Advanced Machine Learning (HW #2)

Due date: June 17

Submitting Homework

Please zip each of your source code into a single compress file and name the file using this format: `<StudentID>.zip`. (rar, 7z, tar.gz, …etc are all acceptable)

And please upload your report in this format: `<StudentID>.pdf`

Totally there are 2 files should be uploaded. If you have a revision, please append the version number after your student id, for instance:

`0123456_v2.zip 0123456_v3.pdf`

Please **don't** upload the dataset, because the file size is large.

1. Bernoulli Mixture Model

In this exercise, you will implement a Bernoulli mixture model for handwriting recognition. Consider a set of binary variables $\mathbf{x} = \{x_i\}$, each of which represents a pixel in an image. Assume each image contains D pixels, and $x_i \sim \text{Bernoulli}(\mu_i)$, so that

$$p(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i} (1 - \mu_i)^{(1-x_i)}$$

where $\boldsymbol{\mu} = \{\mu_i\}$, $0 \le \mu_i \le 1$ for all $i = 1, \ldots, D$. Now, we consider a mixture of Bernoulli distributions as follows to capture the patterns of images

$$p(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \pi_k \, p(\mathbf{x} \mid \boldsymbol{\mu}_k)$$

where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\} = \{\mu_{ik}\}$ and $\boldsymbol{\pi} = \{\pi_k\}$, $0 \le \pi_k \le 1$ and $\sum_{k=1}^{K} \pi_k = 1$.

(1) Please apply the EM algorithm for Bernoulli mixture model on MNIST handwriting digit data set. You can obtain the data as the following links

1. http://yann.lecun.com/exdb/mnist/train-images-idx3-ubyte.gz
2. http://yann.lecun.com/exdb/mnist/train-labels-idx1-ubyte.gz
3. http://yann.lecun.com/exdb/mnist/t10k-images-idx3-ubyte.gz
4. http://yann.lecun.com/exdb/mnist/t10k-labels-idx1-ubyte.gz

where 1 and 2 are training data, 3 and 4 are test data. To save your work, please call `DataPrep` to read the data set. This will help you to convert the images to black and white color with 400 dimensional column vectors. Totally there are 50,000 instances for training and 10,000 for testing. For model visualization, you

can use `ShowModel` to display the training results as an example shown in Figure 1. You are welcome to implement your own visualizer.

(2) Please evaluate the log likelihood per image on training data and test data respectively.

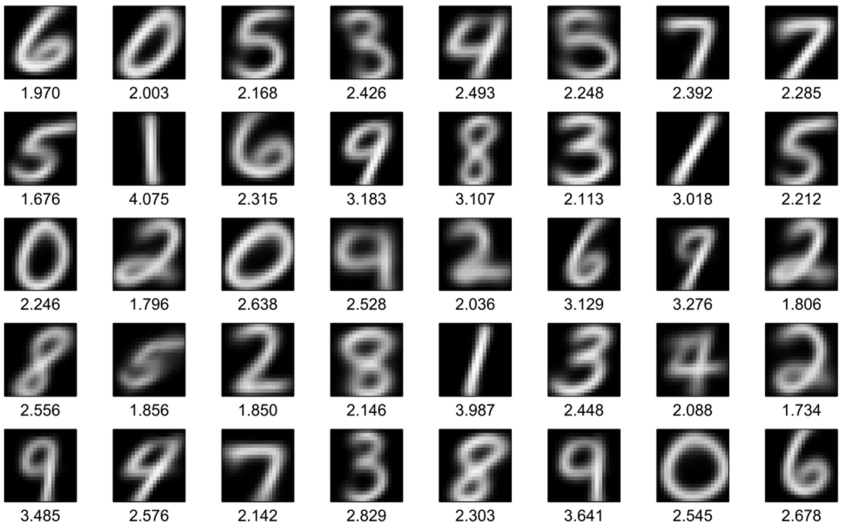

Figure 1. An example of Bernoulli mixture model with $K = 40$.

The images and labels represent the $\boldsymbol{\mu}_k$ and π_k (in percentage) respectively,

Appendix

```
[ train, test ] = DataPrep( path_to_dataset )
    path_to_dataset     the root directory of MNIST dataset
    train               a struct for train data, total 50000 samples
        train.images    dim = 400 x 50000
        train.labels    dim = 50000 x 1
    test                a struct for test data, total 10000 samples
        test.images     dim = 400 x 10000
        test.labels     dim = 10000 x 1

ShowModel( mu, pi, n_rows, n_cols, ind )
    mu          μ, the parameters of Bernoulli distributions
                dim = <# pixels per image> x <# images>
    pi          π, mixing weights
                dim = <# mixture components>
    n_rows      the number of rows for displaying a panel of images.
    n_cols      the number of columns for displaying a panel of images.
    ind         the indices of mixture components
```