Stochastic Discriminative EM (sdEM)

Andrés R. Masegosa[†‡]

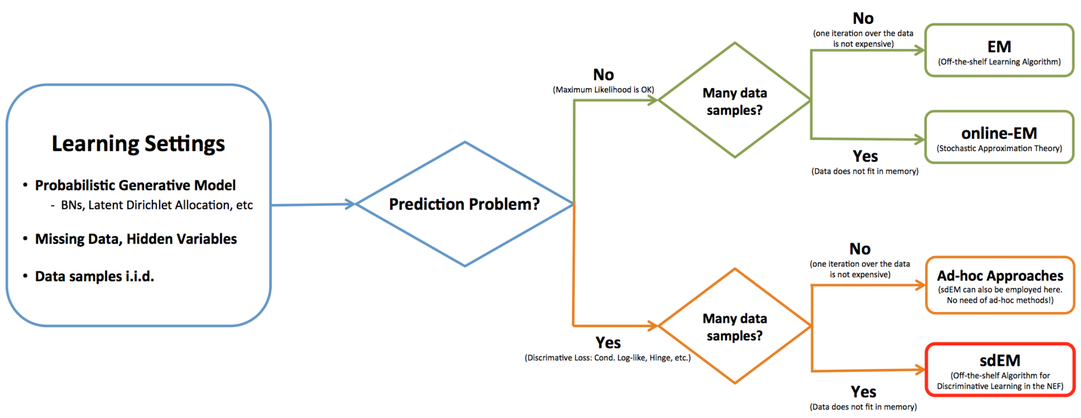
[†] Dept. of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
andres@idi.ntnu.no

[‡] Dept. of Computer Science and A. I.
University of Granada
Granada, Spain
andrew@decsai.ugr.es

Abstract

Stochastic discriminative EM (sdEM) is an online-EM-type algorithm for discriminative training of probabilistic generative models belonging to the natural exponential family. In this work, we introduce and justify this algorithm as a stochastic *natural gradient* descent method, i.e. a method which accounts for the information geometry in the parameter space of the statistical model. We show how this learning algorithm can be used to train probabilistic generative models by minimizing different discriminative loss functions, such as the negative conditional log-likelihood and the Hinge loss. The resulting models trained by sdEM are always generative (i.e. they define a joint probability distribution) and, in consequence, allows to deal with missing data and latent variables in a principled way either when being learned or when making predictions. The performance of this method is illustrated by several text classification problems for which a multinomial naive Bayes and a latent Dirichlet allocation based classifier are learned using different discriminative loss functions, and in a online manner.

1 Introduction



2 Models & Assumptions

Y denotes the random variable (continuous, discrete or vector-value random variable) to be predicted, X denotes the observable predictive variables and Z the non-observable or hidden variables.

- The generative data model belongs to a **natural exponential family**

$$p(y, z, x|\theta) \propto exp(\langle s(y, z, x), \theta\rangle - A_l(\theta))$$

where $\theta \in \Theta$ is the natural parameter, $s(y, z, x) \in \mathcal{S}$ is the vector of sufficient statistics and A_l is the log partition function.

- A **conjugate prior distribution**

$$p(\theta|\alpha) \propto exp(\langle s(\theta), \alpha\rangle - A_g(\alpha))$$

where the sufficient statistics are $s(\theta) = (\theta, -A_l(\theta))$ and the hyperparameter α has two components $(\bar{\alpha}, \nu)$. ν is a positive scalar and $\bar{\alpha}$ is a vector.

- Transformation from **the expectation parameter** $\mu = E[s(y, z, x)|\theta]$ to the **natural parameter** θ expressed as is **available in closed form.** Or, equivalently, the maximum likelihood parameters associated to a given sufficient statistics can be computed in closed form.

3 The sdEM Algorithm

Learning Settings

A data set D with n observations $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ and *discriminative loss function* $\ell(y_i, x_i, \theta)$. Our learning problem consists in minimizing the following objective function:

$$L(\theta) = \sum_{i=1}^{n} \ell(y_i, x_i, \theta) - \ln p(\theta|\alpha) = E\left[\ell(y, x, \theta)|\pi\right] - \frac{1}{n}\ln p(\theta|\alpha) = E\left[\bar{\ell}(y, x, \theta)|\pi\right]$$

where π is the empirical distribution of the data D.

The stochastic updating equation of sdEM

sdEM can be interpreted as a stochastic gradient descent algorithm iterating over the *expectation parameters* and guided by the *natural gradient* in its Riemannian space

$$\mu_{t+1} = \mu_t - \rho_t I(\mu_t)^{-1}\frac{\partial\bar{\ell}(y_t, x_t, \theta(\mu_t))}{\partial\mu}$$

Theorem 1. *In a natural exponential family, the natural gradient of a loss function with respect to the expectation parameters equals the gradient of the loss function with respect to the natural parameters,*

$$I(\mu)^{-1}\frac{\partial\bar{\ell}(y, x, \theta(\mu))}{\partial\mu} = \frac{\partial\bar{\ell}(y, x, \theta)}{\partial\theta}$$

Pseudo-Code Description of sdEM

Require: D is randomly shuffled.

1: $\mu_0 = \bar{\alpha}$; (initialize according to the prior)
2: $\theta_0 = \theta(\mu_0)$;
3: $t = 0$;
4: **repeat**
5: **for** $i = 1, \ldots, n$ **do**
6: **E-Step**: $\mu_{t+1} = \mu_t - \frac{1}{(1+\lambda t)}\frac{\partial\bar{\ell}(y_i, x_i, \theta_t)}{\partial\theta}$;
7: Check-Step: $\mu_{t+1} = Check(\mu_{t+1}, \mathcal{S})$;
8: **M-Step**: $\theta_{t+1} = \theta(\mu_{t+1})$;
9: $t = t + 1$;
10: **end for**
11: **until** convergence
12: **return** $\theta(\mu_t)$;

Recent results in information geometry [25] show that sdEM could also be interpreted as a mirror descent algorithm or proximal gradient method with a Bregman divergence as a proximitiy measure.

4 Discriminative Loss Functions

Negative conditional log-likelihood (NCLL)

This loss function (which is valid for classification, regression, multi-label, etc.) is defined as follows:

$$\ell_{NCLL}(y_t, x_t, \theta) = -\ln p(y_t|x_t, \theta) = -\ln\int p(y_t, z, x_t|\theta)dz + \ln\int p(y, z, x_t|\theta)dydz$$

And its gradient is computed as

$$\frac{\partial\ell_{NCLL}(y_t, x_t, \theta)}{\partial\theta} = -E_z[s(y_t, z, x_t)|\theta] + E_{yz}[s(y, z, x_t)|\theta]$$

The Hinge loss (Hinge) for probabilistic generative models

We build on LeCun et al.'s [21] ideas about energy-based learning for prediction problems. We define the hinge loss (which is only valid for classification) as follows

$$\ell_{Hinge}(y_t, x_t, \theta) = \max(0, 1 - \ln\frac{p(y_t, x_t|\theta)}{p(\bar{y}_t, x_t|\theta)}) \quad (1)$$

where $\bar{y}_t$ denotes here too the most offending incorrect answer, $\bar{y}_t = arg\max_{y\neq y_t} p(y, x_t|\theta)$.

The gradient of this loss function can be simply computed as follows

$$\frac{\partial\ell_{Hinge}(y_t, x_t, \theta)}{\partial\theta} = \begin{cases} 0 & \textbf{if } \ln\frac{p(y_t, x_t|\theta)}{p(\bar{y}_t, x_t|\theta)} > 1 \\ -E_z[s(y_t, z, x_t)|\theta] + E_z[s(\bar{y}_t, z, x_t)|\theta] & \textbf{otherwise} \end{cases}$$

sdEM updating equations for partially observed data

NLL $$\mu_{t+1} = (1 - \rho_t(1 + \frac{\nu}{n}))\mu_t + \rho_t\left(E_z[s(y_t, z, x_t)|\theta(\mu_t)] + \frac{1}{n}\bar{\alpha}\right)$$

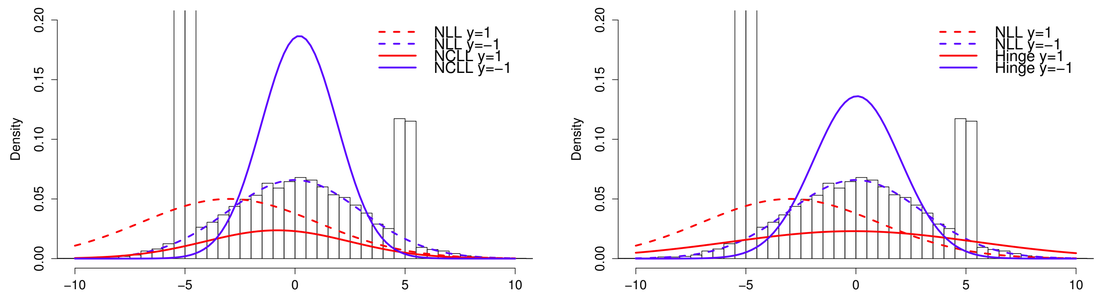
NCLL $$\mu_{t+1} = (1 - \rho_t\frac{\nu}{n})\mu_t + \rho_t\left(E_z[s(y_t, z, x_t)|\theta(\mu_t)] - E_{yz}[s(y, z, x_t)|\theta(\mu_t)] + \frac{1}{n}\bar{\alpha}\right)$$

Hinge $$\mu_{t+1} = (1 - \rho_t\frac{\nu}{n})\mu_t + \rho_t\begin{cases} \frac{1}{n}\bar{\alpha} & \textbf{if } \ln\frac{\int p(y_t, z, x_t|\theta)dz}{\int p(\bar{y}_t, z, x_t|\theta)dz} > 1 \\ E_z[s(y_t, z, x_t)|\theta(\mu_t)] & \\ -E_z[s(\bar{y}_t, z, x_t)|\theta(\mu_t)] + \frac{1}{n}\bar{\alpha} & \textbf{otherwise} \end{cases}$$

where $\bar{y}_t = arg\max_{y\neq y_t}\int p(y, z, x_t|\theta)dz$

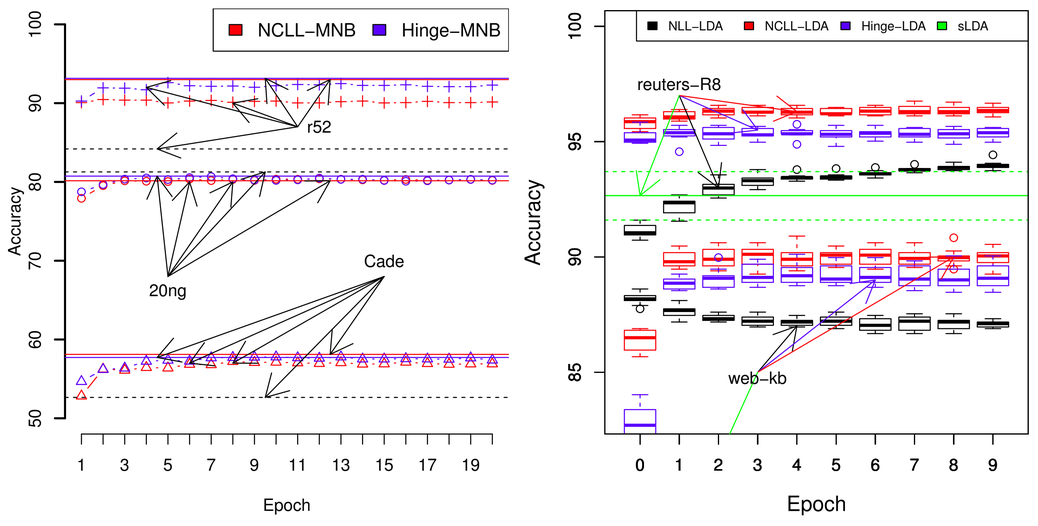
5 Experimental Analysis

Simulated Data



(a) NCLL Loss (b) Hinge Loss

Discriminative Learning with Multinomial-NB and LDA



(a) Discriminative Multinomial Naive Bayes (b) Discriminative Online LDA

Acknowledgements

This work has been partially funded from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 619209 (AMIDST project).