Final Exam

Matthew C. Scicluna
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
Montréal, QC H3T 1J4
matthew.scicluna@umontreal.ca

December 16, 2017

1 Short Problems

1.1 Maximum Entropy Principle

We find the Maximum Entropy distribution on the set $\mathbb{N}$ such that $E(X) = \alpha$. We claim that this is the Geometric Distribution $p(k) = \left(\frac{\alpha}{1+\alpha}\right)^k \frac{1}{1+\alpha}$
PROOF: We want to find the distribution which maximizes the entropy $H(p)$ satisfying the constraints $\mathbb{E}(X) = \alpha$ and $\sum_{i=0}^{\infty} p(i) = 1$. We form the Lagrangian:

$$L(p, \nu, C) = -H(p) + \nu\left(\sum_{i=0}^{\infty} ip(i) - \alpha\right) + C\left(\sum_{i=0}^{\infty} p(i) - 1\right)$$

Taking the derivative w.r.t. $p(k)$ we get:

$$\frac{\partial}{\partial p(k)} L(p, \nu, C) = -\log p(k) - 1 + k\nu + C \tag{1.1}$$

$$\Rightarrow p(k) = \exp\{k\nu\}\exp\{C - 1\} \tag{1.2}$$

And using that $\sum_{i=0}^{\infty} p(i) = 1$ we have that

$$\sum_{i=0}^{\infty} \exp\{i\nu\}\exp\{C-1\} = 1 \Rightarrow \exp\{-C+1\} = \sum_{i=0}^{\infty} \exp\{i\nu\} \tag{1.3}$$

we substitute (1.3) into (1.1) to eliminate C

$$p(k) = \frac{\exp\{k\nu\}}{\sum_{i=0}^{\infty} \exp\{i\nu\}} \tag{1.4}$$

We then solve for α

$$\begin{aligned}
\mathbb{E}(X) = \sum_{k=0}^{\infty} \frac{k \exp\{k\nu\}}{\sum_{i=0}^{\infty} \exp\{i\nu\}} &= \alpha \\
\Rightarrow \sum_{k=0}^{\infty} k \exp\{k\nu\} &= \alpha \sum_{i=0}^{\infty} \exp\{i\nu\} \\
\overset{(a)}{\Rightarrow} \frac{\exp\{\nu\}}{(1-\exp\{\nu\})^2} &= \frac{\alpha}{(1-\exp\{\nu\})} \\
\Rightarrow \exp\{\nu\} &= \frac{\alpha}{1+\alpha}
\end{aligned}$$

Where (a) comes from the geometric series. Finally, we sub this value into (1.4) to get the familiar formula:

$$p(k) = \left(\frac{\alpha}{1+\alpha}\right)^k \frac{1}{1+\alpha} \tag{1.5}$$

1.2 Sampling

We want to sample from $X \mid ||X-y||_2 \leq 1$ with $X \sim \mathcal{N}(0, I_d)$. We propose the following scheme: Sample X from $\mathcal{N}(0, I_d)$ using some method like the Box Muller transform. Check if $||X-y||_2 \leq 1$, and if so, return X. Otherwise, toss the sample and repeat until an X is sampled successfully.
The scheme works because $p(X|Accept) = p(X|\ ||X-y||_2 \leq 1)$, as needed.

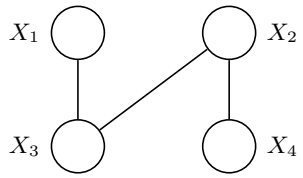
2 Factorization and Markov properties

2.1 UGM from constraints

We find the unique undirected graphical model on four random variables X_1, X_2, X_3, X_4 which satisfy:

1. All distributions satisfy: $X_1 \perp X_2 | (X_3, X_4)$, $X_3 \perp X_4 | (X_1, X_2)$, $X_1 \perp X_4 | X_3$
2. There exist distributions which satisfy: $X_1 \not\perp X_3 | X_2$, $X_2 \not\perp X_4 | X_1$, $X_1 \not\perp X_4$

The graph which satisfies this is:



2.2 DGM

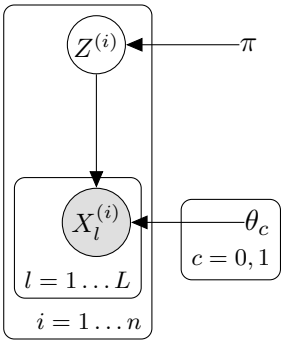
Given the two directed graphical models, it is possible to have a distribution p on $X_1, \cdots, X_{16}$ such that $p \in \mathcal{L}(G_1)$ and $p \in \mathcal{L}(G_2)$. Namely, take

$$p(x_1, \cdots, x_{16}) = \prod_{i=1}^{16} p(x_i) \tag{2.1}$$

This is trivially in $\mathcal{L}(G_1)$ and $\mathcal{L}(G_2)$, as needed.

3 EM algorithm

Let $X \in \{0,1\}^L$. Let $Z \in \{0,1\}$ be a Bernouilli latent variable with $p(z=1) = \pi$. Our NB model is $p(X_l = 1|Z=1) = \theta_1$ and $p(X_l = 1|Z=0) = \theta_0$, where $X = (X_1, \cdots, X_L)$ and the X_l's are conditionally independent given θ_z. Suppose that we have n i.i.d. observations $\{X^{(i)}\}_{i=1}^n$ from this model. We want to estimate parameters $(\pi, \theta_0, \theta_1)$ by maximum likelihood via the EM algorithm.
The full graphical model is as follows:



We derive the EM algorithm to estimate $\theta = (\pi, \theta_0, \theta_1)$. Let $\tau_i = p(Z^{(i)} = 1 \mid X^{(i)})$, $\tau = \sum_{i=1}^n \tau_i$ and $c_i = \sum_{l=1}^L X_l^{(i)}$. For the E-Step, given θ_0, θ_1, π from the previous

iteration, we would compute τ_i using the following formula:

$$\tau_i = p(Z^{(i)} = 1 \mid X^{(i)}) \tag{3.1}$$

$$= \frac{p(X^{(i)} \mid Z^{(i)} = 1)p(Z^{(i)} = 1)}{\sum_{c=0}^{1} p(X^{(i)} \mid Z^{(i)} = c)p(Z^{(i)} = c)} \tag{3.2}$$

$$= \frac{\prod_{l=1}^{L} p(X_l^{(i)} \mid Z^{(i)} = 1)p(Z^{(i)} = 1)}{\sum_{c=0}^{1} \prod_{l=1}^{L} p(X_l^{(i)} \mid Z^{(i)} = c)p(Z^{(i)} = c)} \tag{3.3}$$

$$= \frac{\pi\theta_1^{c_i}(1-\theta_1)^{L-c_i}}{\pi\theta_1^{c_i}(1-\theta_1)^{L-c_i} + (1-\pi)\theta_0^{c_i}(1-\theta_0)^{L-c_i}} \tag{3.4}$$

$$= \frac{1}{1 + \frac{(1-\pi)\theta_0^{c_i}(1-\theta_0)^{L-c_i}}{\pi\theta_1^{c_i}(1-\theta_1)^{L-c_i}}} \tag{3.5}$$

$$= \sigma\left(\log \frac{\pi\theta_1^{c_i}(1-\theta_1)^{L-c_i}}{(1-\pi)\theta_0^{c_i}(1-\theta_0)^{L-c_i}}\right) \tag{3.6}$$

For the M-Step, we compute θ_0, θ_1, π via maximum likelihood using the τ_i's from the previous step. To simplify notation the computations we introduce an index c to indicate class i.e. $\tau_{ic} = p(Z^{(i)} = c \mid X^{(i)})$ and $\pi_c = p(Z^{(i)} = c)$ and proceed with the understanding that $\tau_{i1} = \tau_i$ and $\pi_{i1} = \pi_i$. The objective function to optimize is as follows:

$$\mathbb{E}\left(\ln p(X, Z, \theta)\right) = \sum_{i=1}^{n}\sum_{c=0}^{1} \tau_{ic}[\ln \pi_c + c_i \ln \theta_c + (L - c_i)\ln(1-\theta_c)] \tag{3.7}$$

To find the maximum w.r.t. $\pi = \pi_1$ we take derivative of (3.7) with the following term added to maintain the constraint: $\lambda(\sum_{c=0}^{1} \pi_c - 1)$. This gives us:

$$\frac{\partial}{\partial \pi_1}\mathbb{E}\left(\ln p(X, Z, \theta)\right) = \sum_{i=1}^{n} \frac{\tau_{i1}}{\pi_1} - \lambda = 0 \tag{3.8}$$

$$\Rightarrow \pi_1 = \frac{\sum_{i=1}^{n} \tau_{i1}}{\lambda} \tag{3.9}$$

And satisfying the constraint gives us: $\lambda = \sum_{c=0}^{1}\sum_{i=1}^{n} \tau_{ic} = \sum_{i=1}^{n} 1 = n$
Finally

$$\pi = \frac{\sum_{i=1}^{n} \tau_i}{n} = \frac{\tau}{n} \tag{3.10}$$

To find the maximum w.r.t. θ_c we take derivative of (3.7). This gives us:

$$\frac{\partial}{\partial \theta_c}\mathbb{E}\left(\ln p(X, Z, \theta)\right) = \sum_{i=1}^{n} \tau_{ic}\left[\frac{c_i}{\theta_c} - \frac{L - c_i}{1 - \theta_c}\right] = 0 \tag{3.11}$$

$$\Rightarrow \sum_{i=1}^{n} c_i\tau_{ic} = \sum_{i=1}^{n} L\theta_c\tau_{ic} \tag{3.12}$$

$$\Rightarrow \theta_c = \frac{\sum_{i=1}^{n} c_i\tau_{ic}}{L\sum_{i=1}^{n} \tau_{ic}} \tag{3.13}$$

Finally, after some algebra:

$$\theta_0 = \frac{\sum_{i=1}^{n}(1-\tau_i)c_i}{(n-\tau)L} \tag{3.14}$$

$$\theta_1 = \frac{\sum_{i=1}^{n}\tau_i c_i}{\tau L} \tag{3.15}$$

4 Parallel Chains

We are given a graph and variables X_s and Y_s which are discrete K-valued for $s = 1, \cdots, S$, and Z is a binary random variable. We are given the following factors for a specific distribution $p \in \mathcal{L}(G)$:

1. $p(X_1 = l) = p(Y_1 = l) = \pi_l$ for $l = 1, \cdots, K$
2. $p(X_s = i | X_{s-1} = j) = p(Y_s = i | Y_{s-1} = j) = A_{ij}$ for $i, j = 1, \cdots, K$ and $s = 2, \cdots, S$
3. $p(Z = 1 | X_s = k, Y_s = l) = \begin{cases} p & \text{if } k = l \\ q & \text{o.w.} \end{cases}$

We have that

$$p(Z = 1) = \sum_{i,j} P \circ (A^{S-1}\pi)(A^{S-1}\pi)^T \tag{4.1}$$

Where
$P \in \mathbb{R}^{K \times K}$ and $P_{ij} = \begin{cases} p & i = j \\ q & i \neq j \end{cases}$

We first prove that $p(X_s = i) = [A^{S-1}\pi]_i$. For $s = 2$,

$$\begin{aligned} p(X_2 = i) &= \sum_{j=1}^{K} p(X_2 = i | X_1 = j) p(X_1 = j) \\ &= \sum_{j=1}^{K} A_{ij}\pi_j = [A\pi]_i \end{aligned}$$

Then, by induction,

$$\begin{aligned} p(X_S = i) &= \sum_{j=1}^{K} p(X_S = i | X_{S-1} = j) p(X_{S-1} = j) \\ &= \sum_{j=1}^{K} A_{ij} A_{ij}^{S-2}\pi \\ &= \sum_{j=1}^{K} A_{ij}^{S-1}\pi = [A^{S-1}\pi]_i \end{aligned}$$

Putting it all together:

$$\begin{aligned} p(Z=1) &= \sum_{i=1}^{k}\sum_{i=1}^{j} p(Z=1, X_s=i, Y_s=j) \\ &= \sum_{i=1}^{k}\sum_{i=1}^{j} p(Z=1|X_s=i, Y_s=j)p(X_s=i)p(Y_s=j) \\ &= \sum_{i=1}^{k}\sum_{i=1}^{j} P \circ (A^{S-1}\pi)(A^{S-1}\pi)^T \end{aligned}$$

5 Metropolized Gibbs sampler

Let p be a strictly positive distribution of the random variable $X = (X_1, \cdots, X_n)$, with X_i are K-valued random variables, with $K \geq 3$ and $p_i(z_i|x^t_{-i}) := P(X_i = z_i|X_{-i} = x^t_{-i})$. We have that the Metropolized Gibbs sampler makes a proposal drawn from the transition q_i with

$$q_i((z_i, x^t_{-i})|(x^t_i, x^t_{-i})) := \begin{cases} \frac{p_i(z_i|x^t_{-i})}{1-p_i(x^t_i|x^t_{-i})} & \text{for } z_i \neq x^t_i \\ 0 & \text{for } z_i = x^t_i \end{cases} \tag{5.1}$$

1. We find the acceptance probability $\alpha_i((z_i, x^t_i)|(x^t_i, x^t_{-i}))$ that arises from this proposal distribution q_i in a Metropolis-Hasting algorithm for the target distribution p. For $z_i \neq x^t_i$ we have that:

$$\begin{aligned} \frac{q_i((x^t_i, x^t_{-i})|(z_i, x^t_{-i}))p(z_i, x^t_{-i})}{q_i((z_i, x^t_{-i})|(x^t_i, x^t_{-i}))p(x^t_i, x^t_{-i})} &= \frac{p_i(x^t_i|x^t_{-i})p(z_i, x^t_{-i})}{1-p_i(z_t|x^t_{-i})} \frac{1-p_i(x^t_i|x^t_{-i})}{p_i(z_i|x^t_{-i})p(x^t_i, x^t_{-i})} \\ &= \frac{1-p_i(x^t_i|x^t_{-i})}{1-p_i(z_t|x^t_{-i})} \frac{p_i(x^t_i|x^t_{-i})p(z_i|x^t_{-i})p(x^t_{-i})}{p_i(z_i|x^t_{-i})p(x^t_i|x^t_{-i})p(x^t_{-i})} \\ &= \frac{1-p_i(x^t_i|x^t_{-i})}{1-p_i(z_t|x^t_{-i})} \end{aligned}$$

   And so

$$\alpha_i((z_i, x^t_i)|(x^t_i, x^t_{-i})) = \min\left(1, \frac{1-p_i(x^t_i|x^t_{-i})}{1-p_i(z_t|x^t_{-i})}\right)$$

2. We consider using a cyclic scan Gibbs sampler that samples each variable X_i in a cyclic order. The Markov Chain produced by the Metropolized Gibbs sampler has p as a stationary distribution because the detailed balance equation is satisfied by the construction of α_i. p is unique because the transition kernel A_{ij} is regular. It is obvious that $A_{ij} > 0 \forall i \neq j$, but $A_{ii} > 0$ too since $A_{ii} = \sum_{z_i \neq x^t_i} q_i((z_i, x^t_{-i})|(x^t_i, x^t_{-i}))(1 - \alpha_i((z_i, x^t_i)|(x^t_i, x^t_{-i}))) > 0$.

6 Bayesian point estimates: posterior mean vs. MAP for exponential family

We have an i.i.d sample $(x_1, \cdots, x_n)$ of Bernoulli random variables whose unknown mean parameter is $\mathbb{E}_{X \sim p(\cdot|\mu)}\{X\} = \mu$

1. We want to compute the posterior mean estimate as a function of $\bar{x}$ We first compute $p(\mu|x_1, \cdots, x_n)$. Notice that:

$$p(\mu|x_1, \cdots, x_n) \propto p(x_1, \cdots, x_n|\mu)p(\mu) = \mu^{n\bar{x}}(1-\mu)^{n-n\bar{x}} \cdot 1 \tag{6.1}$$

   We recognize (6.1) as a Beta distribution with $\alpha = n\bar{x} + 1$ and $\beta = n - n\bar{x} + 1$ i.e.

$$p(\mu|x_1, \cdots, x_n) = \frac{\Gamma(n+2)}{\Gamma(n\bar{x}+1)\Gamma(n-n\bar{x}+1)}\mu^{n\bar{x}}(1-\mu)^{n-n\bar{x}} \tag{6.2}$$

   Therefore,

$$\mathbb{E}_{\mu \sim p(\cdot|X)}\{\mu\} = \frac{n\bar{x}+1}{n\bar{x}+1+n-n\bar{x}+1} = \frac{n\bar{x}+1}{n+2} \tag{6.3}$$

2. Since the prior is uniform, we have that the MAP estimate for μ is the MLE. We find the MLE using stationary points by differentiating the log (for computational convenience). We get

$$\begin{aligned} \frac{\partial \log p(\mu|x_1, \cdots, x_n)}{\partial \mu} &= \frac{n\bar{x}}{\mu} - \frac{n-n\bar{x}}{1-\mu} = 0 \\ &\Rightarrow n\bar{x} - \mu n\bar{x} - \mu n + \mu n\bar{x} = 0 \\ &\Rightarrow n\bar{x} = \mu n \\ &\Rightarrow \mu = \bar{x} \end{aligned}$$

3. We express the canonical parameter η of a Bernouilli as a function of its moment parameter μ.

$$p(x|\mu) = \exp\left\{x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right\} \tag{6.4}$$

   Which we can write as

$$p(x|\eta) = \exp\left\{T(x)\eta - A(\eta)\right\} h(x) \tag{6.5}$$

   Where

$$\begin{aligned} T(x) &= x \\ \eta &= \log \frac{\mu}{1-\mu} \\ A(\eta) &= -\log(1-\mu) = \log(1+e^{\eta}) \\ h(x) &= 1 \end{aligned}$$

Note from the above $\eta = \log \frac{\mu}{1-\mu} \Rightarrow \mu = \sigma(\eta)$.

To compute $p_\eta(\eta)$ we use the following increasing transformation $\eta = g(\mu)$ (notice $g^{-1}(\eta) = \sigma(\eta)$). We see that:

$$\begin{aligned} p_\eta(\eta) &= p_\mu(g^{-1}(\eta)) \left| \frac{dg^{-1}(\eta)}{d\eta} \right| \\ &= p_\mu(\sigma(\eta)) \left| \frac{d\sigma(\eta)}{d\eta} \right| \\ &= 1 \cdot \sigma(\eta)(1 - \sigma(\eta)) \end{aligned}$$

As needed.

4. The posterior mean estimate is

$$\begin{aligned} \int \mu(\eta) p(\eta | x_1, \cdots, x_n) d\eta &= \int \sigma(\eta) \frac{p(x_1, \cdots, x_n | \eta) p(\eta)}{p(x_1, \cdots, x_n)} d\eta \\ &= \int \sigma(\eta) \frac{p(x_1, \cdots, x_n | \eta)}{p(x_1, \cdots, x_n)} \underbrace{\sigma(\eta)(1 - \sigma(\eta)) d\eta}_{= d\mu} \\ &= \int \mu \underbrace{\frac{p(x_1, \cdots, x_n | \mu)}{p(x_1, \cdots, x_n)} p(\mu)}_{= p(\mu | x_1, \cdots, x_n)} d\mu \\ &= \mathbb{E}_{\mu \sim p(\cdot | X)} \{\mu\} \\ &= \frac{n\bar{x} + 1}{n + 2} \end{aligned}$$

5. Again, the MAP estimator is the MLE, since the prior is uniform. We use that the MLE is a plug in estimator to get that $\eta^{MAP} = \log \frac{\bar{x}}{1-\bar{x}}$. Then $\mu(\eta^{MAP}) = \bar{x}$.

6. Note that 1. and 3. are the same because the integral is essentially the same (with the only difference being the measure). 2. and 5. are the same since the MLE is the same as MAP for a uniform prior, and the MLE is invariant to continous mappings.