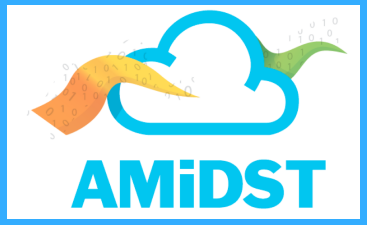
Stochastic Discriminative EM

Andres R. Masegosa[†] [‡]

[†] Dept. of Computer and I.S.
Norwegian University of Science
and Technology
Trondheim, Norway

[‡] Dept. of Computer Science and A. I.
University of Granada
Granada, Spain

Outline



Part I

Introduction

Definitions

- **Prediction Problem**:
  - Y variable to be predicted (discrete, continuous or vector-value).
  - X the predictive variables.

Definitions

- **Prediction Problem**:
  - Y variable to be predicted (discrete, continuous or vector-value).
  - X the predictive variables.

- **Probabilistic Generative Models** (Exponential Family):
  - Our models indexed by θ defines a joint probability $p(y, x|\theta)$.
  - E.g.: Naive Bayes, (Gaussian) Bayesian networks, Dynamic BNs, Latent Dirichlet allocation, etc.

Definitions

- **Prediction Problem**:
  - Y variable to be predicted (discrete, continuous or vector-value).
  - X the predictive variables.
- **Probabilistic Generative Models** (Exponential Family):
  - Our models indexed by θ defines a joint probability $p(y, x|\theta)$.
  - E.g.: Naive Bayes, (Gaussian) Bayesian networks, Dynamic BNs, Latent Dirichlet allocation, etc.
- **Probabilistic Discriminative Models**:
  - The models indexed by θ defines a conditional probability $p(y|x, \theta)$.
  - E.g.: Logistic Regression, Linear Regression, Generalized Linear Models, Conditional Random Fields, etc.

Definitions

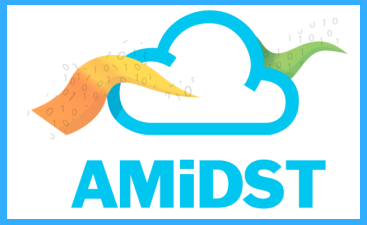
Generative Models

- Missing Data
- Hidden Variables
- Holistic Modelling
- Worst prediction performance
- Model a joint probability (harder and inaccurate)

Discriminative Models

- No Missing Data
- No Hidden Variables
- No Holistic Modelling
- Better prediction performance
- Model a cond. probability (simpler and targeted)

Definitions



Discriminative Learning of Generative Models

- Missing Data
- Hidden Variables
- Holistic Modelling
- Good prediction performance (?)
- Model a prediction-targeted-joint probability distribution

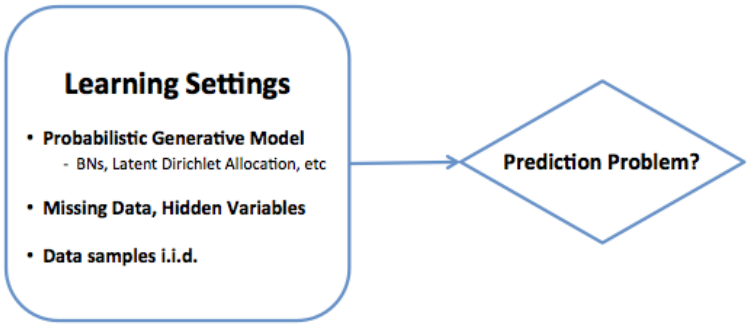
" Frequentist-based Solution"

The BIG picture

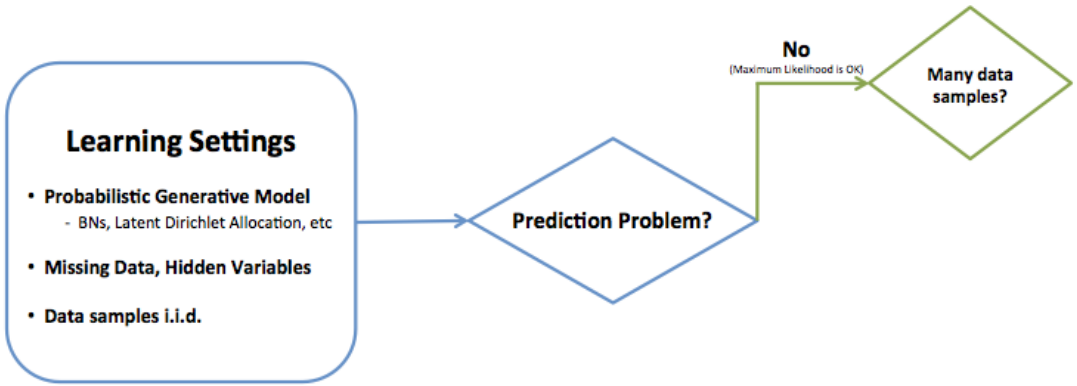
Learning Settings

- **Probabilistic Generative Model**
  - BNs, Latent Dirichlet Allocation, etc
- **Missing Data, Hidden Variables**
- **Data samples i.i.d.**

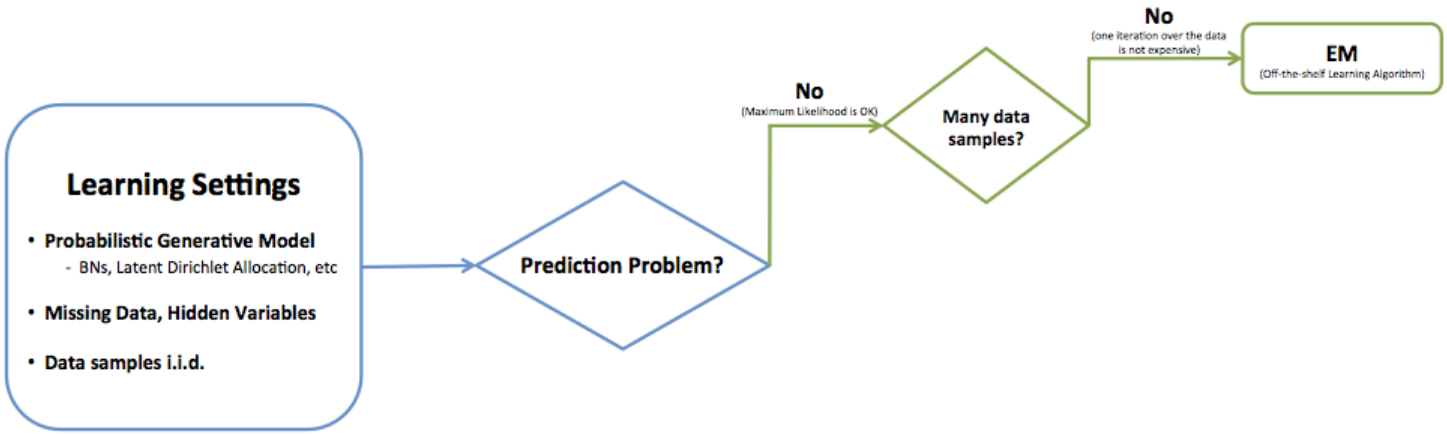
The BIG picture



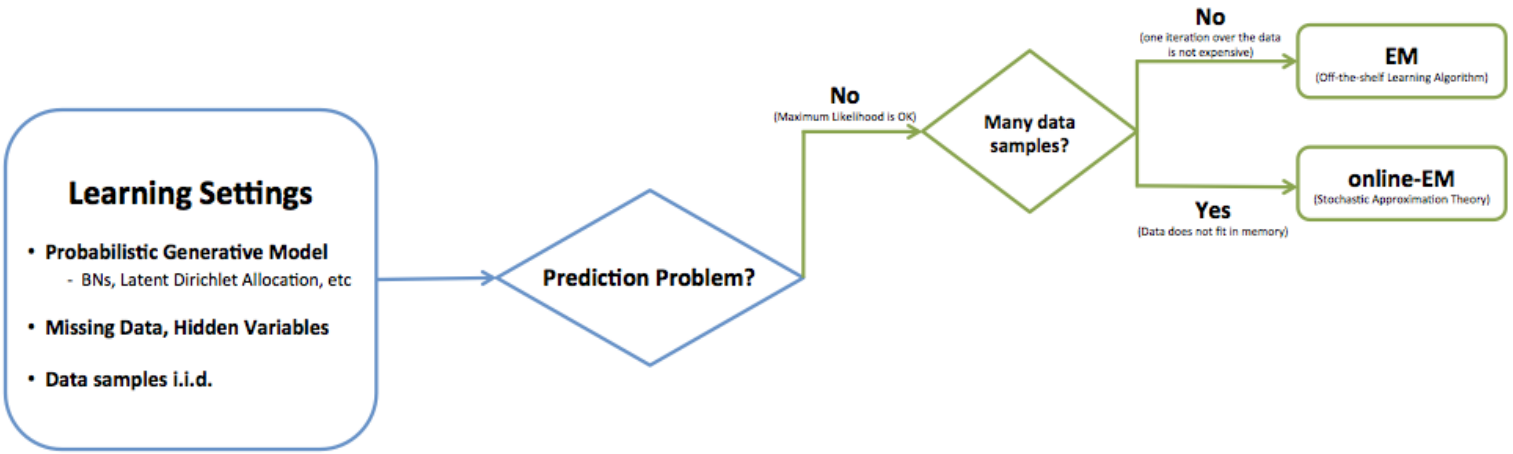
The BIG picture



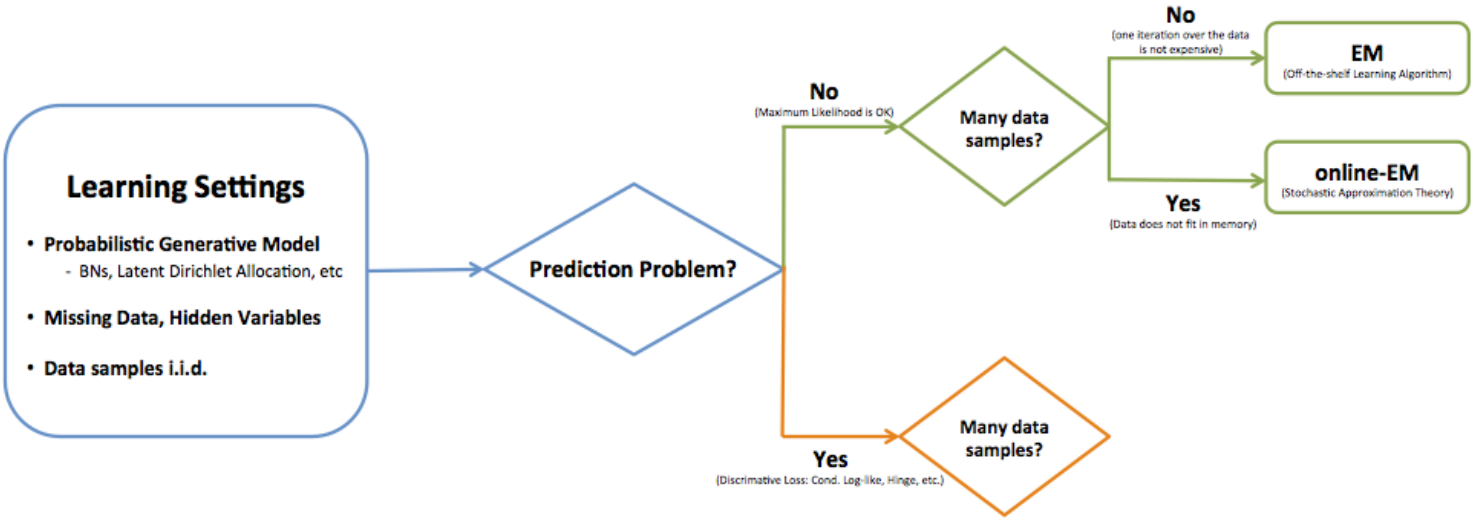
The BIG picture



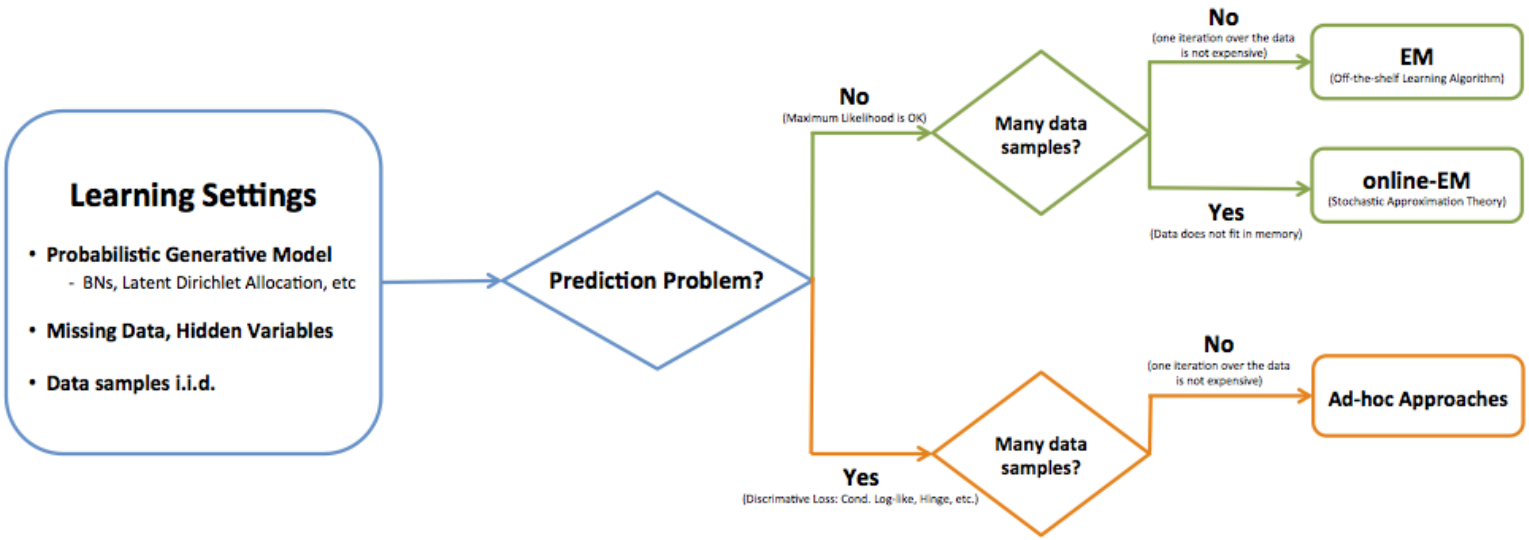
The BIG picture



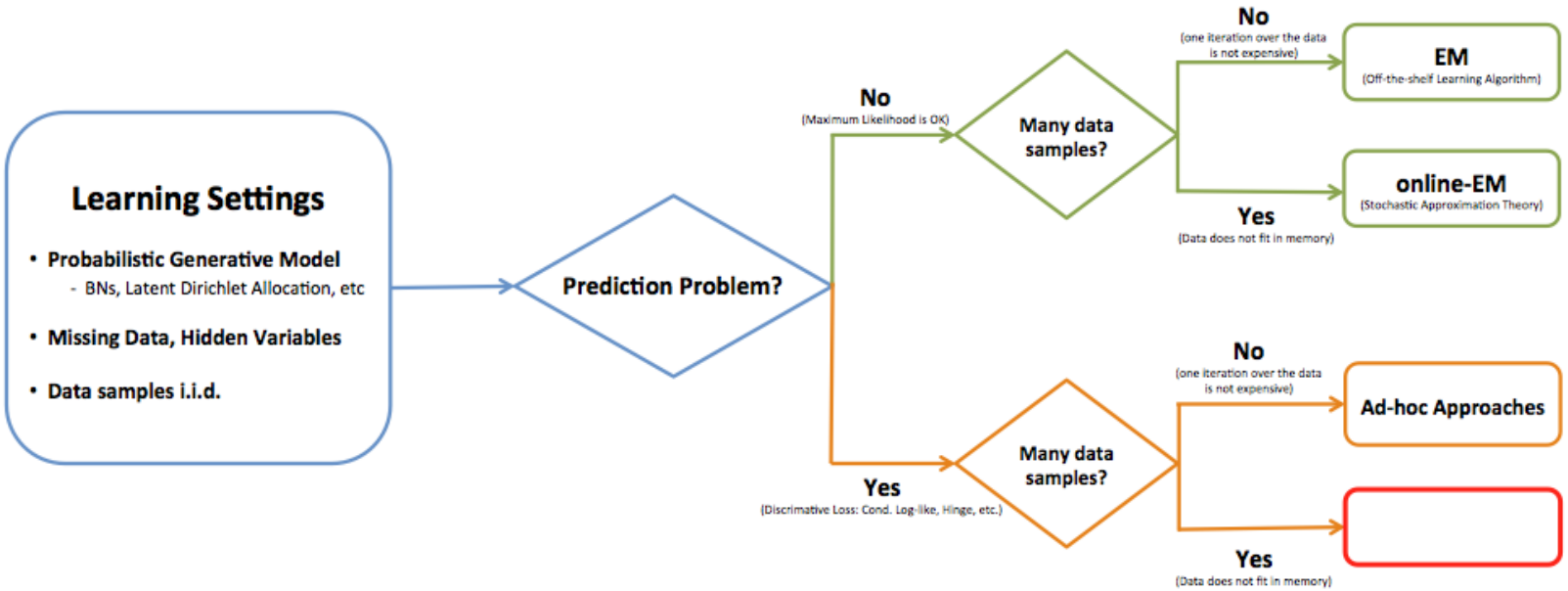
The BIG picture



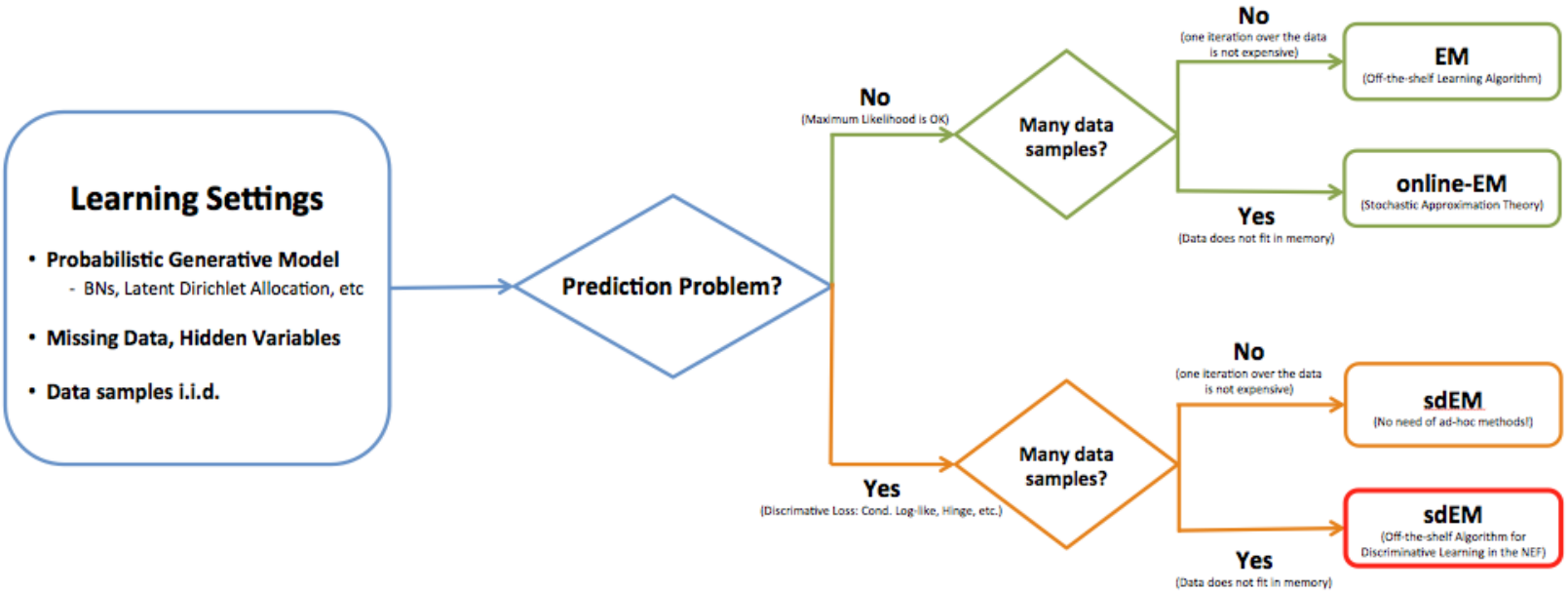
The BIG picture



The BIG picture



The BIG picture



Part II

Previous Knowledge (Maximum Likelihood)

Assumption

- The data-generating model $p(y, x|\theta)$ in **the exponential family** with a natural or canonical representation:

$$p(y, x|\theta) \propto e^{\theta \cdot s(y,x) - A(\theta)}$$

Assumption

- The data-generating model $p(y, x|\theta)$ in **the exponential family** with a natural or canonical representation:

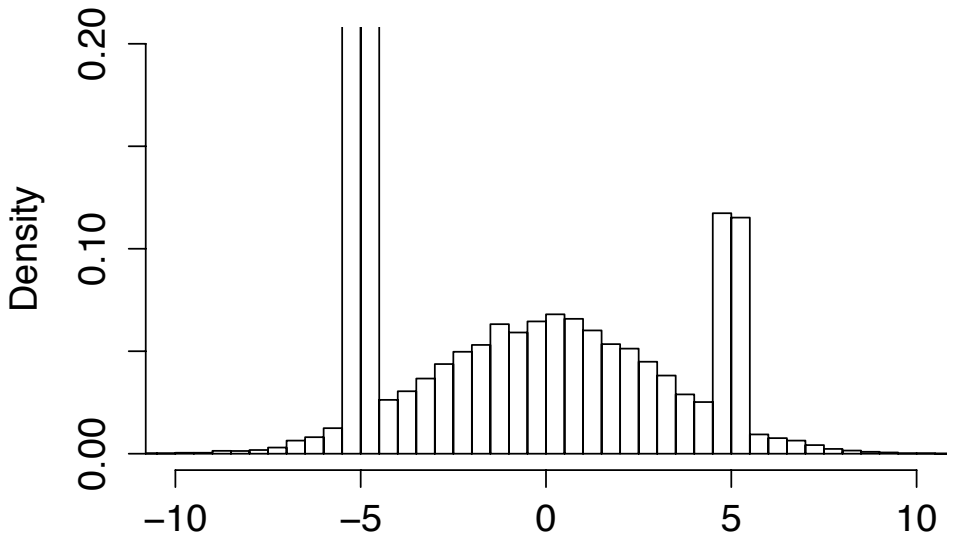
$$p(y, x|\theta) \propto e^{\theta \cdot s(y,x) - A(\theta)}$$

- The **moment or expectation parameters** η are defined:

$$\eta = E[s(y, x)|\theta]$$

  - There is a **1-to-1** relation between θ and η parameters.

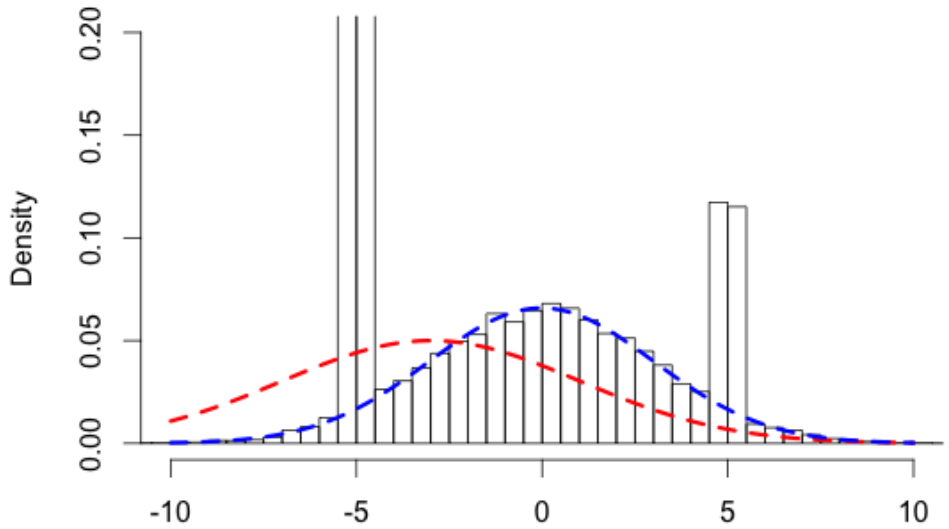
Maximum Likelihood Estimation



Distribution of the data $\pi(x, y) = \pi(x|y)\pi(y)$**:**

- Two classes with equal prior.
- Blue class: $\pi(x|y = -1) \sim N(0, 3)$
- Red class: $\pi(x|y = 1) \sim 0.8 \cdot N(-5, 0.1) + 0.2 \cdot N(5, 0.1)$

Maximum Likelihood Estimation



Generative Learning or Maximum Likelihood:

- The model to be fitted is $p(y, x)$ assumes $p(x|y)$ is univariate Gaussian.
- Prediction Accuracy around 78%

A new look at maximum likelihood estimation

ID	X
1	0
2	0
3	1
4	0
5	1

Estimate $P(X = 0)$

A new look at maximum likelihood estimation

ID	X	$n_i^{(0)}$	$n_i^{(1)}$
1	0	1	0
2	0	2	0
3	1	2	1
4	0	3	1
5	1	3	2
		3/5	2/5

1. Counting...

- $n_{i+1}^{(0)} = n_i^{(0)} + I[x_i == 0]$
- $n_{i+1}^{(1)} = n_i^{(1)} + I[x_i == 1]$

and finally normalize $\bar{n}_N^{(0)} = n_N^{(0)}/N$.

A new look at maximum likelihood estimation

ID	X	$n_i^{(0)}$	$n_i^{(1)}$
1	0	1	0
2	0	2	0
3	1	2	1
4	0	3	1
5	1	3	2
		3/5	2/5

1. **Counting...**
   - $n_{i+1}^{(0)} = n_i^{(0)} + I[x_i == 0]$
   - $n_{i+1}^{(1)} = n_i^{(1)} + I[x_i == 1]$

   and finally normalize $\bar{n}_N^{(0)} = n_N^{(0)}/N$.

2. **Compute parameters from countings**
   - $\theta^{(0)} = \bar{n}_N^{(0)}/(\bar{n}_N^{(0)} + \bar{n}_N^{(1)})$
   - $\theta^{(1)} = \bar{n}_N^{(1)}/(\bar{n}_N^{(0)} + \bar{n}_N^{(1)})$

A new look at maximum likelihood estimation

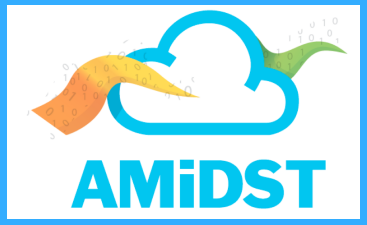
ID	X	$n_i^{(0)}$	$n_i^{(1)}$
1	0	1/1	0/1
2	0	2/2	0/2
3	1	2/3	1/3
4	0	2/4	2/4
5	1	3/5	2/5
		3/5	2/5

1. **Normalized counting:**
   - $\bar{n}_{t+1}^{(0)} = (1-\rho_t)\bar{n}_t^{(0)} + \rho_t I[x_i == 0]$
   - $\bar{n}_{t+1}^{(1)} = (1-\rho_t)\bar{n}_t^{(1)} + \rho_t I[x_i == 1]$

   where $\rho_t = \frac{1}{t}$.

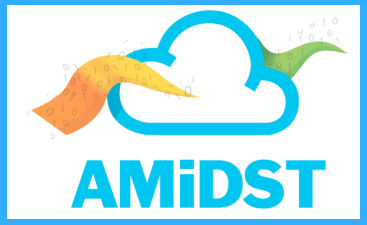
2. **Compute parameters from countings**
   - $\theta^{(0)} = \bar{n}_N^{(0)} / (\bar{n}_N^{(0)} + \bar{n}_N^{(1)})$
   - $\theta^{(1)} = \bar{n}_N^{(1)} / (\bar{n}_N^{(0)} + \bar{n}_N^{(1)})$

A new look at maximum likelihood estimation



- $\bar{n}^{(0)}$ **and** $\bar{n}^{(1)}$ **can also parameterize** $P(X|\bar{n}^{(0)}, \bar{n}^{(1)})$
  - They are the **expectation/moment parameters** of the binomial.

A new look at maximum likelihood estimation



- $\bar{n}^{(0)}$ **and** $\bar{n}^{(1)}$ **can also parameterize** $P(X|\bar{n}^{(0)}, \bar{n}^{(1)})$
  - They are the **expectation/moment parameters** of the binomial.

- Normalized counting is an **iterative updating of the expectation parameters**:

$$\bar{n}_{t+1} = (1 - \rho_t)\bar{n}_t + \rho_t s(x_t)$$

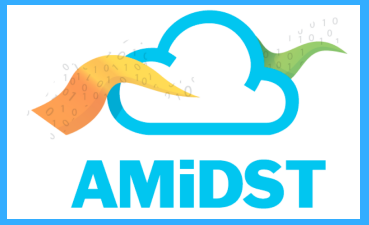
where $s(x) = (I[x == 0], I[x == 1])$ is the **sufficient statistics function**.

A new look at maximum likelihood estimation

- **After some maths....**:

$$\bar{n}_{t+1} = (1-\rho_t)\bar{n}_t + \rho_t s(x_t)$$

A new look at maximum likelihood estimation



- **After some maths....**:

$$\begin{aligned} \bar{n}_{t+1} &= (1-\rho_t)\bar{n}_t + \rho_t s(x_t) \\ &= \bar{n}_t + \rho_t \left(s(x_t) - \bar{n}_t\right) \end{aligned}$$

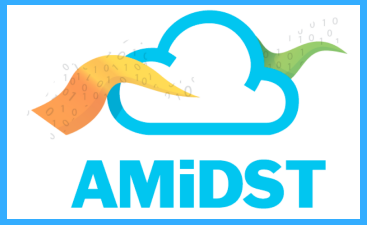
A new look at maximum likelihood estimation

- **After some maths....**:

$$\begin{aligned} \bar{n}_{t+1} &= (1-\rho_t)\bar{n}_t + \rho_t s(x_t) \\ &= \bar{n}_t + \rho_t \left(s(x_t) - \bar{n}_t\right) \\ &= \bar{n}_t + \rho_t \frac{\tilde{\partial} \ln p(x_t|\bar{n}_t)}{\tilde{\partial}\bar{n}} \end{aligned}$$

where $\tilde{\partial}$ denotes the *natural* gradient (Riemanian gemotry).

A new look at maximum likelihood estimation



- **After some maths....**:

$$\begin{aligned} \bar{n}_{t+1} &= (1-\rho_t)\bar{n}_t + \rho_t s(x_t) \\ &= \bar{n}_t + \rho_t \left(s(x_t) - \bar{n}_t\right) \\ &= \bar{n}_t + \rho_t \frac{\tilde{\partial} \ln p(x_t|\bar{n}_t)}{\tilde{\partial}\bar{n}} \end{aligned}$$

  where $\tilde{\partial}$ denotes the *natural* gradient (Riemanian gemotry).

- Sequential updating of "normalized sufficient statistics" is equivalent to a **stochastic natural gradient ascent** method over the moment parameters:

A new look at maximum likelihood estimation

- **After some maths....**:

$$\begin{aligned} \bar{n}_{t+1} &= (1-\rho_t)\bar{n}_t + \rho_t s(x_t) \\ &= \bar{n}_t + \rho_t \left(s(x_t) - \bar{n}_t\right) \\ &= \bar{n}_t + \rho_t \frac{\tilde{\partial} \ln p(x_t|\bar{n}_t)}{\tilde{\partial}\bar{n}} \end{aligned}$$

  where $\tilde{\partial}$ denotes the *natural* gradient (Riemanian gemotry).

- Sequential updating of "normalized sufficient statistics" is equivalent to a **stochastic natural gradient ascent** method over the moment parameters:
  - **Stochastic approximation theory** guarantees the convergence of the above iteration if

$$\sum_t \rho_t = \infty \quad \sum_t \rho_t^2 < \infty$$

Part III

Discriminative Learning with sdEM

Discriminative learning

- **The above algorithm also works for other loss functions**:

$$\bar{n}_{t+1} \quad = \quad \bar{n}_t - \rho_t \frac{\tilde{\partial}\ell(y_t, x_t|\bar{n}_t)}{\tilde{\partial}\bar{n}}$$

  - The convergence is guarantee by **stochastic approximation theory**.

Discriminative learning

Theorem

In the exponential family, the natural gradient of a loss function with respect to the expectation/moment parameters equals the gradient of the loss function with respect to the natural parameters,

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

where $I(\eta)^{-1}$ *is the inverse of the Fisher information matrix for* $p(y, x|\theta)$.

Discriminative learning

- The **negative conditional log-likelihood**,

$$\ell(y_t, x_t|\bar{n}_t) = -\ln p(y_t|x_t, \bar{n}_t) = -\ln p(y_t, x_t|\bar{n}_t) + \ln p(x_t|\bar{n}_t)$$

Discriminative learning

- The **negative conditional log-likelihood**,

$$\ell(y_t, x_t|\bar{n}_t) = -\ln p(y_t|x_t, \bar{n}_t) = -\ln p(y_t, x_t|\bar{n}_t) + \ln p(x_t|\bar{n}_t)$$

- The **updating equation**:

$$\bar{n}_{t+1} \quad = \quad \bar{n}_t + \rho_t \left(s(y_t, x_t) - E_y[s(y, x_t)|\bar{n}_t]\right)$$

Discriminative learning

- The **negative conditional log-likelihood**,

$$\ell(y_t, x_t|\bar{n}_t) = -\ln p(y_t|x_t, \bar{n}_t) = -\ln p(y_t, x_t|\bar{n}_t) + \ln p(x_t|\bar{n}_t)$$

- The **updating equation**:

$$\bar{n}_{t+1} \quad = \quad \bar{n}_t + \rho_t \left(s(y_t, x_t) - E_y[s(y, x_t)|\bar{n}_t]\right)$$

  - For a naive Bayes classifier, the iteration equations are simply expressed:

$$\bar{n}_{t+1}^{(0)} = \bar{n}_t^{(0)} + \rho_t(1 - p(y = 0|x_t)) \quad \text{if } y_t == 0.$$

$$\bar{n}_{t+1}^{(1)} = \bar{n}_t^{(1)} - \rho_t p(y = 1|x_t) \quad \text{if } y_t == 0.$$

Discriminative learning

- The **Hinge** or **max-margin** loss,

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

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

Discriminative learning

- The **Hinge** or **max-margin** loss,

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

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 **updating equation**:

$$\bar{n}_{t+1} = \bar{n}_t + \rho_t \left(s(y_t, x_t) - s(\bar{y}_t, x_t)\right) \quad if \quad \ln \frac{p(y_t, x_t|\theta)}{p(\bar{y}_t, x_t|\theta)} < 1$$

Discriminative learning

- The **Hinge** or **max-margin** loss,

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

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 **updating equation**:

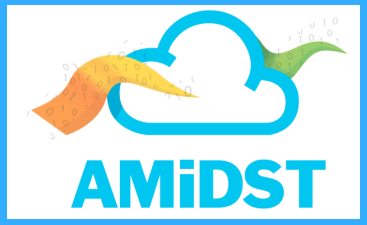
$$\bar{n}_{t+1} = \bar{n}_t + \rho_t \left(s(y_t, x_t) - s(\bar{y}_t, x_t)\right) \quad if \quad \ln \frac{p(y_t, x_t|\theta)}{p(\bar{y}_t, x_t|\theta)} < 1$$

  - For a naive Bayes classifier, the iteration equations are simply expressed:

$$\bar{n}^{(0)}_{t+1} = \bar{n}^{(0)}_t + \rho_t \cdot 1 \quad \text{if } y_t == 0 \text{ and } \ln \frac{p(y_t, x_t|\theta)}{p(\bar{y}_t, x_t|\theta)} < 1$$

$$\bar{n}^{(1)}_{t+1} = \bar{n}^{(1)}_t - \rho_t \cdot 1 \quad \text{if } y_t == 0 \text{ and } \ln \frac{p(y_t, x_t|\theta)}{p(\bar{y}_t, x_t|\theta)} < 1$$

Discriminative learning with hidden variables



- The **p(y,z,x) is in the exponential family** and $s(y, z, x)$ the suff. stats.

Discriminative learning with hidden variables

- The **p(y,z,x) is in the exponential family** and $s(y, z, x)$ the suff. stats.

- The **negative conditional log-likelihood**,

$$\bar{n}_{t+1} \quad = \quad \bar{n}_t + \rho_t \left(E_z[s(y_t, z, x_t)|\bar{n}_t] - E_{yz}[s(y, z, x_t)|\bar{n}_t] \right)$$

Discriminative learning with hidden variables

- The **p(y,z,x) is in the exponential family** and $s(y, z, x)$ the suff. stats.

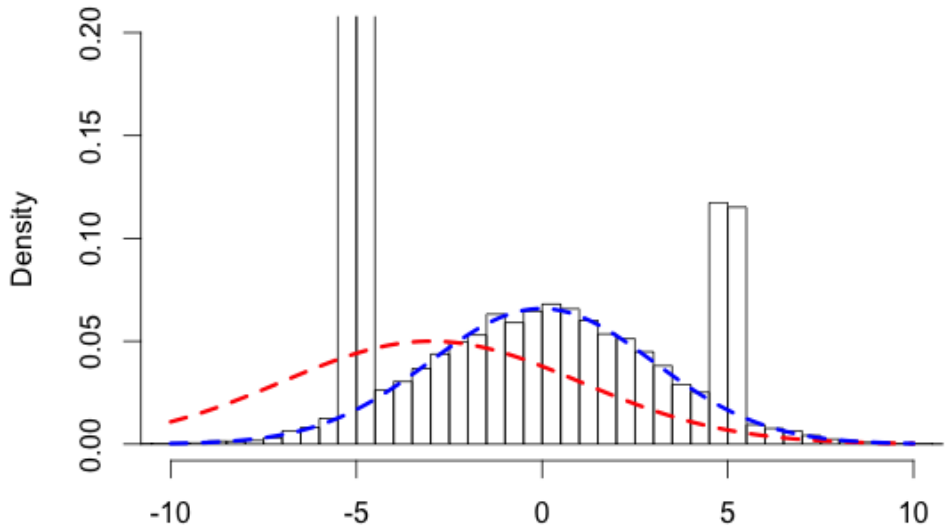
- The **negative conditional log-likelihood**,

$$\bar{n}_{t+1} = \bar{n}_t + \rho_t \left(E_z[s(y_t, z, x_t)|\bar{n}_t] - E_{yz}[s(y, z, x_t)|\bar{n}_t] \right)$$

- The **Hinge loss**:

$$\bar{n}_{t+1} = \bar{n}_t + \rho_t \left(E_z[s(y_t, z, x_t)|\bar{n}_t] - E_z[s(\bar{y}_t, z, x_t)|\bar{n}_t] \right)$$

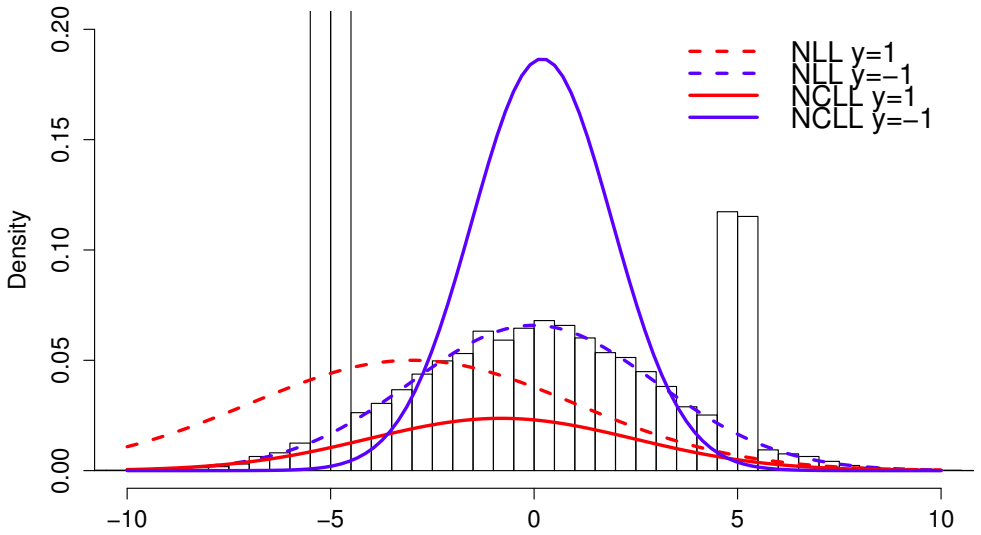
What is discriminative learning?



Generative Learning or Maximum Likelihood:

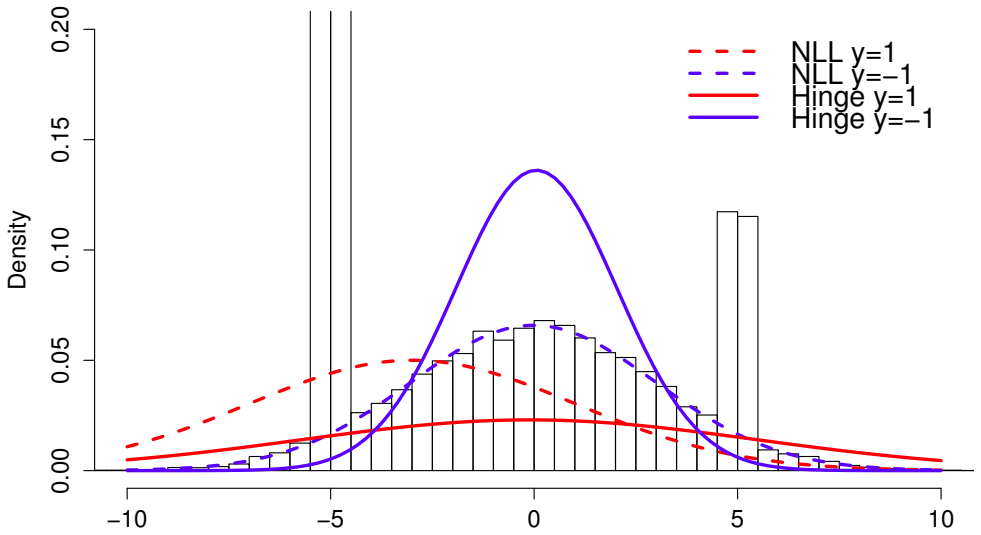
- The model to be fitted is $p(y, x)$ assumes $p(x|y)$ is univariate Gaussian.
- Prediction Accuracy around 78%

What is discriminative learning?



Discriminative Learning with the NCLL loss: 90.4% of accuracy

What is discriminative learning?



Discriminative Learning with the Hinge Loss: 90.6% of accuracy

What's the relationship with EM?

Algorithm 1 Standard EM

1: Choose some θ_0;
2: $t = 0$;
3: **repeat**
4: $\quad n_0 = 0$
5: $\quad$ **for** $i = 1, \ldots, N$ **do**
6: $\quad\quad$ **E-Step**: $n_{i+1} = n_i + (E_z[s(y_i, z, x_i)|\theta_t])$
7: $\quad$ **end for**
8: $\quad \bar{n}_t = n_N/N$
9: $\quad$ **M-Step**: $\theta_{t+1} = \theta(\bar{n}_t)$;
10: $\quad t = t + 1$;
11: **until** convergence
12: **return** $\theta(\bar{n}_t)$;

What's the relationship with EM?

Algorithm 2 Standard EM

1: Choose some θ_0;
2: $t = 0$;
3: **repeat**
4: $\quad \bar{n}_0 = 0$
5: $\quad$ **for** $i = 1, \ldots, N$ **do**
6: $\quad\quad$ **E-Step**: $\bar{n}_{i+1} = (1 - \frac{1}{i})\bar{n}_i + \frac{1}{i} \cdot (E_z[s(y_i, z, x_i)|\theta_t])$
7: $\quad$ **end for**
8: $\quad$ **M-Step**: $\theta_{t+1} = \theta(\bar{n}_N)$;
9: $\quad t = t + 1$;
10: **until** convergence
11: **return** $\theta(\bar{n}_t)$;

What's the relationship with EM?

Algorithm 3 Online EM

Require: D is randomly shuffled.

Choose some θ_0;
$t = 0$;
$\bar{n}_0 = 0$
repeat
  for $i = 1, \ldots, N$ **do**
    E-Step: $\bar{n}_{t+1} = (1 - \rho_t)\bar{n}_t + \rho_t \cdot (E_z[s(y_i, z, x_i)|\theta_t])$
    M-Step: $\theta_{t+1} = \theta(\bar{n}_t)$;
    $t = t + 1$;
  end for
until convergence
return $\theta(\bar{n}_t)$;

What's the relationship with EM?

Algorithm 4 sdEM with NCLL loss

Require: D is randomly shuffled.

Choose some θ_0;
$t = 0$;
$\bar{n}_0 = 0$
repeat
  for $i = 1, \ldots, N$ **do**
    E-Step: $\bar{n}_{t+1} = \bar{n}_t + \rho_t \cdot (E_z[s(y_i, z, x_i)|\theta_t] - E_{yz}[s(y_i, z, x_i)|\theta_t])$
    M-Step: $\theta_{t+1} = \theta(\bar{n}_t)$;
    $t = t + 1$;
  end for
until convergence
return $\theta(\bar{n}_t)$;

Some more details about sdEM

- Employment of a **conjugate prior** $p(\theta|\alpha)$

$$arg\min_{\theta} \sum_{(y_i,x_i)\in D} \ell(y_i, x_i, \theta) + \ln p(\theta|\alpha)$$

  - Guarantees convergence: $\ln p(\theta|\alpha)$ is a log-barrier function.

Some more details about sdEM

- Employment of a **conjugate prior** $p(\theta|\alpha)$

$$arg\min_{\theta} \sum_{(y_i,x_i)\in D} \ell(y_i, x_i, \theta) + \ln p(\theta|\alpha)$$

  - Guarantees convergence: $\ln p(\theta|\alpha)$ is a log-barrier function.

- **Unbiased estimates of the expected sufficient statistics**:

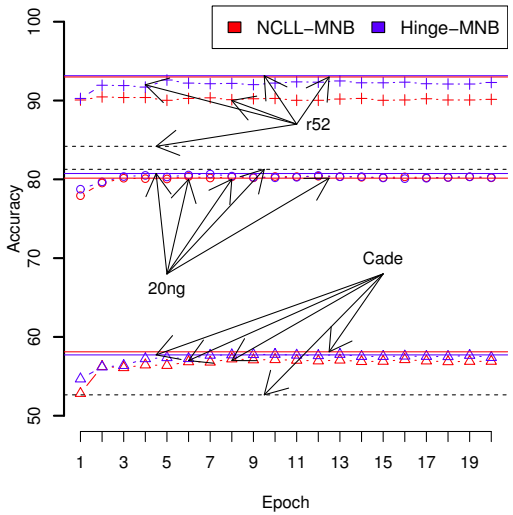
$$E_z[s(y_t, z, x_t)|\theta] = \sum_z p(z|y_t, x_t, \theta)s(y_t, z, x_t)$$

  - Collapsed Gibbs sampling is OK!.
  - Variational inference provides unbiased estimates. How sdEM would work?

Part IV

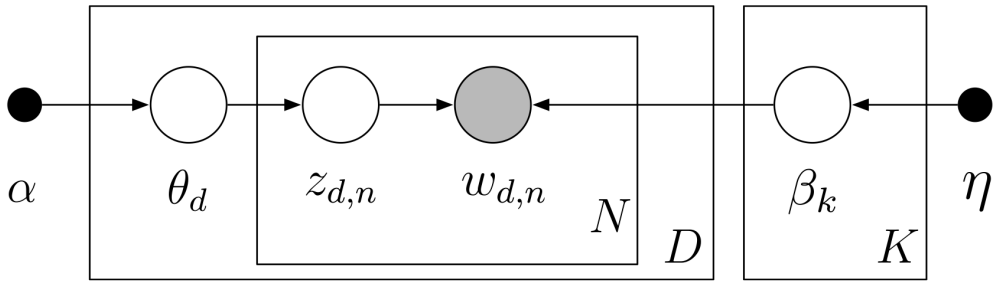
sdEM for Text Classification

Discriminative Multinomial NB



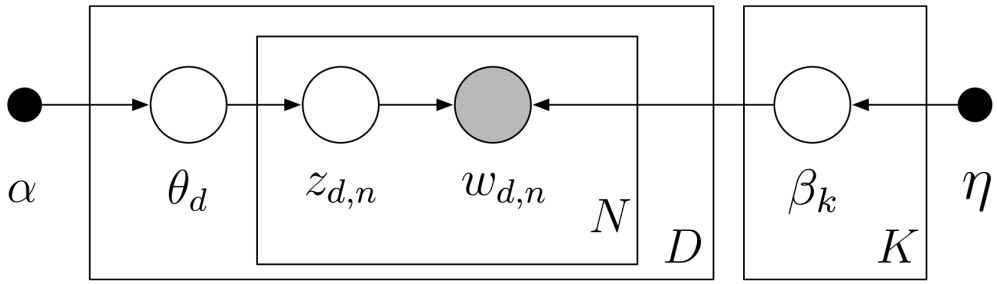
- **Comparison against**: L2-regularized Logistic Regression and the primal L2-regularized SVM (Liblinear toolkit v.18).
- A simple to implement Multinomial NB using sdEM becomes **competitive with highly optimized algorithms**.

Supervised LDA models using sdEM



- LDA belongs to the exponential family.
- Just use sdEM for discriminative training of LDA.
- Online and out-of-core supervised or max-margin LDA model.

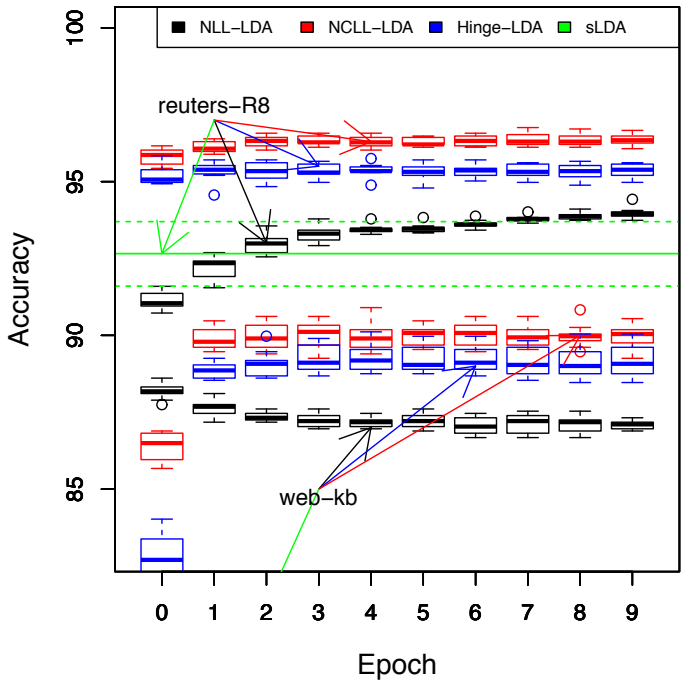
Supervised LDA models using sdEM



- LDA belongs to the exponential family.
- Just use sdEM for discriminative training of LDA.
- Online and out-of-core supervised or max-margin LDA model.

sdEM is an off-the-shelf algorithm for discriminative learning

Supervised LDA models using sdEM



- **Comparison against** sLDA (Blei& McAuliffe, NIPS2007) .
- Outstanding results.

Part V

Conclusions & Future Works

Conclusions

- **sdEM** is a learning algorithm for **discriminative training of generative models**.
- It can be seen as a **stochastic natural gradient descent algorithm** for minimizing general discriminative loss functions.
- It is comparatively **simpler and easier to implement** (and debug) than other ad-hoc approaches.

Conclusions

- **sdEM** is a learning algorithm for **discriminative training of generative models**.
- It can be seen as a **stochastic natural gradient descent algorithm** for minimizing general discriminative loss functions.
- It is comparatively **simpler and easier to implement** (and debug) than other ad-hoc approaches.

It does not solve the inference problem!

Future Works

- **Missing Data**:
  - Logistic Regression with missing data

Future Works

- **Missing Data**:
  - Logistic Regression with missing data = sdEM + NB + NCLL loss

Future Works

- **Missing Data**:
  - Logistic Regression with missing data = sdEM + NB + NCLL loss
  - Linear SVM with missing data

Future Works

- **Missing Data**:
  - Logistic Regression with missing data = sdEM + NB + NCLL loss
  - Linear SVM with missing data = sdEM + NB + Hinge loss

Future Works

- **Missing Data**:
  - Logistic Regression with missing data = sdEM + NB + NCLL loss
  - Linear SVM with missing data = sdEM + NB + Hinge loss

- **Interpretation of a discriminatevely trained generative model**:
  - Discriminatives LDA models uncover topics that defines just the differences between classes.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 619209