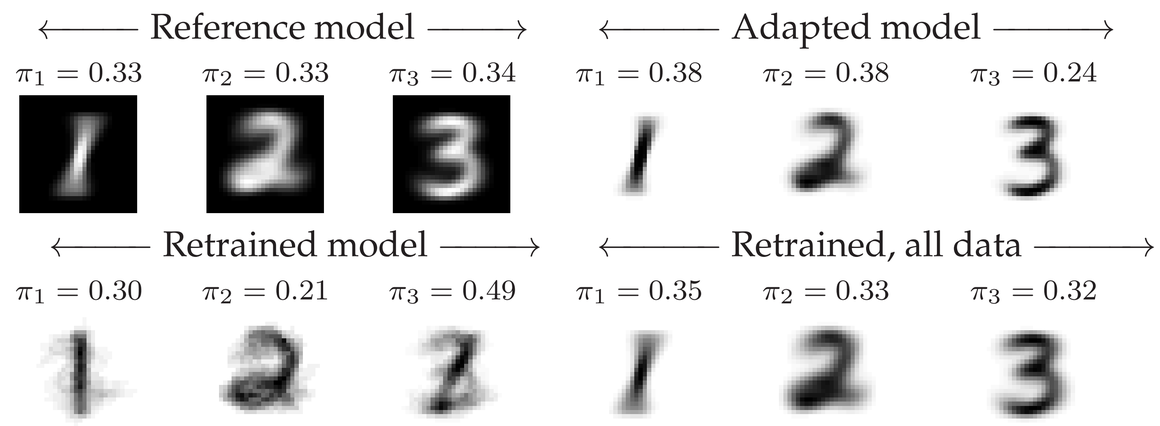
- Our objective function is the log-likelihood of the adaptation data given the constrained MMB model with $3M-1$ free parameters:

$$L\left(\{\widetilde{\pi}_m, a_m, b_m\}_{m=1}^M\right) = \sum_{n=1}^{N} \log \sum_{m=1}^{M} \widetilde{\pi}_m p(\mathbf{x}_n; a_m, b_m)$$

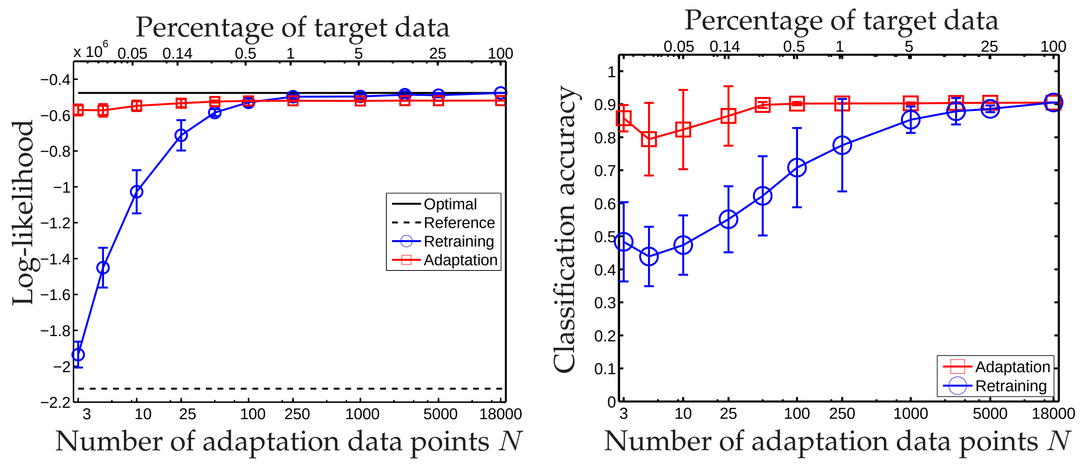
  where $p(\mathbf{x}_n; a_m, b_m)$ is a multivariate Bernoulli with $\widetilde{\mathbf{p}}_m = \sigma(\mathbf{p}_m; a_m, b_m)$.
- Unlike retraining, using a nonlinear transformation makes the M step not solvable in closed form for $\{a_m, b_m\}$. We solve it iteratively using BFGS.
- BFGS increases but (if we exit BFGS early) need not maximize the likelihood within the M step $\Longrightarrow$ generalized EM algorithm (convergence assured by GEM theorem).
- We provide a generalized EM algorithm to maximize the objective function (see details in paper). Computational cost: $\mathcal{O}(NMW)$.

Experiments I - MNIST handwritten digits

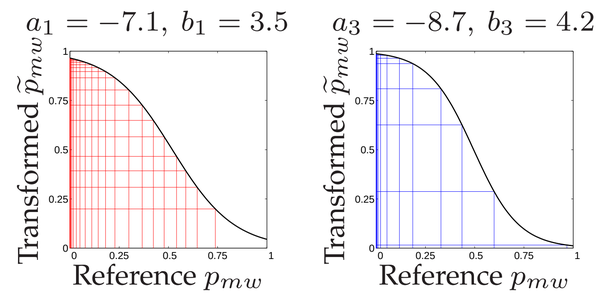
Sample training vectors $\mathbf{x}_n \in \{0,1\}^{28\times 28}$ in the reference (top row) and target (bottom row) datasets. The latter has inverted each pixel.



MMB parameters for the reference model, adaptation ($N = 100$ adaptation points), retraining ($N = 100$) and retraining ($N = 18000$).

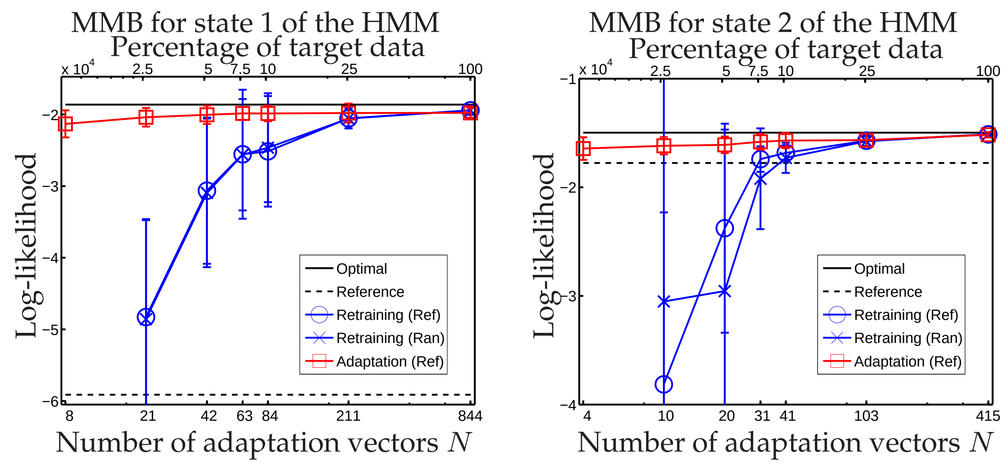


Log-likelihood (left) of retraining and adaptation algorithms, and classification accuracy (right) on test sets.



Estimated sigmoid for each MMB component. They invert the input.

Experiments II - Wireless link datasets



Results (on test sets) of retraining and adaptation algorithms, as a function of the adaptation set size N. Reference model trained with N=1200 (=40minutes) of data.

Summary

- For wireless links, adaptation achieves models with traces of about a minute that are as good as retraining MMBs with traces of hours, greatly simplifying the task of building realistic network simulators.
- Our algorithm applies to adapting MMBs in other settings, and future work will address other, more suitable ways of sharing parameters.
- Matlab code at http://eecs.ucmerced.edu.

Work partially supported by the National Science Foundation under grants IIS–0711186 and CNS–0923586, the California Institute for Energy and Environment under grant MUC–09– 03, and the Center for Information Technology Research in the Interest of Society under grant 442130–19900.

IJCAI 11 TWENTY-SECOND INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE JULY 16-22, 2011 CCIB BARCELONA, CATALONIA (SPAIN) WWW.IJCAI-11.ORG