# CS 228T Problem Set 3

May 6, 2011

*Instructions.* The lengths listed for each problem are suggested *maximum* lengths for typed solutions, not minimum; solving the problems fully in less space is possible. Some questions may be related to published research papers, so do not refer to any outside sources to complete this assignment, in accordance with the honor code. If you work in groups, indicate in your solutions who you worked with.

1. *Dual decomposition for pose segmentation* (17 points, 1 page). Two important problems in computer vision are that of parsing articulated objects (*e.g.*, the human body), called *pose estimation*, and segmenting the foreground and the background, called *segmentation*. Intuitively, these two problems are linked, in that solving either one would be easier if the solution to the other were available. We consider solving these problems simultaneously using a joint model over human poses and foreground/background labels and then using dual decomposition for MAP inference in this model.

   We construct a two-level model, where the high level handles pose estimation and the low level handles pixel-level background segmentation. Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected grid over the pixels. Each node $i \in \mathcal{V}$ represents a pixel. Suppose we have one binary variable $x_i$ for each pixel, where $x_i = 1$ means that pixel $i$ is in the foreground. Denote the full set of these variables by $\mathbf{x} = (x_i)$.

   In addition, suppose we have an undirected tree structure $T = (\mathcal{V}', \mathcal{E}')$ on the parts. For each body part, we have a discrete set of candidate poses that the part can be in, where each pose is characterized by parameters specifying its position and orientation. (These candidates are generated by a procedure external to the algorithm described here.) Define $y_{jk}$ to be a binary variable indicating whether body part $j \in \mathcal{V}'$ is in configuration $k$. Then the full set of part variables is given by $\mathbf{y} = (y_{jk})$, with $j \in \mathcal{V}'$ and $k = 1, \ldots, K$, where $J$ is the total number of body parts and $K$ is the number of candidate poses for each part. Note that in order to describe a valid configuration, $\mathbf{y}$ must satisfy the constraint that $\sum_{k=1}^{K} y_{jk} = 1$ for each $j$.

   Suppose we have the following energy function on pixels:

$$E_1(\mathbf{x}) = \sum_{i \in \mathcal{V}} 1[x_i = 1] \cdot \theta_i + \sum_{(i,j) \in \mathcal{E}} 1[x_i \neq x_j] \cdot \theta_{ij}.$$

   Assume that the $\theta_{ij}$ arises from a metric (*e.g.*, based on differences in pixel intensities), so this can be viewed as the energy for a pairwise metric MRF with respect to $G$.

We then have the following energy function for parts:

$$E_2(\mathbf{y}) = \sum_{p \in \mathcal{V}'} \theta_p(y_p) + \sum_{(p,q) \in \mathcal{E}'} \theta_{pq}(y_p, y_q).$$

Since each part candidate $y_{jk}$ is assumed to come with a position and orientation, we can compute a binary mask in the image plane. The mask assigns a value to each pixel, denoted by $\{w^i_{jk}\}_{i \in \mathcal{V}}$, where $w^i_{jk} = 1$ if pixel $i$ lies on the skeleton and decreases as we move away. We can use this to define an energy function relating the parts and the pixels:

$$E_3(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}'} \sum_{k=1}^{K} 1[x_i = 0, y_{jk} = 1] \cdot w^i_{jk}.$$

In other words, this energy term only penalizes the case where a part candidate is active but the pixel underneath is labeled as background.

Formulate the minimization of $E_1 + E_2 + E_3$ as an integer program and show how you can use dual decomposition to solve the dual of this integer program. Your solution should describe the decomposition into slaves, the method for solving each one, and the update rules for the overall algorithm. Briefly justify your design choices, particularly your choice of inference algorithms for the slaves.

2. *Incremental EM* (17 points). Do exercise 19.17 from the book.

3. *Variational methods for topic models* (25 points, 3 pages). Here, we examine how to use variational methods to fit a complex real-world model. *Topic models* are a popular class of Bayesian models of document corpora and other kinds of data, and we will consider a simple example of such a model; see §17.5.4 for further background and motivation.

   Consider the following model for a single document $\mathbf{w}$:

$$\begin{array}{lcll} \theta & \sim & \text{Dirichlet}(\alpha) & \\ z_n & \sim & \text{Multinomial}(\theta), & n \in \{1, 2, \ldots, N\} \\ w_n & \sim & p(w_n \,|\, z_n, \beta), & n \in \{1, 2, \ldots, N\}, \end{array}$$

   where $n$ indexes words in the document, and each document is assumed for simplicity to have a fixed number of words $N$. Each word $w_n$ is defined to be an item from a vocabulary indexed by $\{1, \ldots, V\}$, and each $w_n$ can be viewed as a binary vector with $V$ elements that has a single entry set to 1 and zeroes elsewhere (*i.e.*, an indicator representation). Each document $\mathbf{w} = (w_1, \ldots, w_N)$ has an associated multinomial distribution, specified by $\theta$, over $K$ topics, and each of the $N$ words in the document are generated by first sampling a topic $z_n$ and then sampling a word $w_n$ using the parameters $\beta$. Here, $\beta_{ij}$ is the probability that word $j$ is generated from topic $i$. The parameters $\alpha$ and $\beta$ are fixed but unknown (hyper)parameters, $\mathbf{z}$ and $\theta$ are latent variables, and $\mathbf{w}$ is observed. The joint distribution is given by

$$\begin{aligned} p(\theta, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta) &= p(\theta \,|\, \alpha) \cdot p(\mathbf{z} \,|\, \theta) \cdot p(\mathbf{w} \,|\, \mathbf{z}, \beta) \\ &= p(\theta \,|\, \alpha) \prod_{n=1}^{N} p(z_n \,|\, \theta) p(w_n \,|\, z_n, \beta). \end{aligned}$$

We are interested both in the posterior over the latent variables, since they have some semantic meaning, and in fitting the model to the data. It is natural to apply EM since it provides a solution to both problems simultaneously. Given a training corpus $\mathcal{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$, the goal is to maximize the (incomplete) log likelihood

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d \,|\, \alpha, \beta).$$

Since this objective is additive, the problem reduces to maximizing $\log p(\mathbf{w} \,|\, \alpha, \beta)$, so in the sequel we refer only to a single document $\mathbf{w}$. The main work in the E-step is in finding the posterior of the hidden variables $p(\theta, \mathbf{z} \,|\, \mathbf{w}, \alpha, \beta)$, which we would like to compute in any case. Unfortunately, the normalizer of this distribution is intractable to compute exactly, so we will use a mean field approximation in the E-step, yielding a variational EM algorithm. In particular, consider the family $\mathcal{Q}$ of approximate posteriors in the form

$$q(\theta, \mathbf{z} \,|\, \gamma, \phi) = q(\theta \,|\, \gamma) \prod_{n=1}^{N} q(z_n \,|\, \phi_n).$$

Here, $\gamma$ is called a variational Dirichlet parameter and the $\phi_n$ are variational multinomial parameters. Inference will involve computing the I-projection of the true posterior onto $\mathcal{Q}$, *i.e.*, finding the variational parameters $\gamma$ and $\phi_n$ that approximate the true posterior best.

(a) The following results will help in deriving explicit solutions for the E-step and the M-step.

   i. Give the (linear) exponential family form of the Dirichlet distribution; specify the natural parameters, the sufficient statistics, and the log partition function.

   ii. Suppose $\theta \sim \text{Dirichlet}(\alpha)$. Using your exponential family parameterization for the Dirichlet distribution, derive a simple closed-form expression for $\mathbb{E}[\log \theta_i]$ in terms of $\alpha$. (You will find that expressions in this form appear in the derivations below.)

(b) Write down the variational EM energy functional for this model. What quantity in the model does this provide a lower bound for? Expand this lower bound using the factorizations of $p$ and $q$. Expand it further in terms of the model parameters $\alpha$ and $\beta$ and the variational parameters $\gamma$ and $\phi_n$.

(c) The (variational) E-step involves maximizing the lower bound in (b) with respect to $\gamma$ and $\phi_n$. We employ an alternating maximization procedure, in which $\phi_n$ is updated to its maximizing value and then $\gamma$ is updated to its maximizing value.

   i. Derive a closed-form update for $\phi_n$ by maximizing the lower bound with respect to $\phi_n$. (Note that this is a constrained optimization problem since each $\phi_n$ is a parameter vector for a multinomial distribution.) Your answer should be in terms of $\beta$ and $\gamma$. You do not need to provide an explicit normalization constant.

   ii. Derive the first-order optimality condition for $\gamma_i$. Show that the update

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$

   satisfies the condition.

Full inference involves alternating between these two updates until the bound converges.

(d) The M-step involves maximizing the lower bound in (b) with respect to the hyperparameters $\alpha$ and $\beta$.

i. Derive a closed-form update for each $\beta_{ij}$. You do not need to provide an explicit normalization constant.

ii. Derive the first-order optimality condition for $\alpha_i$. (Since the solution will depend on the values of $\alpha_j$ for $j \neq i$, there is no closed-form expression for the maximizing value of $\alpha_i$, so this step must be carried out numerically using, *e.g.*, Newton's method.)

(e) Summarize your results by writing out *high-level* pseudocode for fitting the model to a dataset. In this part only, you should account for the fact that there are multiple documents in the training corpus. (This mostly involves adding an extra loop to the pseudocode and specifying which of the parameters above are document-specific.)

*Hints.* At various points in the derivations, you may find the following facts useful:

- The Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_K > 0$ has the density function

$$f(\theta_1, \ldots, \theta_K; \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1}$$

  and is *conjugate* to the multinomial distribution in the sense described in §17.3.2. The domain of $f$ is the $(k-1)$-probability simplex.
- Let $A(\eta)$ be the log partition function for an exponential family distribution $p$ with sufficient statistics $\tau$ and natural parameters $\eta$. If $X \sim p$, then $\nabla A(\eta) = \mathbb{E}_p[\tau(X)]$.
- The gamma function $\Gamma$ is a continuous extension of the factorial function, and has the property that $\Gamma(x) = (x-1)!$ for positive integers $x$. The first derivative of the log gamma function is the *digamma function* and is denoted by $\Psi(x)$. The derivative of the digamma function is called the *trigamma function* and is denoted $\Psi'(x)$. You can leave these functions unexpanded because efficient numerical methods are available to evaluate them.

4. *Bayesian nonparametrics* (25 points, 3 pages). In this problem, we will explore *latent feature models*, which go beyond latent class models. Latent feature models allow each training example to be represented by an *unbounded* number of latent features rather than just a single latent class variable, as in the case of mixture models. There are several situations where this idea is useful, but one way to think about it is that the infinite feature models we will discuss here allow for each example to belong to multiple clusters at once. This can be accomplished in a nonparametric Bayesian way by constructing a prior distribution on infinite binary matrices.

   Let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be the dataset. Suppose each instance $\mathbf{x} \in \mathbb{R}^n$ can be modeled in terms of latent binary features $\mathbf{f} \in \{0,1\}^K$ according to a generative model $p(\mathbf{x} \mid \mathbf{f})$. (For this problem, $p(\mathbf{x} \mid \mathbf{f})$ can be treated as a black box; its specific form does not matter.)

   Let $\mathbf{F} \in \{0,1\}^{N \times K}$ be the (latent) binary matrix obtained from placing $\mathbf{f}_i$ into the $i$th row. Here, $f_{ik} = 1$ if feature $k$ is active in example $i$. The full generative model over all instances

is then specified by the latent feature prior $p(\mathbf{F})$ and the likelihood

$$p(\mathbf{X} \mid \mathbf{F}) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \mathbf{f}_i).$$

For now, assume $K$ is finite. We specify the prior distribution over $\mathbf{F}$ by assuming each example $i$ possesses feature $k$ with probability $\pi_k$. Given hyperparameter $\alpha$, the model is

$$\begin{aligned} \pi_k \mid \alpha &\sim \text{Beta}(\alpha/K, 1) \\ f_{ik} \mid \pi_k &\sim \text{Bernoulli}(\pi_k), \end{aligned}$$

for $k = 1, \ldots, K$ and $i = 1, \ldots, N$.

*Note.* Part (c) is more difficult than the other parts, but it is possible to do all the other parts without doing it, so make sure you try (d) and (e) even if you get stuck on (c).

(a) Find an explicit expression for $p(\mathbf{F} \mid \alpha)$. Your answer should be in terms of beta or gamma functions and should contain no unevaluated integrals.

*Note.* The beta distribution with parameters $\alpha$ and $\beta$ has the density

$$f(x; \alpha, \beta) = \frac{1}{\text{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where the *beta function*

$$\text{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx$$

is the normalizing constant. The beta distribution is the Dirichlet distribution of order 2 and is *conjugate* to the Bernoulli distribution in the sense described in §17.3.2.

(b) Show how to perform Gibbs sampling in the *finite* latent feature model from part (a). Explicitly, find a simple equation for $p(f_{ik} \mid \mathbf{f}_{-ik})$, where $\mathbf{f}_{-ik}$ is the set of examples, other than $i$, with feature $k$ active.

*Notes.* It suffices to compute $p(f_{ik} \mid \mathbf{f}_{-ik})$ rather than $p(f_{ik} \mid \mathbf{F}_{-(ik)})$ because the columns of $\mathbf{F}$ are generated independently under the prior $p(\mathbf{F})$; this means we need not condition on features other than $k$.

You may find the following definitions useful. Let $m_k = \sum_{i=1}^{N} f_{ik}$ be the number of examples with feature $k$ active, and let $m_k^{(i)}$ be the number of examples, excluding example $i$, with feature $k$ active.

(c) We now move to the case of infinite features. This can be derived from an underlying stochastic process called the *Indian buffet process* (IBP), much like infinite latent class models can be derived from the Chinese restaurant process.

In the IBP, $N$ customers enter a restaurant one after another. Each encounters a buffet consisting of infinitely many dishes in a line. The first starts at the left and takes a serving from each dish, stopping after a Poisson($\alpha$) number of dishes. The $i$th customer

moves along the buffet and samples dishes in proportion to their popularity, *i.e.*, he tries a dish with probability $m_k/i$, where $m_k$ is the number of previous customers who have sampled a dish. Having reached the end of all previously sampled dishes, the $i$th customer then tries a Poisson($\alpha/i$) number of new dishes. The entries $f_{ik}$ in our (now infinite) matrix $\mathbf{F}$ indicate whether the $i$th customer tried the $k$th dish.

If $\mathbf{F} \sim \text{IBP}(\alpha)$, prove that

$$p(\mathbf{F} \mid \alpha) = \alpha^T \cdot \exp(-\alpha H_N) \cdot \prod_{i=1}^{N} \frac{1}{K_1^{(i)}!} \cdot \prod_{k=1}^{T} \frac{(m_k - 1)!(N - m_k)!}{N!},$$

where $H_N = \sum_{i=1}^{N}(1/i)$ is the $N$th harmonic number, $T$ is the total number of dishes sampled by the $N$ customers, $K_1^{(i)}$ be the number of new dishes sampled by the $i$th customer, and $m_k$ is the number of customers who tried dish $k$.

*Hint.* Think about all customers at once and regroup terms. As above, let $m_k^{(i)}$ be the number of customers, excluding $i$, who have sampled $k$. In the algebra, it may be useful to take $m_k^{(i)}$ to be 1 if $i$ is the first to sample $k$.

(d) In latent class models, the particular values of the class variables did not have any particular meaning (*i.e.*, they were exchangeable). In this case, the features do not have any predetermined meaning, so we want to treat a matrix $\mathbf{F}'$ which simply reorders the columns of $\mathbf{F}$ as equivalent to $\mathbf{F}$, *i.e.*, $p(\mathbf{F} \mid \alpha) = p(\mathbf{F}' \mid \alpha)$ if $\mathbf{F}$ and $\mathbf{F}'$ are the same up to permuting columns. This will play the same role for binary matrices here as partitions did for assignment vectors in the Dirichlet process mixture model.

Let lof($\cdot$) be a function that maps a binary matrix to a *left-ordered* binary matrix, in which the columns are ordered from left to right in decreasing order of the magnitude of the binary number expressed by that column. Let $[\mathbf{F}]$ be the equivalence class of $\mathbf{F}$ under this relation, *i.e.*, $[\mathbf{F}]$ is the set of all binary matrices that map to the same left-ordered matrix as $\mathbf{F}$. The number of matrices in any given $[\mathbf{F}]$ is

$$\frac{\prod_{i=1}^{N} K_1^{(i)}!}{\prod_{h=0}^{2^N-1} K_h!},$$

where $K_h$ is the number of columns with identical entries $h$.

Using this fact and part (c), find a formula for $p([\mathbf{F}] \mid \alpha)$. Make sure your expression does not depend on the customer ordering; this implies that the distribution is exchangeable.

(e) Show how to perform Gibbs sampling for the infinite model you defined in part (d). What is its relation to the expression you found in part (b)?

*Hint.* Use exchangeability.

5. *Sampling methods for the marginal likelihood* (16 points, 2 pages). This problem considers two simple Monte Carlo methods for estimating the marginal likelihood, which is the basis for Bayesian model selection. Note that neither method is ideal, but several better methods based on these ideas are in common use.

Let $p(D \mid \theta)$ be the data likelihood and let $p(\theta)$ be the prior over the model parameters. Assume that $p(D \mid \theta)$ can be evaluated reasonably efficiently.

(a) Consider a normalized importance sampling estimator for the marginal likelihood, in which we generate $M$ parameter samples $\theta_m$ from some proposal distribution $q(\theta)$ and then estimate the marginal likelihood as

$$\hat{p}(D) = \frac{\sum_{m=1}^{M} w_m p(D \mid \theta_m)}{\sum_{m=1}^{M} w_m},$$

where $w_m = p(\theta_m)/q(\theta_m)$. Suppose we sample a set of $\theta_m$ from the posterior $p(\theta \mid D)$ (possibly with MCMC), so the proposal $q(\theta)$ is $p(\theta \mid D)$.

i. Simplify the form of the estimator to derive a formula for the marginal likelihood that uses only quantities that can be computed efficiently. Your expression should take the form of a kind of average of terms that are easy to compute.

ii. Give a reason why we would sample from the posterior rather than the prior, even though the prior is generally easier to sample from.

iii. Describe some serious drawbacks of this method when posterior samples are used.

*Hint.* Think about the relationship among the prior, the data, the posterior, and the marginal likelihood.

(b) Let

$$p_t(\theta \mid D) = (1/Z_t) p(D \mid \theta)^t p(\theta),$$

where

$$Z_t = \int_{\Theta} p(D \mid \theta)^t p(\theta) \, d\theta.$$

The parameter $t$ can be viewed as a temperature.

i. Show that

$$\log p(D) = \int_0^1 \mathbb{E}_{p_t}[\log p(D \mid \theta)] \, dt.$$

*Hint.* Consider $d/dt \log Z_t$.

ii. Describe how you would approximate the integral above by evaluating its value at several discrete values of $t$, *i.e.*, $0 = t_1 < t_2 < \cdots < t_k = 1$. In particular, describe how to apply one of the MCMC techniques discussed previously in the course to estimate the quantities necessary for approximating the integral. The method should be more efficient than running $k$ independent chains.