

## Homework #4

Professor: Eric Xing

Due Date: November 10, 2008

**1 Expectation Maximization (EM) [24 Points, Mark]**

The *expectation maximization* (EM) algorithm is one of the most important tools in machine learning. It allows us to create models that include *hidden (latent)* variables. When our models contain parameters that depend on these unknown variables, we cannot compute estimates of these parameters in the usual way. EM gives us a way to estimate these parameters.

We will make this more concrete by considering a simple example. Suppose I have two unfair coins. The first lands on heads with probability  $p$ , and the second lands on heads with probability  $q$ . Imagine  $N$  tosses, where for each toss I choose to use the first coin with probability  $\pi$  and choose to use the second with probability  $1 - \pi$ . The outcome of each toss  $i$  is  $x_i \in \{0, 1\}$ . Suppose I tell you the outcomes of the  $N$  tosses, for example  $x = \{x_1, x_2, \dots, x_N\}$ , but I don't tell you which coins I used on which toss.

Given only the outcomes,  $x$ , your job is to compute estimates for  $\theta$  which is the set of all parameters,  $\theta = \{p, q, \text{ and } \pi\}$ . It is pretty remarkable that this can be done at all.

To compute these estimates, we will create a *latent* variable  $Z$  where  $z_i \in \{0, 1\}$  indicates the coin used for the  $n^{\text{th}}$  toss. For example  $z_2 = 1$  indicates the first coin was used on the second toss.

We define the *incomplete* data log-likelihood as  $\log \mathbb{P}(x|\theta)$  and the *complete* data log-likelihood as  $\log \mathbb{P}(x, z|\theta)$ .

1. (3 pts) The incomplete log-likelihood of the data is given by  $\log \mathbb{P}(x|\theta) = \log \left( \sum_z \mathbb{P}(x, z|\theta) \right)$ . Use Jensen's inequality to show that a lower bound on the incomplete log-likelihood is given by:

$$\log \mathbb{P}(x|\theta) \geq \sum_z g(z) \log \left\{ \frac{\mathbb{P}(x, z|\theta)}{g(z)} \right\}$$

where  $g(z)$  is an arbitrary probability distribution over the latent variable  $Z$ .

2. (3 pts) Show that  $\log \mathbb{P}(x|\theta) = \mathcal{Q}(g(z), \theta) + KL(g(z) || \mathbb{P}(z|x, \theta))$  where

$$\begin{aligned} \mathcal{Q}(g(z), \theta) &= \sum_z g(z) \log \left\{ \frac{\mathbb{P}(x, z|\theta)}{g(z)} \right\} \\ KL(g(z) || \mathbb{P}(z|x, \theta)) &= - \sum_z g(z) \log \left\{ \frac{\mathbb{P}(z|x, \theta)}{g(z)} \right\} \end{aligned}$$

3. (3 pts) Since  $KL(g(z) || \mathbb{P}(z|x, \theta)) \geq 0$ , we see that  $\mathcal{Q}(g(z), \theta)$  is a lower bound to the incomplete data log-likelihood. To maximize this lower bound we will iteratively maximize

with respect to the probability distribution  $g(z)$  and then the parameters  $\theta$ . Let  $\theta_t$  be the parameters at iteration  $t$ . Show that the maximum with respect to  $g(z)$  (for a fixed  $\theta_t$ ) is:

$$\max_{g(z)} \mathcal{Q}(g(z), \theta_t) = \mathbb{E}_{\mathbb{P}(z|x, \theta_t)}[\log \mathbb{P}(x, z|\theta_t)] + H(\mathbb{P}(z|x, \theta_t))$$

which is just the expected value of the log-likelihood of the complete data plus the entropy of the posterior distribution of  $Z$ . This is the *expectation* part of EM.

4. (3 pts) Show that  $\mathbb{E}(z_i|x_i, \theta) = \mathbb{P}(z_i = 1|x_i, \theta)$
5. (3 pts) Use Bayes rule to compute  $\mathbb{P}(z_i = 1|x_i, \theta_t)$
6. (3 pts) Write down the complete log-likelihood  $\log \mathbb{P}(x, z|\theta)$ .
7. (3 pts) E-Step: Show that the expected log-likelihood of the complete data  $\mathcal{Q}(\theta|\theta_t) = \mathbb{E}_{\mathbb{P}(z|x, \theta_t)}[\log \mathbb{P}(x, z|\theta)]$  is given by

$$\begin{aligned} \mathcal{Q}(\theta|\theta_t) = & \sum_{i=1}^N \mathbb{E}[z_i|x_i, \theta_t][\log \pi + x_i \log p + (1 - x_i) \log(1 - p)] + \\ & (1 - \mathbb{E}[z_i|x_i, \theta_t][\log(1 - \pi) + x_i \log q + (1 - x_i) \log(1 - q)]) \end{aligned}$$

8. (3 pts) M-Step: Describe the process you would use to obtain the update equations for  $p_{t+1}, q_{t+1}, \pi_{t+1}$

## 2 Kmeans [Hanghang, 20 points]

Given  $N$  data points  $x_i, (i = 1, \dots, N)$ , Kmeans will group them into  $K$  clusters by minimizing the distortion function  $J = \sum_{n=1}^N \sum_{k=1}^K r_{n,k} \|x_n - \mu_k\|^2$ , where  $\mu_k$  is the center of the  $k^{\text{th}}$  cluster; and  $r_{n,k} = 1$  if  $x_n$  belongs to the  $k^{\text{th}}$  cluster and  $r_{n,k} = 0$  otherwise. In this exercise, we will use the following iterative procedure

- Initialize the cluster center  $\mu_k, (k = 1, \dots, K)$ ;
- Iterate until convergence
  - Update the cluster assignments for every data point  $x_n$ :  $r_{n,k} = 1$  if  $k = \operatorname{argmin}_j \|x_n - \mu_j\|^2$ ;  $r_{n,k} = 0$  otherwise.
  - Update the center for each cluster  $k$ :  $\mu_k = \frac{\sum_{n=1}^N r_{n,k} x_n}{\sum_{n=1}^N r_{n,k}}$

### (1) Convergence of Kmeans [10 pts]

Prove that the above procedure will converge in finite steps.

- *hints: consider whether or not the number of possible cluster assignments is finite.*

### (2) Kmeans and GMM [10 pts]

Remember in GMM,  $p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$ , where  $\pi_k = p(z_k = 1)$  is the prior for the  $k^{\text{th}}$  component; and  $\mu_k, \Sigma_k$  are the mean and covariance matrix for  $k^{\text{th}}$  component respectively. In the E-step, we will update  $p(z_k = 1|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^K \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}$

Now suppose that

- (1)  $\Sigma_k = \epsilon \mathbf{I}$  where  $\epsilon$  is some *given* number;
- (2)  $\pi_k \neq 0$  ( $k = 1, \dots, K$ );
- (3)  $\|x_n - \mu_i\| \neq \|x_n - \mu_j\|$  for any  $i \neq j$ .

Under the above assumptions, prove that when  $\epsilon \rightarrow 0$ ,  $p(z_k = 1|x_n) = r_{n,k}$ , where  $r_{n,k}$  is the cluster assignment used in Kmeans.

### 3 Hidden Markov Models [36 Points, Jerry]

In this problem, we are going to implement Forward algorithm, Forward-Backward algorithm, and Viterbi algorithm of Hidden Markov Models.

Consider the Dishonest Casino Problem in the lecture slides, where a fair die has an equal probability of  $1/6$  to get each of the numbers from 1 to 6, a loaded die has a probability of  $1/2$  to get 6 and  $1/10$  to get each of the rest, both of the dies have equal chance to be used initially and casino player switches back-and-forth between fair and loaded die once every 20 turns.

More formally,

$$P(1 | F) = P(2 | F) = P(3 | F) = P(4 | F) = P(5 | F) = P(6 | F) = \frac{1}{6}$$

$$P(1 | L) = P(2 | L) = P(3 | L) = P(4 | L) = P(5 | L) = \frac{1}{10}, P(6 | L) = \frac{1}{2}$$

$$P(y_1 = F) = P(y_1 = L) = \frac{1}{2}$$

$$P(y_t = F | y_{t-1} = F) = P(y_t = L | y_{t-1} = L) = 0.95$$

$$P(y_t = L | y_{t-1} = F) = P(y_t = F | y_{t-1} = L) = 0.05$$

, where  $y_t$  is the hidden state of the  $t$ -th turn.

We observe a sequence of rolls by the casino player as follows.

1245526462146146136136661664661636616366163616515615115146123562344

1. (12 pts) What is the probability of observing this sequence, given our model of how the casino works? First, fill out the Forward algorithm. Then, report your result and include the code of the Forward algorithm. (Hint: You can load the sequence from the attached data file. You are free to write codes from scratch if you don't like the provided template.)
2. (12 pts) Which rolls in the above sequence have probability of more than 0.5 to be resulted from loaded dice? You will need to implement the Forward-Backward algorithm in order to answer this question, which can be written based on the Forward algorithm. Please include the core part of the algorithm in your write-up. In addition, report the value of  $P(y_{10} = L | X)$ .

3. (12 pts) Implement the Viterbi algorithm to find the most probable sequence of fair/loaded dice that leads to the observation. Are the turns predicted to be a result of loaded dice identical to the ones in last sub-question? Report the most probable sequence, and include the code. (Hint: You can report your decoding result in the form of a sequence of letter F and L.)

## 4 ! Move to HW5 ! Bayesian Networks [20 Points, Suyash]

**Please do not submit your solution to this question in HW4.**

A graduate student takes her car to a dishonest mechanic, who claims that it requires \$1, 103 in repairs before tax. Doubting this diagnosis, the clever graduate student decides to verify it with a graphical model. Suppose that the car can have three possible problems: brake trouble (B), muffler trouble (M), and low oil (O), all of which are a priori marginally independent. The diagnosis is performed by testing for four possible symptoms: squealing noises (Sq), smoke (Sm), shaking (Sh), and engine light (Li). The conditional probabilities of these symptoms are related to the underlying causes as follows. Squealing depends only on brake problems, whereas smoke depends on brake problems and low oil. Shaking is related to brake problems and muffler problems, whereas the engine light depends only on the oil level.

1. (2 pts) Draw the graphical model for this problem.
2. (2 pts) Write the joint distribution of all the variables from the graphical model.
3. (3 pts) From this joint distribution, what can you conclude is the exact number of independent parameters that you would require to describe this joint distribution. Assume all variables are boolean.
4. (3 pts) If you were to assume no conditional independences, how many parameters would you require to specify the joint distribution assuming boolean variables?
5. (2 pts) What is the markov blanket of variable Sm ?
6. (2 pts) What is the smallest set of variables X that d-separates B and M,O?
7. (3 pts) What is an example of the “explaining away” phenomenon in the graph?
8. (3 pts) For which car problems do we gain information by observing the engine light (Li = 1)? Suppose that we see the car belching oily clouds of smoke (Sm = 1). What does seeing the engine light (Li= 1) tell us now?

**Please do not submit your solution to this question in HW4.**