# EEE 485-585 FALL 2018 PROBLEM SET 4

**Due Date: 18 December 2018, 17:30**

## Question 1 [25 pts]

Download ps4q1.mat from Moodle. Standardize the data such that it becomes zero mean and unit variance.

**(a)** [1 pt] Plot the standardized data.

**(b)** [2 pts] Write down the code for the $K$-means clustering algorithm (for general $K$) for $l_2$ norm (squared Euclidian distance as loss) $l_1$ norm (absolute distance as loss).

**(c)** [7 pts] Run the $K$-means algorithm for $K = 2$ until convergence for both $l_2$ and $l_1$ norm versions. Assume that the initial prototype vectors are $\boldsymbol{\mu}_1 = (-1, 1)$ and $\boldsymbol{\mu}_2 = (1, -1)$. Plot the resulting clusters at the end of each maximization step for all steps until convergence. (The figure for $K$-means must look like the figure for the $K$-means example that we did in class. The clusters must be shown in a two dimensional plot).

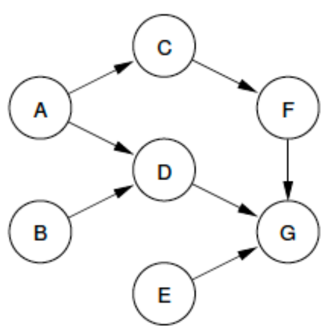
**(d)** [7 pts] Run the $K$-means algorithm for $K = 3$ until convergence for both $l_2$ and $l_1$ norm versions. Assume that the initial prototype vectors are $\boldsymbol{\mu}_1 = (-1, 1)$, $\boldsymbol{\mu}_2 = (1, -1)$ and $\boldsymbol{\mu}_3 = (0, -1)$. Plot the resulting clusters at the end of each maximization step for all steps until convergence. (The figure for $K$-means must look like the figure for the $K$-means example that we did in class. The clusters must be shown in a two dimensional plot).

**(e)** [4 pts] Comment on the differences between the clusters formed using $l_2$ and $l_1$ norm versions. Which one you think is more suitable for this dataset?

**(f)** [4 pts] Comment on the differences between the clusters formed using $K = 2$ and $K = 3$. Which one you think is more suitable for this dataset?

## Question 2 [25 pts]

Consider the following directed graphical model.




**(a)** [4 pts] Write the joint probability of the random variables above in factored form.

**(b)** [4 pts] Prove if $A$ and $E$ are independent.

**(c)** [4 pts] Prove if $A$ and $E$ are conditionally independent given $D$ and $C$.

**(d)** [4 pts] Compute the Markov blanket of $D$.

**(e)** [9 pts] Find expressions for $p(E|G)$ and $p(E|G, A, B)$ in terms of the probabilities in the factored form of the joint distribution. Simplify as much as possible.

## Question 3 [25 pts]

For this question, consider the RBM in the lecture notes. You are only given the energy function $E(\boldsymbol{x}, \boldsymbol{h})$ and the form of the joint distribution $p(\boldsymbol{x}, \boldsymbol{h}) = \exp(-E(\boldsymbol{x}, \boldsymbol{h}))/Z$, where $Z$ is the partition function.

**(a)** [5pts] Prove that

$$p(h_i = 1|\boldsymbol{x}) = \phi\left(\sum_{j=1}^{m} w_{ij}x_j + c_i\right)$$
$$p(x_j = 1|\boldsymbol{h}) = \phi\left(\sum_{i=1}^{n} w_{ij}h_i + b_j\right)$$

where

$$\phi(z) = \frac{1}{1 + e^{-z}}$$

You must show all steps of your derivation to get full credit.

**(b)** [5pts] Prove that

$$p(\boldsymbol{x}) = \frac{1}{Z}e^{\boldsymbol{b}^T\boldsymbol{x} + \sum_{i=1}^{n}\log(1+\exp(c_i + \sum_{j=1}^{m} w_{ij}x_j))}$$

You must show all steps of your derivation to get full credit.

**(c)** [15pts] Given a training sample $\boldsymbol{x}^t = [x_1^t, x_2^t, \ldots, x_m^t]^T$ consider the log-likelihood:

$$l(\theta) = \log p(\boldsymbol{x}^t|\theta)$$

where $\theta = \{w_{ij}, b_j, c_i\}$ is the set of all parameters. Prove the following:

$$\frac{\partial l(\theta)}{\partial w_{ij}} = p(h_i = 1|\boldsymbol{x}^t)x_j^t - \sum_{\boldsymbol{x}} p(\boldsymbol{x})p(h_i = 1|\boldsymbol{x})x_j$$

$$\frac{\partial l(\theta)}{\partial b_j} = x_j^t - \sum_{\boldsymbol{x}} p(\boldsymbol{x})x_j$$

$$\frac{\partial l(\theta)}{\partial c_i} = p(h_i = 1|\boldsymbol{x}^t) - \sum_{\boldsymbol{x}} p(\boldsymbol{x})p(h_i = 1|\boldsymbol{x})$$

You must show all steps of your derivation to get full credit.

## Question 4 [25 pts]

In this question you will build a Restricted Boltzmann Machine (RBM). Then, you will train two logistic regression classifiers (LRC) without any regularization. The first one will be trained by using the pixels of the images as features, and the second one will be trained using the output at the hidden layer of the RBM as inputs.

Specifically, let $X_{train}$ and $X_{test}$ denote your training and test data, respectively. You will train your RBM using $X_{train}$, and learn a transformation, $f(\cdot)$ that maps the visible input to the latent (hidden) outputs. You will then train two LRCs, one with input $X_{train}$ and another one with input $f(X_{train})$.

After the training phase, you will report the accuracy of LRC trained with raw pixels on $X_{train}$ and $X_{test}$, and the accuracy of LRC trained with the output of the RBM on $f(X_{train})$ and $f(X_{test})$. Simulation details are explained below:

The RBM will have binary visible and hidden units. It should be trained using contrastive divergence (CD). Download the dataset ps4q4.mat from Moodle . You can use the whole training dataset if you want to or you can use randomly selected 10000 samples from the training set (since the dataset is large and it can be time consuming to train using the entire training set). You must use the entire test set for evaluation. Include all the simulation code with your submission to get credit from this question. The parameters of your RBM should be as follows:

- Learning rate = 0.05
- Epochs = 20
- Batch size = 10
- K value for CD = try different values and see the effect, even though 1 might work surprisingly well.
- # of hidden components = see below

When selecting the number of hidden components use at least 5 different values. You will notice that setting this value too low will make the performance of the RBM+LRC actually worse than using LRC on pixels directly. Include at least one such case when reporting accuracies, and include at least one case where the performance of RBM+LRC is higher than using only LRC. Also for each value of number of hidden components, plot the resulting weights for each hidden component as images (you will have # of hidden componentsx28x28 images, use subplot when plotting each of them, so that they will be on the same figure). Note that if you are selecting a different subset of training data for each experiment, you need to also report the LRC only score for each experiment. For logistic regression, you can use one of the off-the-shelf software packages.