## Sparse Classification RBM

Weidong Liang

## Mathematical Formulation

Parameter Set:

## Model Explanation

Y: click (y=1) or no click (y=0).

: feature value of the i-th feature class (there are total of C feature classes).

:i-th hidden unit value.

## Derivation of Properties

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
| Note that:  Hence: |

|  |
| --- |
| Note that  Hence: |

|  |
| --- |
| Note that:  Hence: |

|  |
| --- |
|  |

## Derivation of Learning Algorithm

Generative Learning

|  |
| --- |
|  |

Discriminative Learning

|  |
| --- |
|  |

Hybrid Learning

Derivatives of Gradients

Parameter Set:

|  |
| --- |
|  |

Parameter Updates

|  |
| --- |
|  |

|  |
| --- |
|  |

Using CD-k Approximation

|  |
| --- |
|  |

Combine All

|  |
| --- |
|  |

CD-K Updates

|  |
| --- |
| For t = 0, …, k – 1 do  For i = 1, …, H do sample  For i = 1, …, C do sample  Do Sample  For i = 1,…H do |

## Maximum Likelihood and the Delta Rule

* Maximum Likelihood
* Mini-batch Gradient Ascent (Delta Rule)

## Learning Algorithm

|  |
| --- |
| //CD-k, only if  For t = 0, …, k – 1 do  For j = 1, …, H do sample [O(C+H)]  For i = 1, …, C do sample (Computation Intensive! O(**H\*V**))  Do Sample [O(H)]  For i = 1,…H do [O(H)]  //Gradient Calculation  [O(H\*C)]  //Parameter Update  [Note, storing two sets of parameters would be memory hungry, store the differences only??] |

Note:

Efficient calculation of the following is the key to the performance of CD-k!

Naïve implementation requires O(|X|\*|H|)

Use Mini-batch + Cache Strategy??

Nonexact sampling??MCMC???Importance Sampling, rejection sampling etc to avoid the normalization constant? ????

What else??

## SparseClassRBM Verses Logistic Regression

Note that in logisitc regression, we have:

And for SparseClassRBM, we have

Setting H=1, ,d=1,U = 0, we have

which is a form of logistic regression.

Therefore, we can view logistic regression as a special form of RBM with less variables.