Assignment 4: EM and Combining Models

Due April 6 at 11:59pm
90 marks total

This assignment is to be done individually.

Important Note: The university policy on academic dishonesty (cheating) will be taken very seriously in this course. You may not provide or use any solution, in whole or in part, to or by another student.

You are encouraged to discuss the concepts involved in the questions with other students. If you are in doubt as to what constitutes acceptable discussion, please ask! Further, please take advantage of office hours offered by the instructor and the TA if you are having difficulties with this assignment.

DO NOT:

- Give/receive code or proofs to/from other students
- Use Google to find solutions for assignment

DO:

- Meet with other students to discuss assignment (it is best not to take any notes during such meetings, and to re-work assignment on your own)
- Use online resources (e.g. Wikipedia) to understand the concepts needed to solve the assignment

---

Question 1 (20 marks)

Question 9.4 in PRML.

Question 2 (35 marks)

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 $\boldsymbol{x} = (x_1, \ldots, x_D)$, each of which has Bernoulli distribution parameter μ_i, the distribution is:

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

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

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

The equations for EM updates of a mixture of Bernoulli distribution are given as equations 9.56-9.60 in PRML.

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.m` 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.m` contains code for reading these images, and turning them into binary values.
- 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 visualization 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.

Question 3 (35 marks + 5 bonus marks)

In this question you will implement AdaBoost for discriminating between 4s and other digits. A code skeleton `boost_digits.m` is provided. You need to fill in the details of AdaBoost. Unfortunately, the book and the slides have somewhat different versions. The one in the slides is the official version from the original Adaboost paper (see `http://en.wikipedia.org/wiki/Adaboost`). To avoid confusion and to help us grade, please use the one in the book, which is also slightly easier to implement. We'll give bonus marks for a theoretical argument that the two versions are equivalent, or an empirical result where you implement both and show that you get the same answers.

In particular, a function `findWeakLearner.m` is provided. This function chooses the best decision stump (feature, threshold, parity) to minimize weighted 0-1 loss.

Each decision stump is returned from `findWeakLearner.m` as d, p, θ. This represents the weak learner:

$$y_m(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad p\boldsymbol{x}(d(1), d(2)) > p\theta \\ -1 & \text{otherwise} \end{cases}$$

Here $\boldsymbol{x}(d(1), d(2))$ gives you the grayscale value of the pixel at location d_1, d_2 in image $\boldsymbol{x}$.

In your report, provide the final plot of training error and test error produced using AdaBoost. Also include the visualization of the final classifier produced using `visualizeClassifier.m`

Note that a second data file `digits10000.mat` is provided if you wish to experiment with more data than the 1000 in `digits.mat`.

Submitting Your Assignment

You should create a report with the answers to questions and figures described above in PDF format. Make sure it is clear what is shown in each figure. **DO NOT INCLUDE SOURCE CODE.**

Submit your assignment using the *new* online assignment submission server at: `https://courses.cs.sfu.ca`.