10-701 Machine Learning - Spring 2012
Problem Set 4

Out: March 21st
In: April 9th

Byron Boots (beb@cs.cmu.edu)
School Of Computer Science, Carnegie Mellon University

- Homework will be done individually: each student must hand in their own answers. It is acceptable for students to collaborate in figuring out answers and helping each other solve the problems. We will be assuming that, as participants in a graduate course, you will be taking the responsibility to make sure you personally understand the solution to any work arising from such collaboration. You also must indicate on each homework with whom you collaborated.

- Homework is due at the beginning of class on the due date. For programming questions, please submit your code and any plot/figures to the AFS submission folder:
`/afs/cs.cmu.edu/academic/class/10701-s12-users/yourandrewid/hw4`
where `yourandrewid` is your Andrew ID. To copy the files to the folder, you can follow these steps:

  - Copy the homework file (pdf format) and other code and data files as needed to your home directory on an Andrew machine, e.g.
```
scp probset4.pdf yourandrewid@linux.andrew.cmu.edu:yourhomepath
password :  ********
```
  - Start an Andrew session, e.g.
```
ssh -l yourandrewid linux.andrew.cmu.edu
password :  *********
```
  - Invoke aklog and copy the files over to your dedicated course folder. You have to insert and listing rights for this folder, e.g.
```
aklog cs.cmu.edu
mkdir /afs/cs.cmu.edu/academic/class/10701-s12-users/yourandrewid/hw4
cp probset4.pdf /afs/cs.cmu.edu/academic/class/
10701-s12-users/yourandrewid/hw4/
```

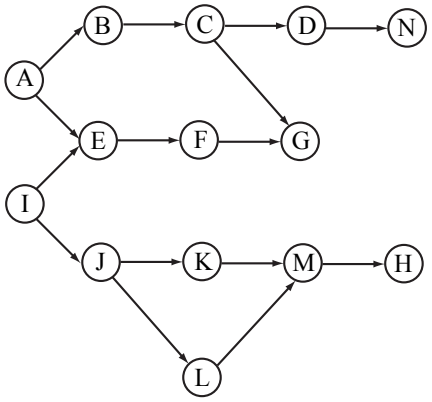
- Please also include a `README` file that describes your submission.

- You may not overwrite your submission — if you want to update a file please submit a file with a new filename, keeping in mind a 20 MB space limit per student. The file with the newest timestamp prior to the due date and time will be evaluated.

1 Bayesian Networks [50 points]

1.1 Independence [14 points]

In this question we analyze how a probabilistic graphical model encodes probabilistic dependence assumptions. Given the graphical model below, which of the following statements are true, regardless of the conditional probability distributions? **[2 points each]**



(a) $P(D, H) = P(D)P(H)$

(b) $P(A, I) = P(A)P(I)$

(c) $P(A, I|G) = P(A|G)P(I|G)$

(d) $P(J, G|F) = P(J|F)P(G|F)$

(e) $P(J, M|K, L) = P(J|K, L)P(M|K, L)$

(f) $P(E, C|A, G) = P(E|A, G)P(C|A, G)$

(g) $P(E, C|A) = P(E|A)P(C|A)$

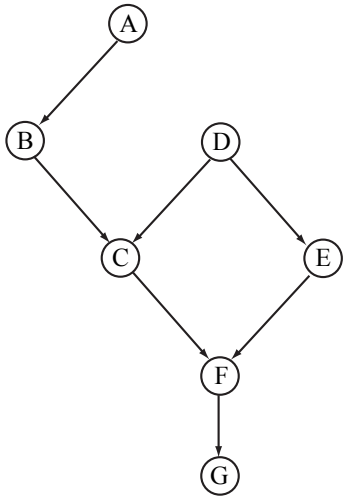
1.2 Constructing a Network [6 points]

Consider three binary variables x, y, and z with the following joint distribution:

x	y	z	$p(x, y, z)$
0	0	0	0.135
0	0	1	0.09
0	1	0	0.005
0	1	1	0.02
1	0	0	0.1125
1	0	1	0.075
1	1	0	0.1125
1	1	1	0.45

Show that the joint distribution $p(x, y, z)$ can be represented by a Bayes net that has just two edges.

1.3 Variable Elimination [20 points]



In this question we are going to practice variable elimination on the Bayesian network given above. All of the variables are binary valued $\{T, F\}$. The conditional probabilities are:

$$P(A = T) = 0.7 \quad P(D = T) = 0.6$$

$$P(B = T|A = T) = 0.9, \quad P(B = T|A = F) = 0.3$$

$$P(C = T|D = T, B = T) = 0.9, \quad P(C = T|D = T, B = F) = 0.5$$
$$P(C = T|D = F, B = T) = 0.7, \quad P(C = T|D = F, B = F) = 0.2$$

$$P(E = T|D = T) = 0.1, \quad P(E = T|D = F) = 0.4$$

$$P(F = T|E = T, C = T) = 0.2, \quad P(F = T|E = T, C = F) = 0.1$$
$$P(F = T|E = F, C = T) = 0.9, \quad P(F = T|E = F, C = F) = 0.2$$

$$P(G = T|F = T) = 0.5, \quad P(G = T|F = F) = 0.1$$

(a) **Joint Probability [3 points]** Write down the formula for the joint probability distribution that makes the same conditional independent assumptions as the above graph.

(b) **Variable Elimination [7 points]** Using variable elimination, compute the probability $P(G = T|A = T) =?$. Show your work.

(c) **Variable Elimination [10 points]** Using variable elimination, compute the probability $P(G = T|A = T, D = T) =?$. Show your work.

1.4 Admissible Bayesian Networks [10 points]

Provide an upper and lower bound on the number of possible Bayesian Networks with n nodes and explain them. Be as tight as possible (remember that a Bayesian Network is a Directed Acyclic Graph).

2 Semi-supervised Learning [25 points]

We begin by looking at the problem of Bernoulli Naive Bayes classification with one binary class variable Y and 3 binary feature variables X_1, X_2, X_3. For the Naive Bayes classifier, we would like to learn the best choice of parameters for $P(Y), P(X_1 \mid Y), P(X_2 \mid Y)$, and $P(X_3 \mid Y)$. Assume Y, $X_1 \mid Y$, $X_2 \mid Y$, and $X_3 \mid Y$ are all Bernoulli variables and let us denote the Bernoulli parameters as[1]

$$\theta_{Y=y} = P(Y = y), \quad \theta_{X_1=x_1|Y=y} = P(X_1 = x_1|Y = y),$$
$$\theta_{X_2=x_2|Y=y} = P(X_2 = x_2|Y = y), \quad \theta_{X_3=x_3|Y=y} = P(X_3 = x_3|Y = y).$$

(a) **[2 points]** Write the log-probability of X and Y in terms of the parameters θ first for a single example $(X_1 = x_1, X_2 = x_2, X_3 = x_3, Y = y)$, then for n i.i.d. examples $(X_1^i = x_1^i, X_2^i = x_2^i, X_3^i = x_3^i, Y^i = y^i)$ for $i = 1, ..., n$.

(b) **[3 points]** Next derive the maximum likelihood estimate of the parameters $\theta_Y = \arg\max_\theta \sum_{i=1}^n \log P(Y^i \mid \theta)$ and $\theta_{x_j=x_j|Y=1} = \arg\max_\theta \sum_{i=1}^n \sum_j \log P(X_j^i \mid Y^i, \theta)$.

Next, consider the case where the class value Y is never directly observed but it is approximately observed using a sensor. Let Z be the binary variable representing the sensor values. One morning you realize the sensor value is missing in some of the examples. From the sensor specifications, you learn that the probability of missing values is four times higher when $Y = 1$ than when $Y = 0$. More specifically, the exact values from the sensor specifications are:

$$P(Z \text{ missing}|Y = 1) = .08, \quad P(Z = 1|Y = 1) = .92$$
$$P(Z \text{ missing}|Y = 0) = .02, \quad P(Z = 0|Y = 0) = .98$$

[1] We only need $\theta_Y = P(Y = 1), \theta_{X_i|Y=y} = P(X_i = 1|Y = y), ...$ since $\theta_{Y=0} = 1 - \theta_{Y=1}, ...$, but the set of θs defined here should help you notationally.

(c) **[2 points]** Draw a Bayes net that represents this problem with a node Y that is the unobserved label, a node Z that is either a copy of Y or has the value "missing", and the three features X_1, X_2, X_3.

(d) **[3 points]** What is the probability of the unobserved class label being 1 given no other information, i.e., $P(Y = 1|Z = \text{"missing"})$? Derive the quantity using the Bayes rule and write your final answer in terms of $\theta_{Y=1}$, our estimate of $P(Y = 1)$.

(e) **[5 points]** Write the log-probability of X, Y and Z given θ, in terms of θ and $P(Z|Y)$, first for a single example $(X_1 = x_1, X_2 = x_2, X_3 = x_3, Z = z, Y = y)$, then for n i.i.d. examples $(X_1^i = x_1^i, X_2^i = x_2^i, X_3^i = x_3^i, Z^i = z^i, Y^i = y^i)$ for $i = 1, ..., n$.

(f) **[10 points]** Provide the E-step and M-step for performing expectation maximization of θ for this problem.

In the E-step, compute the distribution $Q_{t+1}(Y|Z, X)$ using

$$Q_{t+1}(Y = 1|Z, X) = E[Y|Z, X_1, X_2, X_3, \theta_t]$$

using your Bayes net from part (d) and conditional probability from part (e) for the unobserved class label Y of a *single* example.

In the M-step, compute

$$\theta_{t+1} = \text{argmax}_\theta \sum_{i=1}^{n} \sum_{y} Q(Y^i = y|Z^i, X^i) \log P(X_1^i, X_2^i, X_3^i, Y^i, Z^i|\theta)$$

using *all* of the examples $(X_1^1, X_2^1, X_3^1, Y^1, Z^1), ..., (X_1^n, X_2^n, X_3^n, Y^n, Z^n)$. Note: it is OK to leave your answers in terms of $Q(Y|Z, X)$.

3 K-means and GMMs [25 points]

3.1 K-Means [5 points]

Consider the data set in Figure 1. The '+' symbols indicate data points and the (centers of the) circles A, B, and C indicate the starting cluster centers. Show the results of running the K-means algorithm on this data set. To do this, use the remaining figures, and for each iteration, indicate which data points will be associated with each of the clusters, and show the locations of the updated class centers. If a cluster center has no points associated with it during the cluster update step, it will not move. Use as many figures as you need until the algorithm converges.

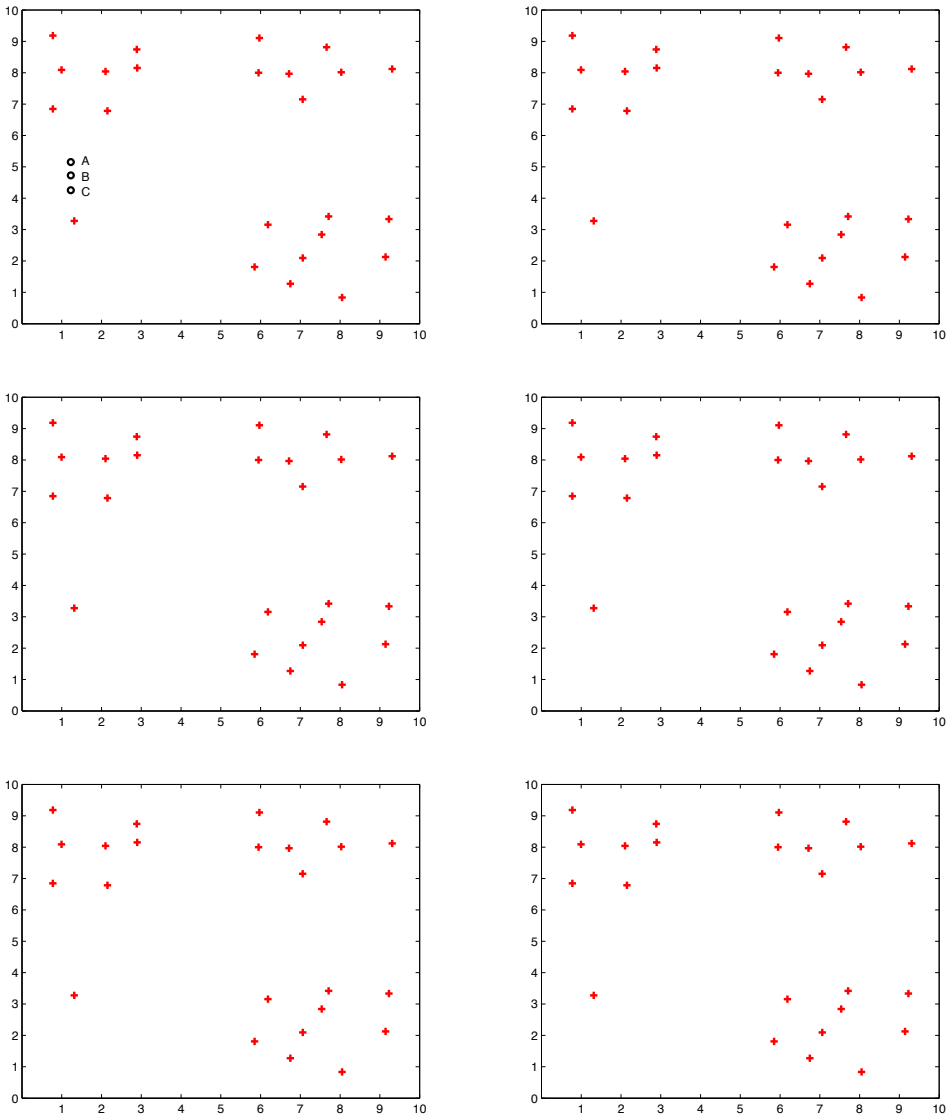


Figure 1: K-Means data set

3.2 Gaussian Mixture Models [20 points]

In this problem we will be implementing Gaussian mixture models and working with the digits data set. The provided data set is a Matlab file consisting of 5000 10×10 pixel hand written digits between 0 and 9. Each digit is a greyscale image represented as a 100 dimensional row vector (the images have been down sampled from the original 28×28 pixel images. The variable X is a 5000×100 matrix and the vector Y contains the true number for each image. Please submit your code and include in your write-up a copy of the plots that you generated for this problem.

(a) **Implementation [10 points]** Implement Expectation Maximization (EM) for the axis aligned Gaussian mixture model. Recall that the axis aligned Gaussian mixture model uses the Gaussian Naive Bayes assumption that, given the class, all features are conditionally independent Gaussians. The specific form of the model is given in Equation 1.

$$\begin{aligned} Z_i &\sim Multinomial(p_1, \ldots, p_K) \\ X_i \mid Z_i = z &\sim N\left(\begin{bmatrix} \mu_1^z \\ \vdots \\ \mu_d^z \end{bmatrix}, \begin{bmatrix} (\sigma_1^z)^2 & 0 & \ldots & 0 \\ 0 & (\sigma_2^z)^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & (\sigma_d^z)^2 \end{bmatrix} \right) \end{aligned} \tag{1}$$

Remember, code should be written and turned in *individually*.

(b) **[5 points]** Run EM to fit a Gaussian mixture model with 16 Gaussians on the digits data. Plot each of the means using `subplot`$(4, 4, i)$ to save paper.

(c) **[5 points] Evaluating Performance** Evaluating clustering performance is difficult. However, because we have information about the ground truth data, we can roughly assess clustering performance. One possible metric is to label each cluster with the majority label for that cluster using the ground truth data. Then, for each point we predict the cluster label and measure the mean 0/1 loss. For the digits data set, report your loss for settings $k = 1, 10$ and 16.