CS228: Probabilistic Graphical Models

Homework 5

Luke Jaffe

Due: 03/18/2017
Submitted: 03/18/2017

Problem 1: Bayesian Inference

Problem 1) Bayesian Inference

a) $P(D|\Theta) = \prod_{i=1}^{M} P(x[i]|\Theta)$, $\Theta \sim Dirichlet(\alpha_1, \dots, \alpha_k)$

$\Theta \sim Dirichlet(\alpha_1, \dots, \alpha_k)$ if $P(\Theta) \propto \prod_k \Theta_k^{\alpha_k - 1}$

$M[i]$ is # times $x[m] = x^i$ appears in dataset, $\alpha = \sum_i \alpha_i$, $X[M+1] \perp D \,|\, \Theta$

posterior $P(\Theta | x[1], \dots, x[m])$ given by $Dirichlet(\alpha_1', \dots, \alpha_k')$

where $\alpha_k' = \alpha_k + \sum_{j=1}^{M} 1\{x[j] = x^k\}$

Show that $P(x[M+1] = x^i | D) = \frac{M[i] + \alpha_i}{M + \alpha}$

$E[Dirichlet(\alpha_1, \dots, \alpha_k)] = \frac{\alpha_i}{\sum_k \alpha_k}$

The predictive probability is the expected value of the posterior

$P(x[M+1] = x^i | D) = E[Dirichlet(\alpha_1', \dots, \alpha_k')]$

$$= \frac{\alpha_i'}{\sum_k \alpha_k'} = \frac{\alpha_i + \sum_{j=1}^{M} 1\{x[j] = x^i\}}{\sum_k [\alpha_k + \sum_{j=1}^{M} 1\{x[j] = x^k\}} = \frac{\alpha_i + M[i]}{\sum_k \alpha_k + \sum_k \sum_{j=1}^{M} 1\{x[j] = x^k\}}$$

$$= \frac{M[i] + \alpha_i}{M + \alpha}$$ QED

b) Show how to compute $P(X[M+1]=x^i, X[M+2]=x^j \mid D)$

If we recognize the gamma function as the kernel of the distribution, we can see

$$E[X_i] = \frac{\Gamma(\alpha)}{\Gamma(\alpha_i)\prod_{i=2}^{K}\Gamma(\alpha_i)} \int_0^1 \cdots \int_0^{1-\sum_{j=1}^{K-1} x_j} x_i^{\alpha_i} \prod_{j=2}^{K} x_j^{\alpha_j - 1} \, dx_1 \ldots dx_K$$

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha_i)\prod_{i=2}^{K}\Gamma(\alpha_i)} \cdot \frac{\Gamma(\alpha_i+1)\prod_{i=2}^{K}\Gamma(\alpha_i)}{\Gamma(\alpha+1)}$$

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha+1)} \cdot \frac{\Gamma(\alpha_i+1)}{\Gamma(\alpha_i)} = \frac{\alpha_i}{\alpha}$$

Now for the cross-moment:

$$E[X_i X_j] = \frac{\Gamma(\alpha)}{\Gamma(\alpha+2)} \cdot \frac{\Gamma(\alpha_i+1)\Gamma(\alpha_j+1)}{\Gamma(\alpha_i)+\Gamma(\alpha_j)} = \frac{\alpha_i \alpha_j}{\alpha(\alpha+1)}$$

Plugging in posterior dirichlet:

$$P(X[M+1]=x^i, X[M+2]=x^j \mid D) = \boxed{\frac{(\alpha_i + M[i])(\alpha_j + M[j])}{(\alpha+M)(\alpha+M+1)}}$$

c) $P(X[M+1]=x^i \mid D) \cdot P(X[M+2]=x^i \mid D) =$

$$= \frac{(M[i]+\alpha_i)(M[i]+\alpha_i)}{(M+\alpha)(M+\alpha)}$$

The ratio of the approximation to correct value is:

$$\frac{(M[i]+\alpha_i)(M[i]+\alpha_i)}{(M+\alpha)(M+\alpha)} \cdot \frac{(M+\alpha)(M+\alpha+1)}{(M[i]+\alpha_i)(M[i]+\alpha_i)}$$

$$= \frac{(M+\alpha)(M+\alpha+1)}{(M+\alpha)(M+\alpha)} = \frac{M+\alpha+1}{M+\alpha} = \boxed{1+\frac{1}{M+\alpha}}$$

Thus, as $M \to \infty$, the error $\to 1$.

For small M, the error can be up to 2.

Problem 2: Programming Assignment

A) Gaussian mixture model

i. Parameter estimates:

```
MLE estimates for PA part A.i:
pi: 0.57
mu_0: [-0.99437209 -1.11730233]
mu_1: [ 1.04922807  0.98085965]
sigma_0: [[ 0.315455    0.29233619]
          [ 0.29233619  0.8328346 ]]
sigma_1: [[ 0.79217671  0.20035059]
          [ 0.20035059  0.25443312]]
```

ii. Parameter estimates:

```
MLE init:
{'pi': 0.58618066426826065, 'sigma_1': array([[ 0.72032607,  0.14375235],
       [ 0.14375235,  0.30904765]]), 'sigma_0': array([[ 0.36106206,  0.30946665],
       [ 0.30946665,  0.75617395]]), 'mu_1': array([ 0.98809485,  0.99678744]),
'mu_0': array([-1.04528972, -1.02646853])}

Rand init 1:
{'pi': 0.58469304369785646, 'sigma_1': array([[ 0.71692698,  0.14185481],
       [ 0.14185481,  0.30804894]]), 'sigma_0': array([[ 0.36273897,  0.31211656],
       [ 0.31211656,  0.76146101]]), 'mu_1': array([ 0.99144466,  0.99839573]),
'mu_0': array([-1.04272223, -1.02148551])}

Rand init 2:
{'pi': 0.58447227483326591, 'sigma_1': array([[ 0.71651377,  0.14162268],
       [ 0.14162268,  0.30792467]]), 'sigma_0': array([[ 0.36305983,  0.3126073 ],
       [ 0.3126073 ,  0.76234312]]), 'mu_1': array([ 0.99190709,  0.99861054]),
'mu_0': array([-1.04229193, -1.0207145 ])}
```

That that the convergence criterion suggested in the problem statement was used, not the one in the starter code (absolute difference of eps).

All methods converged to the around the same log likelihood ~-2570, and arrived at roughly the same parameters. I had some runs where the rand inits arrived at wildly different parameters than the MLE init.

Problem 2) Programming Assignment

i. Estimate The parameters $\pi, \mu_0, \mu_1, \Sigma_0, \Sigma_1$ by MLE

Let $T = M+N$, $T[0]$ is # $Z_{ij}=0$, $T[1]$ is # $Z_{ij}=1$

$$\pi = \frac{1}{T} \cdot \sum_{i,j} Z_{ij}$$ indicator function syntax used

$$\mu_0 = \frac{1}{T[0]} \cdot \sum_{i,j} X_{ij}[Z_{ij}=0], \quad \mu_1 = \frac{1}{T[1]} \cdot \sum_{i,j} X_{ij}[Z_{ij}=1]$$

$$\Sigma_0 = \frac{1}{T[0]-1} \sum_{i,j} (X_{ij}-\mu_0)^T (X_{ij}-\mu_0)[Z_{ij}=0]$$

$$\Sigma_1 = \frac{1}{T[1]-1} \sum_{i,j} (X_{ij}-\mu_1)^T (X_{ij}-\mu_1)[Z_{ij}=1]$$

ii. Implement a Gaussian-Mixtures EM algorithm, estimate Θ

$$ll(D;\Theta) = \sum_{i,j} \log p(X_{ij}|\Theta)$$

$$P(X_{ij}|\Theta) = (1-\pi) N(X_{ij}|\mu_0, \Sigma_0) + \pi N(X_{ij}|\mu_1, \Sigma_1)$$

E-step: Expand and weight dataset

Let W_Z hold weights for $Z=0,1$

$$W\_Z[0]_{ij} = (1-\pi) N(X_{ij}|\mu_0, \Sigma_0)$$
$$W\_Z[1]_{ij} = \pi N(X_{ij}|\mu_1, \Sigma_1)$$

Normalize weights so $W\_Z[0]_{ij} + W\_Z[1]_{ij} = 1$ for all i,j

M-step: Estimate parameters

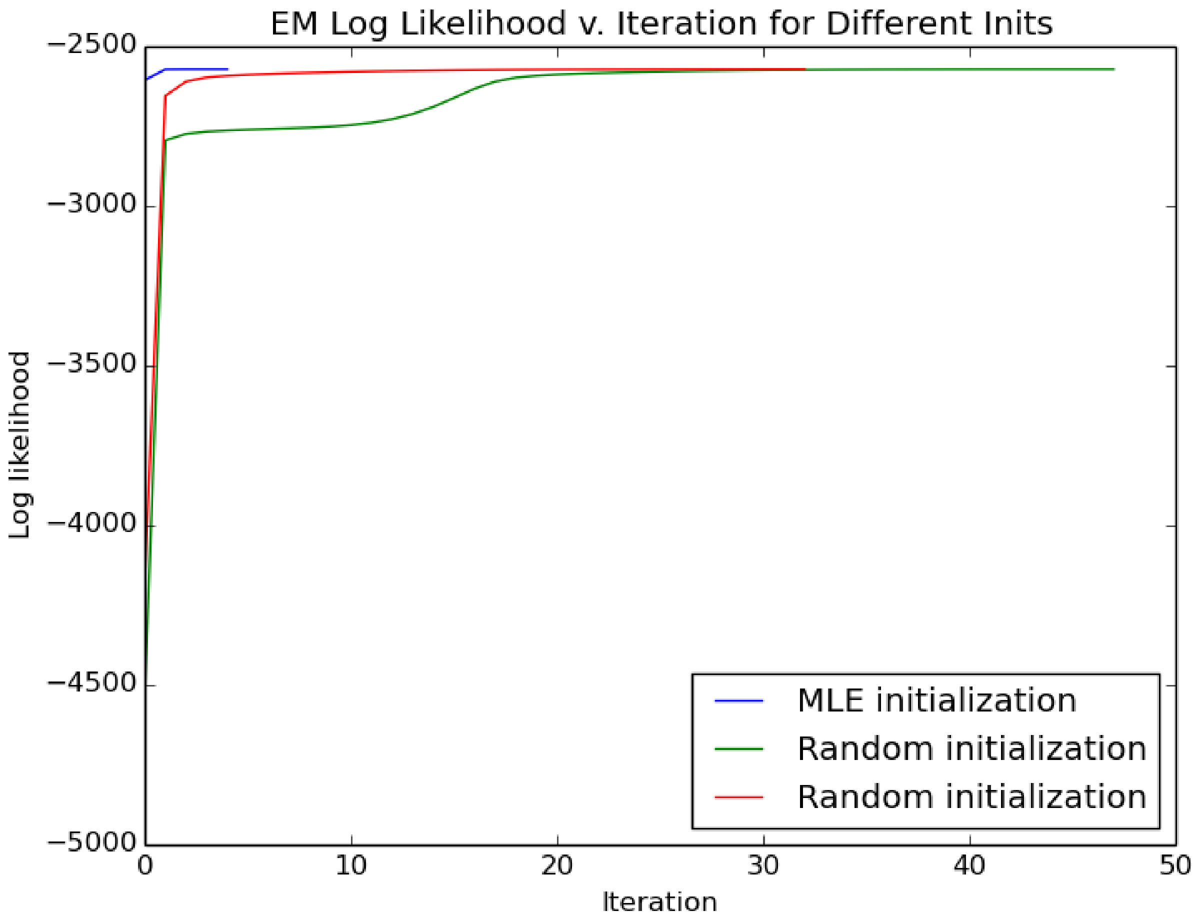
$$\pi = \frac{1}{T} \sum_{i,j} W\_Z[1]_{ij}$$

$$\Sigma_0 = \frac{\sum_{i,j} (X_{ij}-\mu_0)^T (X_{ij}-\mu_0) \cdot W\_Z[0]_{ij}}{\sum_{i,j} W\_Z[0]_{ij} - 1}$$

$$\mu_0 = \frac{\sum_{i,j} X_{ij} \cdot W\_Z[0]_{ij}}{\sum_{i,j} W\_Z[0]_{ij}}$$

$$\Sigma_1 = \frac{\sum_{i,j} (X_{ij}-\mu_1)^T (X_{ij}-\mu_1) \cdot W\_Z[1]_{ij}}{\sum_{i,j} W\_Z[1]_{ij}}$$

$$\mu_1 = \frac{\sum_{i,j} X_{ij} \cdot W\_Z[1]_{ij}}{\sum_{i,j} \cdot W\_Z[1]_{ij}}$$



B) Geography-aware mixture model

i. Parameter estimates:

```
MLE estimates for PA part B.i:
MLE phi: 0.6
MLE lambda: 0.93
```

B) Geography-aware mixture model

i. $\ell\ell(D;\theta) = \sum_{i,j} \log p(X_{ij} \mid \theta)$

$$p(X_{ij} \mid \theta) = \phi\lambda \mathcal{N}(X_{ij} \mid \mu_1, \Sigma_1) + \phi(1-\lambda)\mathcal{N}(X_{ij} \mid \mu_0, \Sigma_0) + (1-\phi)\lambda \mathcal{N}(X_{ij} \mid \mu_0, \Sigma_0) + (1-\phi)(1-\lambda)\mathcal{N}(X_{ij} \mid \mu_1, \Sigma_1)$$

$$= \mathcal{N}(X_{ij} \mid \mu_0, \Sigma_0) \cdot (\phi(1-\lambda) + (1-\phi)\lambda) + \mathcal{N}(X_{ij} \mid \mu_1, \Sigma_1) \cdot (\phi\lambda + (1-\phi)(1-\lambda))$$

$$y_i = \mathbb{I}\left(\sum_{j=1}^{M} z_{ij} \geq \frac{M}{2}\right)$$

$$\phi = \frac{1}{N} \sum_i y_i$$

$$\lambda = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{M} \sum_{j=1}^{M} z_{ij} = \frac{1}{N \cdot M} \sum_{i,j} z_{ij}$$

ii. * indicates marked for $p > 0.5$

ii.

$$P(y_i = 1 \mid x_{i,1:m}) = \frac{p(y_i = 1) \sum_{z_{i,1:m}} \prod_{j=1}^{m} p(x_{ij} \mid z_{ij}) \, p(z_{ij} \mid y_i = 1)}{P(x_{i,1:m})}$$

$$= \left[\phi \prod_{j=1}^{m} \sum_{z_{ij}} p(x_{ij} \mid z_{ij}) \, p(z_{ij} \mid y_i = 1)\right] / P(x_{i,1:m})$$

$$= \left[\phi \prod_{j=1}^{m} \left[p(x_{ij} \mid z_{ij} = 0) \cdot p(z_{ij} = 0 \mid y_i = 1) + p(x_{ij} \mid z_{ij} = 1) P(z_{ij} = 1 \mid y_i = 1)\right]\right] / P(x_{i,1:m})$$

$$= \left[\phi \prod_{j=1}^{m} \left[\mathcal{N}(x_{ij} \mid \mu_0, \Sigma_0) \cdot (1-\lambda) + \mathcal{N}(x_{ij} \mid \mu_1, \Sigma_1) \cdot \lambda\right]\right] / P(x_{i,1:m})$$

$$= \boxed{\frac{\phi \prod_{j=1}^{m} \mathcal{N}_0 \cdot (1-\lambda) + \mathcal{N}_1 \cdot \lambda}{\phi \prod_{j=1}^{m} \mathcal{N}_0 \cdot (1-\lambda) + \mathcal{N}_1 \cdot \lambda + (1-\phi) \prod_{j=1}^{m} \mathcal{N}_0 \cdot \lambda + \mathcal{N}_1 \cdot (1-\lambda)}}$$

$$P(z_{ij} = 1 \mid x_{i,1:m}) = \frac{P(z_{ij} = 1, x_{i,1:m})}{P(x_{i,1:m})} = \frac{\sum_{y_i} p(y_i) p(z_{ij} = 1 \mid y_i) P(x_{i,1:m} \mid z_{ij} = 1)}{P(x_{i,1:m})}$$

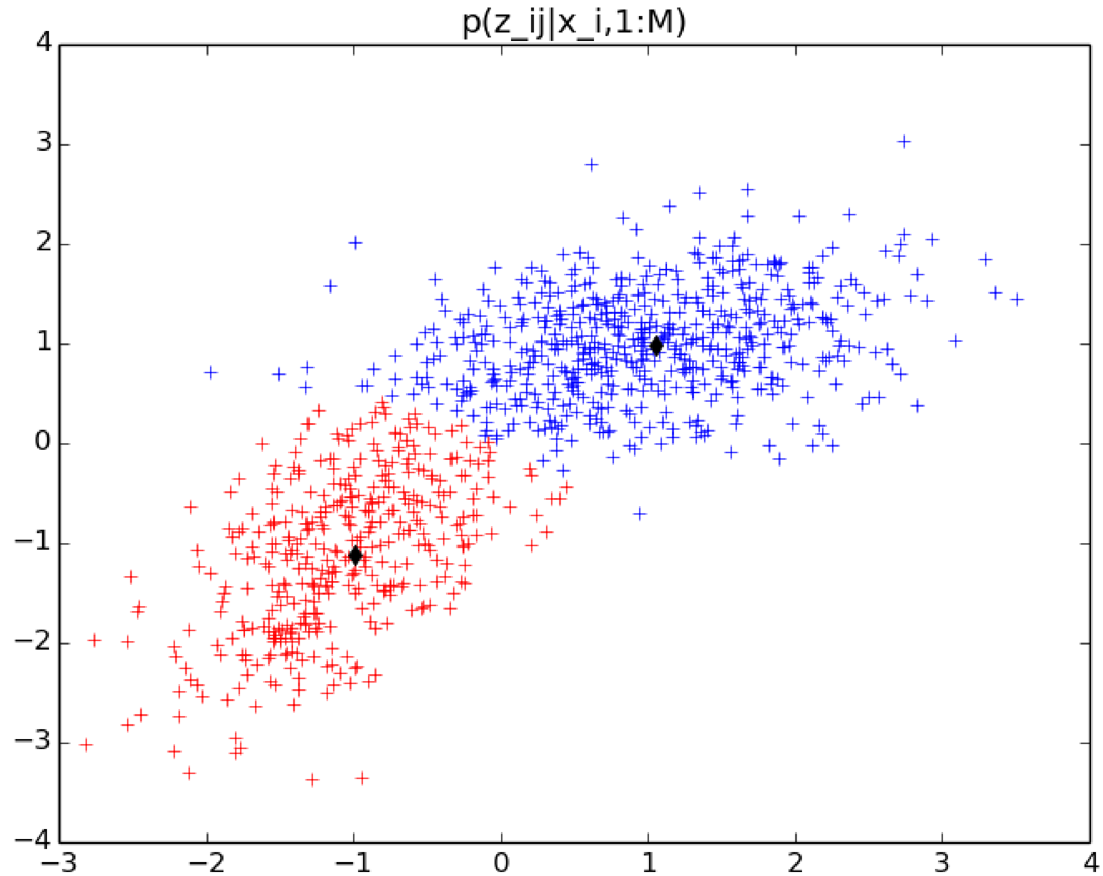
$$= \frac{P(x_{ij} \mid z_{ij} = 1) P(x_{i,1:m \neq j} \mid z_{ij} = 1) \sum_{y_i} P(y_i) p(z_{ij} = 1 \mid y_i)}{P(x_{i,1:m})}$$

$$= \left[P(x_{ij} \mid z_{ij} = 1) P(x_{i,1:m \neq j}) \sum_{y_i} P(y_i) P(z_{ij} = 1 \mid y_i)\right] / P(x_{i,1:m})$$

$$= \frac{P(x_{ij} \mid z_{ij} = 1) P(x_{i,1:m \neq j}) \sum_{y_i} P(y_i) P(z_{ij} = 1 \mid y_i)}{P(x_{i,1:m \neq j}) \left(P(x_{ij} \mid z_{ij} = 1) \sum_{y_i} P(y_i) P(z_{ij} = 1 \mid y_i) + P(x_{ij} \mid z_{ij} = 0) \sum_{y_i} P(y_i) P(z_{ij} = 0 \mid y_i) \right)}$$

$$= \boxed{\frac{\mathcal{N}(x_{ij} \mid \mu_1, \Sigma_1) \cdot [\phi\lambda + (1-\phi)(1-\lambda)]}{\mathcal{N}(x_{ij} \mid \mu_1, \Sigma_1) \cdot [\phi\lambda + (1-\phi)(1-\lambda)] + \mathcal{N}(x_{ij} \mid \mu_0, \Sigma_0) [\phi \cdot (1-\lambda) + (1-\phi)\lambda]}}$$

Precinct	p(yi|xi,1:M)
* 0	1.000
* 1	1.000
2	0.000
* 3	1.000
4	0.000
* 5	1.000
* 6	1.000
* 7	1.000
* 8	1.000
* 9	1.000
10	0.000
11	0.000
12	0.000
* 13	1.000
* 14	1.000
15	0.000
16	0.000
* 17	1.000
* 18	1.000
* 19	1.000
20	0.000
* 21	1.000
* 22	1.000
23	0.000
* 24	1.000
25	0.000
26	0.000
27	0.000
* 28	1.000
29	0.000
* 30	1.000
* 31	1.000
* 32	1.000
33	0.000
* 34	1.000
35	0.000
* 36	1.000
37	0.000
38	0.000
* 39	1.000
40	0.000
* 41	1.000
* 42	1.000
43	0.000
* 44	1.000
45	0.000
* 46	1.000
47	0.000
* 48	1.000
49	0.000


p(z_ij|x_i,1:M)
4
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4

iii.

iii. $\ell\ell(D;\theta) = \sum_{i,j} \log P(x_{i,j} \mid \theta)$

$$P(x_{i,j} \mid \theta) = N(x_{i,j} \mid \mu_0, \Sigma_0) \cdot (\phi \cdot (1-\lambda) + (1-\phi)\cdot\lambda)$$
$$+ N(x_{i,j} \mid \mu_1, \Sigma_1) \cdot (\phi\lambda + (1-\phi)\cdot(1-\lambda))$$

E-step: Expand dataset and weight points

Compute $P(y_i \mid x_{i,1:m}, \theta)$ (See part B ii)
Compute $P(z_{i,j} \mid x_{i,1:m}, \theta)$ (See part B ii)
Compute $P(y_i, z_{i,j} \mid x_{i,1:m}, \theta)$

$$P(y_i, z_{i,j} \mid x_{i,1:m}, \theta) = \frac{P(y_i, z_{i,j}, x_{i,1:m})}{P(x_{i,1:m})} = \frac{P(y_i)P(z_{i,j} \mid y_i)P(x_{i,j} \mid z_{i,j})P(x_{i,1:m \neq j})}{P(x_{i,1:m})}$$

$$= \frac{P(x_{i,1:m \neq j})\, P(y_i)P(z_{i,j} \mid y_i)P(x_{i,j} \mid z_{i,j})}{P(x_{i,1:m \neq j}) \cdot \sum_{y_i, z_{i,j}} P(y_i)P(z_{i,j} \mid y_i)P(x_{i,j} \mid z_{i,j})}$$

$$= \frac{P(y_i)P(z_{i,j} \mid y_i)P(x_{i,j} \mid z_{i,j})}{\sum_{y_i, z_{i,j}} P(y_i)P(z_{i,j} \mid y_i)P(x_{i,j} \mid z_{i,j})}$$

M-step: Estimate Parameters

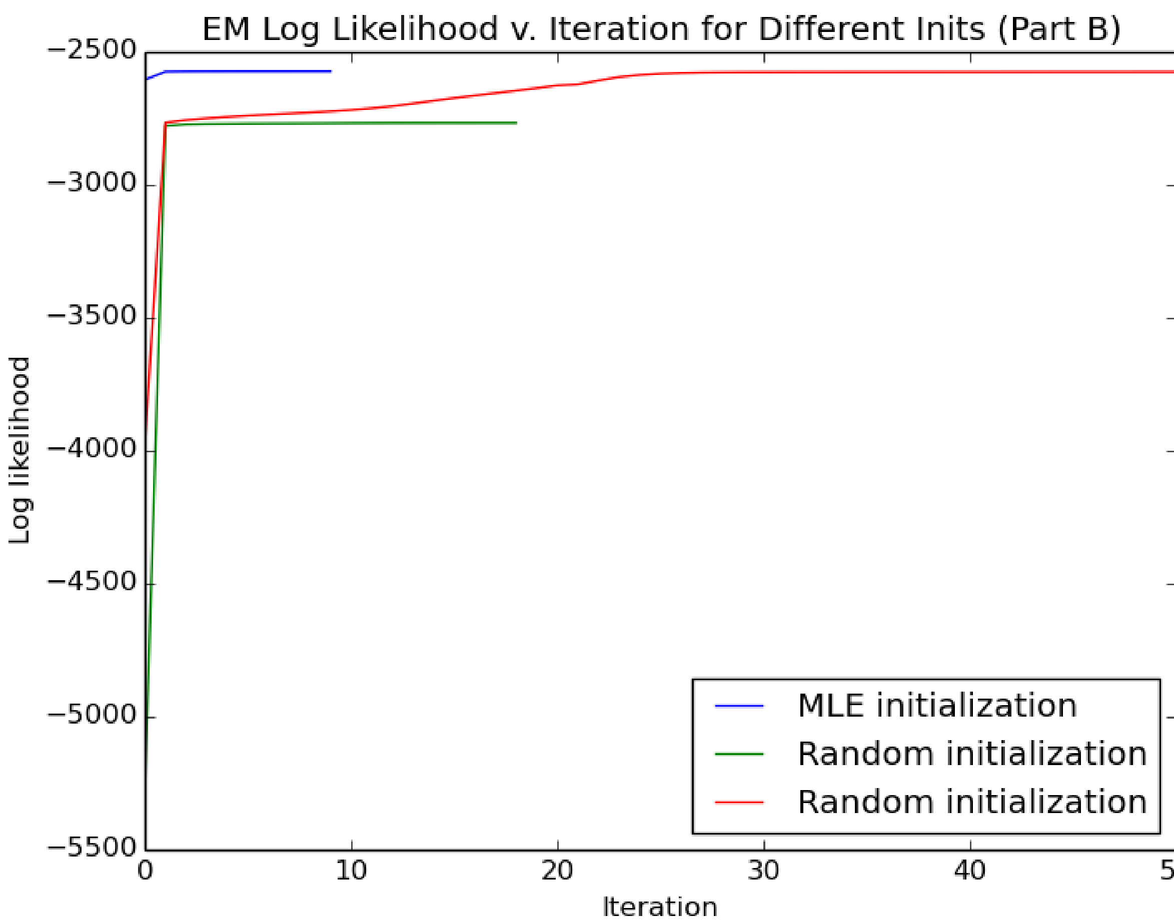
$$\phi = \frac{1}{N} \cdot \sum_i P(y_i \mid x_{i,1:m}, \theta)$$

$$\lambda = \frac{1}{N \cdot M} \cdot \sum_{i,j} p(y_i = 0, z_{i,j} = 0 \mid x_{i,1:m}, \theta) + p(y_i = 1, z_{i,j} = 1 \mid x_{i,1:m}, \theta)$$

Parameter updates for $\mu_0, \mu_1, \Sigma_0, \Sigma_1$ are the same as shown for part A ii, except weighted using $P(z_{i,j} \mid x_{i,1:m}, \theta)$ as computed above.

iv. Parameter estimates:

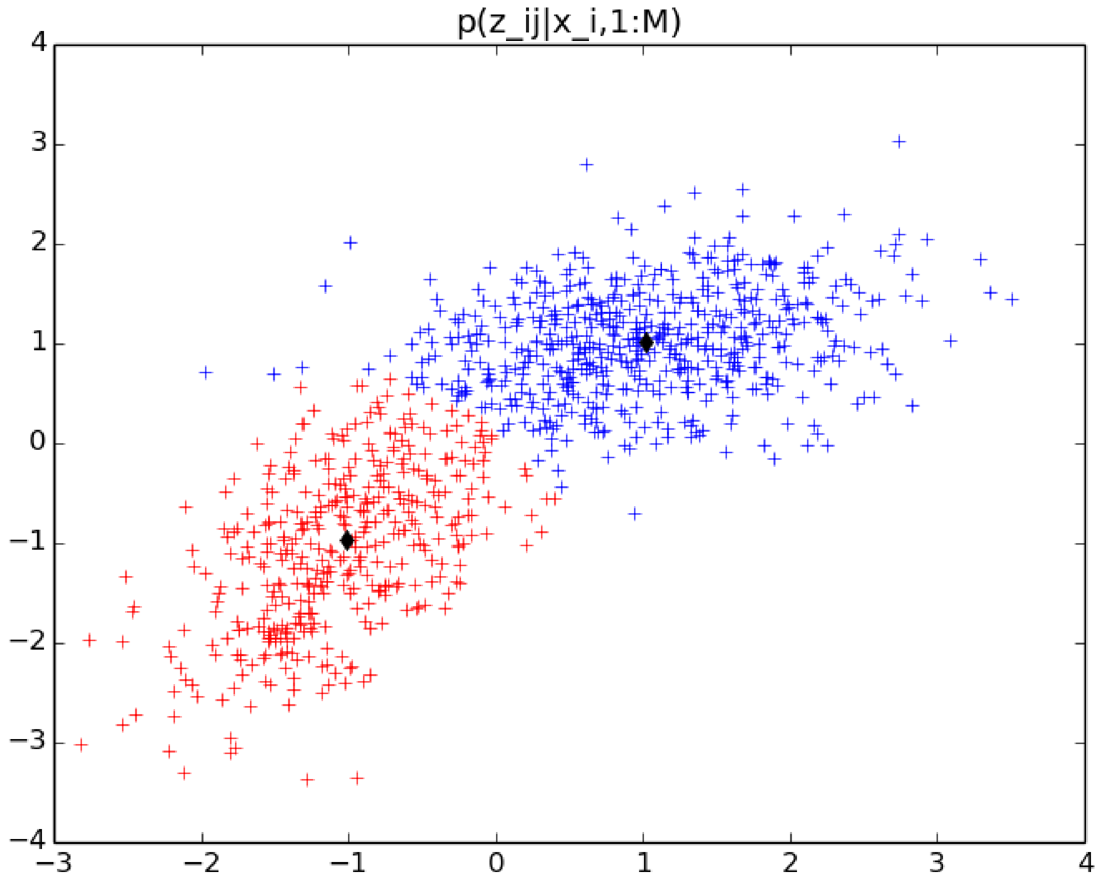
```
MLE init:
{'phi': 0.56000005452278112, 'sigma_1': array([[ 0.69270523,  0.12859102],
       [ 0.12859102,  0.30099677]]), 'sigma_0': array([[ 0.38317011,  0.34278206],
       [ 0.34278206,  0.81501996]]), 'mu_1': array([ 1.01946161,  1.01102893]),
'mu_0': array([-1.01584381, -0.97456872]), 'lambda': 0.93656267888912736}

Rand init 1:
{'phi': 0.0, 'sigma_1': array([[ 0.53861911, -0.04532715],
       [-0.04532715,  0.00839436]]), 'sigma_0': array([[ 1.57358974,  1.22068669],
       [ 1.22068669,  1.4939656 ]]), 'mu_1': array([-1.04553966,  0.65314159]),
'mu_0': array([ 0.15288455,  0.15693989]), 'lambda': 0.99479020426822873}

Rand init 2:
{'phi': 0.56013618371793228, 'sigma_1': array([[ 0.41309118,  0.38771891],
       [ 0.38771891,  0.89215085]]), 'sigma_0': array([[ 0.66396172,  0.11593109],
       [ 0.11593109,  0.29447559]]), 'mu_1': array([-0.9792153 , -0.91169995]),
'mu_0': array([ 1.05498714,  1.02379431]), 'lambda': 0.27571526114628297}
```



For this trial, neither random initialization reached the same parameter estimates as the MLE init. Only one of the random trials converged to the same log likelihood. Perhaps this EM is less stable since there are more parameters, would need to run more trials to test empirically.

v. * indicates marked for p > 0.5

```
-------------------------
|Precinct |p(yi|xi,1:M) |
-------------------------
|* 0      | 1.000       |
|* 1      | 1.000       |
|  2      | 0.000       |
|* 3      | 1.000       |
|  4      | 0.000       |
|* 5      | 1.000       |
|* 6      | 1.000       |
|* 7      | 1.000       |
|* 8      | 1.000       |
|* 9      | 1.000       |
|  10     | 0.000       |
|  11     | 0.000       |
|  12     | 0.000       |
|* 13     | 1.000       |
|* 14     | 1.000       |
|  15     | 0.000       |
|  16     | 0.000       |
|* 17     | 1.000       |
|* 18     | 1.000       |
|* 19     | 1.000       |
|  20     | 0.000       |
|* 21     | 1.000       |
|* 22     | 1.000       |
|  23     | 0.000       |
|* 24     | 1.000       |
|  25     | 0.000       |
|  26     | 0.000       |
|  27     | 0.000       |
|* 28     | 1.000       |
|  29     | 0.000       |
|* 30     | 1.000       |
|* 31     | 1.000       |
|* 32     | 1.000       |
|  33     | 0.000       |
|* 34     | 1.000       |
|  35     | 0.000       |
|* 36     | 1.000       |
|  37     | 0.000       |
|  38     | 0.000       |
|* 39     | 1.000       |
|  40     | 0.000       |
|* 41     | 1.000       |
|* 42     | 1.000       |
|  43     | 0.000       |
|* 44     | 1.000       |
|  45     | 0.000       |
|* 46     | 1.000       |
|  47     | 0.000       |
|* 48     | 1.000       |
|  49     | 0.000       |
```


p(z_ij|x_i,1:M)
4
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4