050035	**Machine Learning**	**Homework 5**
Summer 2014		**Assigned May 18**
		Due 15:00 May 30

You are to complete this individually (as with all the homeworks in this class). No code from the internet is permitted. All you need is provided.

Question 1 (30 point)

In this question you will implement expectation maximization (EM) for the mixture of Bernoulli distributions model (see Sec. 9.3.3 in PRML).

The Bernoulli distribution models binary variables. Given a vector of D binary variables $\mathbf{x} = (x_1, \ldots, x_D)$, each of which has Bernoulli distribution parameter μ_i , the distribution is:

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

As with the Gaussian, we can consider a *mixture of Bernoulli distributions* model:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k) \tag{2}$$

The equations for EM updates of a mixture of Bernoulli distribution are given as equations 9.56-9.60 in the book by Bishop.

You will implement these updates to fit such a mixture to a collection of handwritten digit images from the MNIST dataset.

- Start by downloading the code skeleton `mob_em.py` and data `digits.mat` from the course website (in the tarball).
- `digits.mat` contains 1000 images from MNIST, of the digits 0-9.
- `mob_em.py` contains code for reading these images, and turning them into binary values. (there is also a matlab version)
- Fill in the code for making the EM updates of the responsibilities `resp` and Bernoulli mixture parameters `Pi` and `Mu`.
- Visualization of the current `Mu` parameters is provided.

In your report include visualizations of the final `Mu` parameters learned using K=5, 10, and 20 mixture components. Briefly comment on the differences using different numbers of components. Note that EM depends on initialization, you may get different results each time you run it.

Please submit your modified `mob_em.py`, along with any other files you use.

Submitting Your Assignment

The assignment must be submitted online at cewebs. In order to simplify grading, you must adhere to the following structure. You must submit two files:

1. You must create an assignment report in PDF format, called `report.pdf`. This report must contain the figures / explanations requested.

2. You must create a tar gzip archive called `code.tgz` containing your code.

   This must contain a single directory called `code` (no sub-directories, no leading path names), in which all of your files must appear[1]. There must be the scripts with the specific names referred to, as well as any other files you create and name.

This homework was adopted from Greg Mori.

[1] This includes the data files and others which are provided as part of the assignment.