STAT 241: HOMEWORK #4

MIKE SCHACHTER

Problem 1. Consider data where $N = 196$ individuals are distributed multinomially into four categories, giving data $y = \{120, 16, 22, 38\}$. Let the model for this data be a multinomial distribution with $\{\frac{1}{2} + \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi\}$, for $\pi \in (0, 1)$. Let the complete data be $x = (x_1, x_2, x_3, x_4, x_5)$, where $y_1 = x_1 + x_2$, $y_2 = x_3$, $y_3 = x_4$, and $y_4 = x_5$. Use EM to solve for π and run the algorithm for 10 steps.

I adapted an answer provided in section 1.4.2 of McLachlan 2008, "The EM Algorithm and Extensions". First we'll assume that the complete data is distributed multinomially as $\theta = \{\frac{1}{2}, \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi\}$. The first step of the problem is to write out the likelihood of the complete data x:

$$L_c(x|\pi) = \frac{N!}{\prod_{k=1}^{5} x_k!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{1}{4}\pi\right)^{x_2} \left(\frac{1}{4}(1-\pi)\right)^{x_3} \left(\frac{1}{4}(1-\pi)\right)^{x_4} \left(\frac{1}{4}\pi\right)^{x_5}$$

Then take the log:

$$\begin{aligned} l_c(x|\pi) &= log\, N! - \sum_{i=1}^{5} log\, x_i! + x_1 log\, \frac{1}{2} + (x_2 + x_5) log\, \pi + (x_3 + x_4) log\, (1-\pi) \\ &- (x_2 + x_5 + x_3 + x_4) log\, 4 \end{aligned}$$

Because in the M-step we're going to differentiate with respect to π, we'll define a function on everything that's not a function of π:

$$g(x) = log\, N! - \sum_{i=1}^{5} log\, x_i! + x_1 log\, \frac{1}{2} - (x_2 + x_5 + x_3 + x_4) log\, 4$$

So the complete log likelihood becomes:

$$l_c(x|\pi) = g(x) + (x_2 + x_5) log\, \pi + (x_3 + x_4) log\, (1-\pi)$$

The latent variables in the complete log likelihood are x_1 and x_2. In the E-step, we take the expectation of the complete log likelihood with respect to these latent variables, assuming a set of initialized parameters θ^t and incomplete data $y = (y_1, y_2, y_3, y_4)$:

$$\mathbb{E}\left[l_c(x|\pi) \,|\, \theta^t, y\right] = \mathbb{E}\left[g(x) \,|\, \theta^t, y\right] + (\mathbb{E}\left[x_2 \,|\, \theta^t, y\right] + y_4) log\, \pi + (y_2 + y_3) log\, (1-\pi)$$

Where relevant, x_i was replaced with y_i. We're going to ignore $\mathbb{E}\left[g(x) \,|\, \theta^t, y\right]$, because it's not a function of π, it's a function of π^t, which is constant. It will be differentiated out in the M-step. But we do need to evaluate $\mathbb{E}\left[x_2 \,|\, \theta^t, y\right]$. In order to do so, note that $y_1 = x_1 + x_2$. The total weight of x_1 and x_2 is $\frac{1}{2} + \frac{1}{4}\pi$, so the probability of x_2 is $\frac{1}{2}(\frac{1}{2} + \frac{1}{4}\pi)^{-1}$, and the conditional expectation is:

$$\mathbb{E}\left[x_2 \,|\, \theta^t, y\right] = y_1 \frac{1}{2}(\frac{1}{2} + \frac{1}{4}\pi^t)^{-1}$$

Let $\bar{x}_2^t = \mathbb{E}\left[x_2 \,|\, \theta^t, y\right]$, the expected complete log likelihood becomes:

$$\mathbb{E}\left[l_c(x|\pi) \,|\, \theta^t, y\right] = \mathbb{E}\left[g(x) \,|\, \theta^t, y\right] + (\bar{x}_2^t + y_4) log\, \pi + (y_2 + y_3) log\,(1-\pi)$$

For the M-step we differentiate this with respect to π:

$$\begin{aligned} \frac{\partial}{\partial \pi}\left[\mathbb{E}\left[l_c(x|\pi) \,|\, \theta^t, y\right]\right] &= \frac{\partial}{\partial \pi}\left[(\bar{x}_2^t + y_4) log\, \pi + (y_2 + y_3) log\,(1-\pi)\right] \\ &= \frac{1}{\pi}(\bar{x}_2^t + y_4) - \frac{1}{1-\pi}(y_2 + y_3) \end{aligned}$$

Then we set this to zero, and solve for π:

$$\begin{aligned} \frac{1}{\pi}(\bar{x}_2^t + y_4) &= \frac{1}{1-\pi}(y_2 + y_3) \\ &\cdots \\ \pi &= \frac{\bar{x}_2^t + y_4}{x_2 + y_4 + y_2 + y_3} \end{aligned}$$

So now we have our E and M steps:

E-step:

$$\bar{x}_2^t = y_1 \frac{1}{2}(\frac{1}{2} + \frac{1}{4}\pi^t)^{-1}$$

M-step:

$$\pi^{t+1} = \frac{\bar{x}_2^t + y_4}{x_2 + y_4 + y_2 + y_3}$$

The first 10 iterations of the algorithm, with initial guess $\pi^0 = 0.1$, are:

(1) x2=1.00, old_pi=5.714, new_pi=0.100
(2) x2=2.00, old_pi=25.324, new_pi=0.535
(3) x2=3.00, old_pi=28.570, new_pi=0.625
(4) x2=4.00, old_pi=28.974, new_pi=0.637
(5) x2=5.00, old_pi=29.022, new_pi=0.638
(6) x2=6.00, old_pi=29.028, new_pi=0.638
(7) x2=7.00, old_pi=29.029, new_pi=0.638
(8) x2=8.00, old_pi=29.029, new_pi=0.638
(9) x2=9.00, old_pi=29.029, new_pi=0.638
(10) x2=10.00, old_pi=29.029, new_pi=0.638

Problem 2. Consider an undirected graphical model pairwise factorization of the form:

$$p(x_1, ..., x_d; \psi) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t)$$

(1) Compute ML estimates $\{\psi_s, \psi_{st}\}$ for (i) a tree-structured graph with $E = \{(1,2),(2,3),(3,4)\}$ and (ii) the fully connected graph with all $\begin{pmatrix} 4 \\ 2 \end{pmatrix}$ edges. Which graph gives the higher likelihood? Is higher likelihood better?

See the "problem2.4a.txt" file attached with this document for documentation on the IPF code and clique potentials computed with the tree and full models. The likelihood for the tree model was -79.9, the likelihood for the full model was -77.9. Although the full model has a higher likelihood, it's not necessarily better. It might be overfitting the data. It might be a good idea to try graphs of intermediate complexity and see how their likelihoods compare.

(2) For graph (i) in (a), show that ML estimates $\{\hat{\psi}_s, \hat{\psi}_{st}\}$ can be written in closed form as $\hat{\psi}_s(x_s) = \bar{\mu}_s(x_s)$ and $\hat{\psi}_{st}(x_s x_t) = \frac{\bar{\mu}_{st}(x_s,x_t)}{\bar{\mu}_s(x_s)\bar{\mu}_t(x_t)}$. Does the fully connected graph have the same closed-form solution?

Tried this one for a bit, ran into problems with partition function...

(3) Assume we know that the node compatibility functions are constant; $\psi_s(x_s) = 1$ for all $s \in V$, and $\psi_{st} = \psi_{uv}$ for all $(s,t),(u,v) \in E$. Describe a modified IPF algorithm for computing ML estimates.

The update for IPF is:

$$\psi_c^{t+1}(x_c) = \psi_c^t(x_c)\frac{\bar{p}(x_c)}{p^t(x_c)}$$

In this case, we want the marginal to ultimately be $p^t(x_c) = k$, a constant. Say the initial values of each edge potential are constant. Then $p^t(x_{st}) = p^t(x_{uv})$ for all $(s,t),(u,v) \in E$. However, if the data is noisy, the empirical marginals $\bar{p}(x_c)$ will not be equal for all cliques, there will be some $\bar{p}(x_{st}) \neq \bar{p}(x_{uv})$. So we can't do that. But if we initialize each $\psi_c^0(x_c)$ so that:

$$\frac{\bar{p}(x_{st})}{p^0(x_{st})} = \frac{\bar{p}(x_{uv})}{p^0(x_{uv})} \quad \forall (s,t),(u,v) \in E$$

then each update will keep the potentials equal to eachother.

(4) Assume the graph T is a tree, but the edge set is unknown. Let $\hat{\psi}(T)$ be the ML estimate of the compatibility functions for all verticies and edges. Let $l(\hat{\psi}(T))$ be the maximized log-likelihood for T, and choose the best tree as:

$$T^* \in argmax_T l(\hat{\psi}(T))$$

Show that any tree T^* must be a maximum weight spanning tree, in the sense that:

$$\sum_{(s,t)\in E(T^*)} D(\bar{\mu}_{st} \parallel \bar{\mu}_s\bar{\mu}_t) \geq \sum_{(s,t)\in E(T)} D(\bar{\mu}_{st} \parallel \bar{\mu}_s\bar{\mu}_t)$$

The log-likelihood function for a tree is:

$$l(\psi) = \sum_{n=1}^{N} \sum_{s\in V_T} \psi_s(x_s^n) \sum_{(u,v)\in V_T} \psi_{uv}(x_u^n, x_v^n) - NlogZ$$

Assume equal initial values for the single node potentials across trees. The IPF update for edge potentials is:

$$\psi_{uv}^{t+1}(x_u, x_v) = \psi_{uv}^t(x_u, x_v)\frac{\bar{p}(x_u, x_v)}{p^t(x_u, x_v)}$$

where $\bar{p}(x_u, x_v) = \bar{u}_{uv}$. Edges with high empirical marginals will wind up with greater values for the final potential function, and contribute more to the log likelihood. Because we assumed single node potentials were

initialized at the same values, and because they have the same empirical marginals regardless of the tree, the KL distance for the optimal tree T^* between $\bar{u}_{uv}$ and $\bar{u}_u$ and $\bar{u}_v$ for a given edge will on average be higher than a sub-optimal tree, so that the overall sum of KL distances will be higher for T^* than any sub-optimal tree.

Problem 3. For each function $A(\theta)$, compute the conjugate dual $A^*(\mu)$ and specify an example of an exponential family for which $A(\theta)$ is the cumulant function. Show how the conjuate dual is related to the entropy $H(p) = -\int_{\mathcal{X}} p(x;\theta) log\, p(x;\theta)\, dx$.

Note: I adopted notation from Section 3 of Wainwright and Jordan (2008) instead of $f(u)$ because I think it helps the intuition behind the problem.

(1) $A(\theta) = log(1 + exp(\theta))$. This is the cumulant of the Bernoulli distribution. To see this, let's rewrite the density function as an exponential family:

$$\begin{aligned} p(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ &= exp\{log(\mu^x(1-\mu)^{1-x})\} \\ &= exp\{xlog(\mu) + (1-x)log(1-\mu)\} \\ &= exp\{xlog\left(\frac{\mu}{1-\mu}\right) + log(1-\mu)\} \end{aligned}$$

From this we let the canonical parameter $\theta = log\left(\frac{\mu}{1-\mu}\right)$. If we invert this relationship we get $\mu = (1+e^{-\theta})^{-1} = \sigma(\theta)$, the logistic sigmoid, and the cumulant function becomes:

$$\begin{aligned} -A(\theta) &= log(1-\mu) \\ &= log(1-\sigma(\theta)) \\ &= -log(1+e^{\theta}) \end{aligned}$$

The definition of conjugate dual is:

$$A^*(\mu) = sup_\theta\{\mu\theta - A(\theta)\}$$

We maximize the expression in the sup by taking the derivative with respect to θ:

$$\frac{\partial}{\partial\theta}\left[\mu\theta - A(\theta)\right] = \theta - \sigma(\theta)$$

We then set this expression to zero, and get $\theta^* = log\left(\frac{\mu}{1-\mu}\right)$. Then we plug this maximum into the expression for the conjugate dual to get the final form:

$$\begin{aligned} A^*(\mu) &= \mu log\left(\frac{\mu}{1-\mu}\right) - log(1 + \frac{\mu}{1-\mu}) \\ &\ldots \\ &= \mu log\mu + (1-\mu)log(1-\mu) \end{aligned}$$

The expression for the conjugate dual is the same as the negative entropy for the Bernoulli distribution.

(2) $A(\theta) = -log(\theta)$. This is the cumulant of the exponential distribution. To see this:

$$\begin{aligned} p(x|\lambda) &= \lambda e^{-\lambda x} \\ &= exp\{-\lambda x + log\lambda\} \end{aligned}$$

The canonical parameter is $\theta = -\lambda$, and the cumulant becomes $A(\theta) = -log\lambda = -log(-\theta)$. We then differentiate the conjugate dual expression with respect to θ:

$$\frac{\partial}{\partial\theta}\left[\mu\theta - A(\theta)\right] = \lambda + \frac{1}{\theta}$$

From this we get $\theta^* = -\frac{1}{\lambda}$, which we then plug back in to the conjugate dual expression to get:

$$\begin{aligned} A^*(\lambda) &= \lambda\left(-\frac{1}{\lambda}\right) - A\left(\frac{1}{\lambda}\right) \\ &= -1 + log\left(\frac{1}{\lambda}\right) \\ &= -1 - log\lambda \end{aligned}$$

The entropy of the exponential distribution is $H(\lambda) = 1 - log\lambda$, at least according to wikipedia. Sorry for not doing the integral... So we have some weirdness, where the conjugate dual is not quite equal to the negative entropy. Either wikipedia is wrong, or I'm wrong (as well as Wainwright and Jordan, see table 3.2)....

(3) $A(\theta) = \frac{1}{2}\theta^T\Sigma^{-1}\theta$. This is going to be messy and probably incorrect... This cumulant function comes from a multivariate Gaussian with a fixed covariance matrix. To see this, we can turn the density for the MVG into an exponential family:

$$\begin{aligned} p(x|\mu,\Sigma) &= \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\} \\ &= \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu)\} \\ &= exp\{-\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu - log(2\pi^{d/2}|\Sigma|^{1/2})\} \\ &= h(x)exp\{x^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu\} \end{aligned}$$

Let $\theta = \mu$. Then the cumulant function is $A(\theta) = \frac{1}{2}\mu^T\Sigma^{-1}\mu$. The derivative of the conjugate dual expression is:

$$\frac{\partial}{\partial\theta}\left[\mu\theta - A(\theta)\right] = \mu - \Sigma^{-1}\theta$$

Setting this to zero gives $\theta^* = \Sigma\mu$. We plug this back in to get the conjugate dual expression:

$$A^*(\mu) = \mu^T\Sigma\mu - \frac{1}{2}\mu^T\Sigma\mu = \frac{1}{2}\mu^T\Sigma\mu$$

Can't say this looks anything like the entropy of a multivariate Gaussian though... I'm going to have to wait and see the answer when the solutions come out.