CS 228 Problem Set 5

Hugh Zhang

March 17, 2017

Problem 1

Note that from the hint, we can see that the Dirichlet distribution is a distributed beta, so all the lecture theorems apply.

$$\begin{aligned} & P(X[M+1] = x^i \mid D) \\ = & \int_\theta P(X[M+1] = x^i \mid \theta) * P(\theta \mid D) d\theta \\ = & \int_\theta \theta_i * P(\theta \mid D) d\theta \\ = & E_{\theta|D}[\theta_i] = \frac{\alpha_i + M[i]}{M + \alpha} \end{aligned}$$

with the last inequality taken straight from lecture.

1.2

$$\begin{aligned} & P(X[M+1] = x^i, X[M+2] = x^j \mid D) \\ = & P(X[M+1] = x^i \mid D) * P(X[M+2] = x^i \mid X[M+1] = x^j, D) \end{aligned}$$

But notice, the second part, is just augmenting the data set by one more example. Thus, our probabilities are just

$$\frac{\alpha_i + M[i]}{M + \alpha} * \frac{\alpha_j + M[j] + 1(i = j)}{M + \alpha + 1}$$

1.3

$$
\begin{aligned}
& P(X[M+1] = x^i, X[M+2] = x^j \mid D) \\
& /P(X[M+1] = x^i \mid D) * P(X[M+2] = x^i \mid D) \\
= \; & (\frac{\alpha_i + M[i]}{M+\alpha} * \frac{\alpha_j + M[j] + 1(i=j)}{M+\alpha+1}) / (\frac{\alpha_i + M[i]}{M+\alpha} * \frac{\alpha_j + M[j]}{M+\alpha}) \\
= \; & \frac{(\alpha_j + M[j] + 1(i=j))}{(\alpha_j + M[j])} * \frac{(M+\alpha+1))}{(M+\alpha)} \\
= \; &
\end{aligned}
$$

The difference here is if you update your Dirichlet counts after you sample the M+1th sample when you do exact inference and don't if you do approximate. Thus, if your data set is large, then not updating doesn't matter, but if it is small it might make a large difference.

With the function above, both factors of the product clearly have the +1 become irrelevant as M goes to infinity and dwarfs it.

Problem 2

Let T be the total number of examples.

$$\pi = \sum_{ij} \frac{1[Z_{ij} = 1]}{T} = 0.57$$

$$\mu_0 = \sum_{ij} \frac{1[Z_{ij} = 0] * X_{ij}}{1[Z = 0]} = [-0.99437209, -1.11730233]$$

$$\mu_1 = \sum_{ij} \frac{1[Z_{ij} = 1] * X_{ij}}{1[Z = 1]} = [1.04922807, 0.98085965]$$

$$\sigma_0 = \sum_{ij} \frac{1[Z_{ij} = 0] * (X_{ij} - \mu_0)(X_{ij} - \mu_0)^T}{1[Z = 0]} = [[0.30811884, 0.28553768][0.28553768, 0.81346635]]$$

$$\sigma_1 = \sum_{ij} \frac{1[Z_{ij} = 1] * (X_{ij} - \mu_1)(X_{ij} - \mu_1)^T}{1[Z = 1]} = [[0.77827888, 0.19683566][0.19683566, 0.24996938]]$$

2.A2

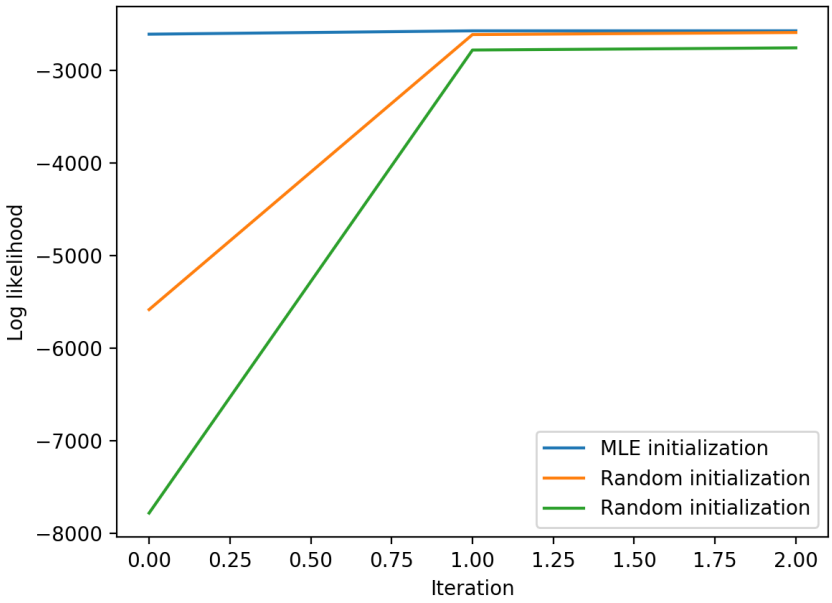

Figure 1: Log likelihood functions

MLE start:

```
{'pi': 0.58616150703175707, 'sigma_1': matrix([[ 0.72112418,  0.14499114],
[ 0.14499114,  0.30825171]]), 'sigma_0': matrix([[ 0.36212646,  0.31082931],
[ 0.31082931,  0.75836101]]),
'mu_1': matrix([[ 0.98729587,  0.99618266]]),
'mu_0': matrix([[-1.04406391, -1.02551825]])}
```

Random 1 start:

```
{'pi': 0.51535490062201894, 'sigma_1': matrix([[ 0.5953022 ,  0.54367732],
[ 0.54367732,  1.01903159]]), 'sigma_0': matrix([[ 0.6061935 ,  0.05204946],
[ 0.05204946,  0.24784284]]), \\ 'mu_1': matrix([[-0.80990003, -0.72964472]]),
'mu_0': matrix([[ 1.16379378,  1.10503744]])}
```

Random 2 start:

```
{'pi': 0.18461282486659261, 'sigma_1': matrix([[ 1.66246823,  1.30203827],
[ 1.30203827,  1.3425169 ]]), 'sigma_0': matrix([[ 1.31417998,  1.17278467],
[ 1.17278467,  1.51739247]]), \\ 'mu_1': matrix([[ 1.07480467,  0.22227733]]),
'mu_0': matrix([[-0.06350569,  0.14531717]])}
```

One of the random starts seems to have found approximately the same maxima that MLE found, but the other one ended up in a slightly worse local minimum.

2.B1

See part 2.A1 for other formulas

$$\phi = \sum_i \frac{1[Y_i = 1]}{N} = 0.6$$
$$\lambda = \sum_{ij} \frac{1[Z_{ij} = Y_i]}{M * N} = .93$$

2.B2

$$\begin{aligned}
& P(Y_i = 1 \mid X_{i,1\ldots M}) \\
= \; & \frac{\sum_Z P(Y_i = 1, Z, X_{i,1\ldots M})}{P(X_{i,1\ldots M})} \\
= \; & \frac{P(Y_i = 1) \sum_{z_i,1\ldots M} \prod_{j=1}^{M} p(x_{ij} \mid z_{ij}) * p(z_{ij} \mid y_{ij})}{P(X_{i,1\ldots M})} \\
= \; & \frac{P(Y_i = 1) \prod_{j=1}^{M} \sum_{z_i,1\ldots M} p(x_{ij} \mid z_{ij}) * p(z_{ij} \mid y_{ij})}{P(X_{i,1\ldots M})} \\
= \; & \frac{P(Y_i = 1) \prod_{j=1}^{M} \sum_{z_i,1\ldots M} p(x_{ij} \mid z_{ij}) * p(z_{ij} \mid y_{ij})}{\sum_y P(Y_i = y) \prod_{j=1}^{M} \sum_{z_i,1\ldots M} p(x_{ij} \mid z_{ij}) * p(z_{ij} \mid y_{ij})}
\end{aligned}$$

And this last part is polynomial to calculate not exponential because we pushed the sum into the product so we can calculate the Z's independently (we can see they are independent from the Bayesian tree)

0	1.0	✓
1	1.0	✓
2	1.10687543644e-11	
3	1.0	✓
4	1.79599220851e-16	
5	1.0	✓
6	1.0	✓
7	1.0	✓
8	1.0	✓
9	1.0	✓
10	4.11317609256e-11	
11	2.27210526945e-09	
12	4.73776541146e-15	
13	1.0	✓
14	0.999999999913	✓
15	6.63072111609e-12	
16	1.95330106508e-14	
17	1.0	✓
18	1.0	✓
19	1.0	✓
20	2.58691416694e-16	
21	1.0	✓
22	0.999999999968	✓
23	2.65859289441e-11	
24	1.0	✓
25	5.71712529964e-10	
26	8.61995863343e-15	
27	3.6809211713e-15	
28	1.0	✓
29	1.03774348656e-12	
30	1.0	✓
31	1.0	✓
32	1.0	✓
33	1.16563125037e-11	
34	1.0	✓
35	2.03129334437e-12	
36	1.0	✓
37	9.35379152442e-14	
38	1.79920271419e-13	
39	1.0	✓
40	1.14212187728e-14	
41	1.0	✓
42	1.0	✓
43	1.45394221333e-14	
44	1.0	✓
45	8.28479865838e-10	
46	0.999999999987	✓
47	4.28728057356e-14	
48	0.999999999975	✓
49	9.21925967782e-05	

Since everything is indexed by i and precinct specific, for this derivation I will just drop the i and only refer to the j index (if necessary).

$$
\begin{aligned}
& P(Z_j = 1 \mid X_1 \dots M) \\
= \; & \frac{\sum_{Z_{-j}} \sum_Y P(Y, Z_{1\dots M}, X_{1\dots M})}{P(X_{1\dots M})} \\
= \; & \frac{\sum_Y P(Y) p(x_j \mid z_j) * p(z_j \mid y_j) \sum_{Z_{-j}} \prod_{-j} p(x_j \mid z_j) * p(z_j \mid y_j)}{P(X_{1\dots M})} \\
= \; & \frac{\sum_Y P(Y) p(x_j \mid z_j) * p(z_j \mid y_j) \prod_{-j} \sum_{Z_{-j}} p(x_j \mid z_j) * p(z_j \mid y_j)}{P(X_{1\dots M})} \\
= \; & \frac{\sum_Y P(Y) p(x_j \mid z_j) * p(z_j \mid y_j) \prod_{-j} \sum_{Z_{-j}} p(x_j \mid z_j) * p(z_j \mid y_j)}{\sum_Y P(Y) \prod_z \sum_Z p(x_j \mid z_j) * p(z_j \mid y_j)}
\end{aligned}
$$

Where this, like the equation above is polynomial and not exponential as desired because we pushed in the sums.

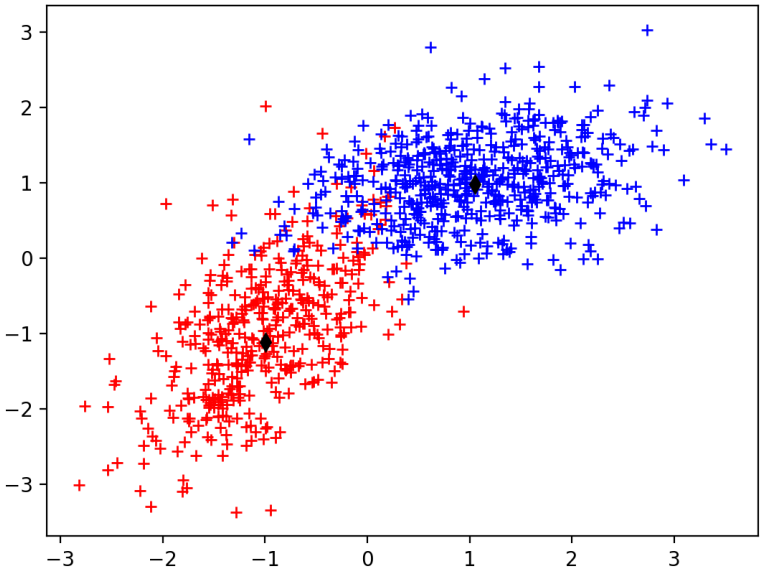

Figure 2: Scatter plot with MLE estimation

2B3

For a given precinct

$$
\begin{aligned}
& \sum_D \log P(X_{1\ldots M}) = \sum_D \log \sum_Y \sum_Z P(X_{1\ldots M}, Y, Z_{1\ldots M}) \\
= & \sum_D \log \sum_Y P(Y) \prod_{j=1\ldots M} \sum_{Z_j} P(Z_j \mid Y) * P(X_j \mid Z_j) \\
= & \sum_D \log \sum_Y P(Y) \prod_{j=1\ldots M} \sum_{Z_j} P(Z_j \mid Y) * P(X_j \mid Z_j)
\end{aligned}
$$

Where

$$
\begin{aligned}
& P(Y = 1) = \phi \\
& P(Z_j \mid Y) = 1(Z = Y) * \lambda + 1(Z \neq Y) * (1 - \lambda) \\
& P(X_j \mid Z_j) = N(X \mid \mu_{Z_j}, \sigma_{Z_j})
\end{aligned}
$$

For $P(Y, Z \mid X)$

$$
\begin{aligned}
& P(Y, Z_{1\ldots M} \mid X_1 \ldots M) \\
= & \frac{P(Y, Z_{1\ldots M}, X_1 \ldots M)}{P(X_1 \ldots M)} \\
= & \frac{P(Y) \prod_z p(x_j \mid z_j) * p(z_j \mid y_j)}{\sum_Y P(Y) \prod_z \sum_Z p(x_j \mid z_j) * p(z_j \mid y_j)}
\end{aligned}
$$

This is computable for similar reasons to the calculations above.

M step:

μ, σ have very similar update equations as before.

$$
\begin{aligned}
\phi &= \frac{\sum_N P(Y_n \mid X_{n,1\ldots M})}{N} \\
\lambda &= \frac{\sum_{i,j} P(Y_i = 0, Z_{ij} = 0 \mid X_{n,1\ldots M}) + P(Y_i = 1, Z_{ij} = 1 \mid X_{n,1\ldots M})}{M * N}
\end{aligned}
$$

2B4

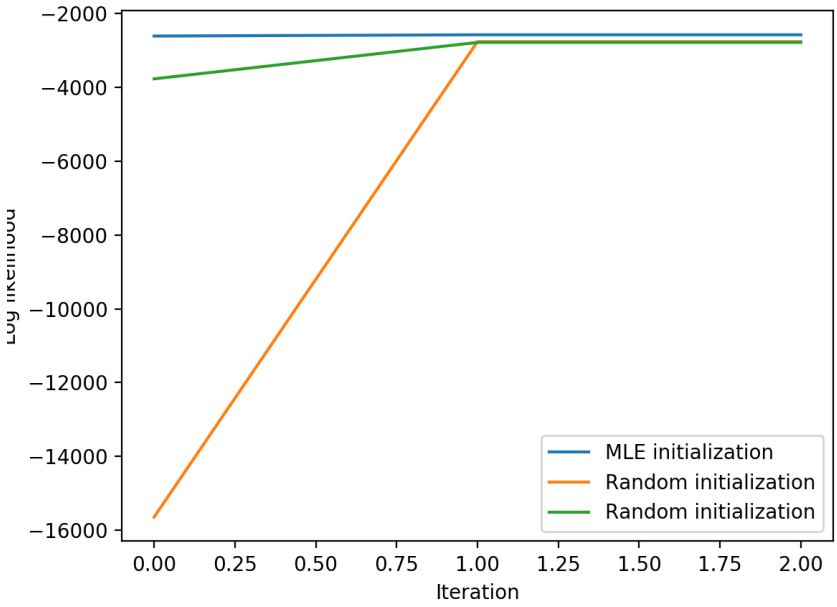

Figure 3: Log likelihood functions

MLE start:

```
{'phi': 0.5600018439242308, 'sigma_1': matrix([[ 0.69145661,  0.13607514],
[ 0.13607514,  0.29779632]]), 'sigma_0': matrix([[ 0.39440707,  0.35034264],
[ 0.35034264,  0.84212032]]), 'mu_1': matrix([[ 1.0205135 ,  1.00876657]]),
'mu_0': matrix([[-1.00910006, -0.96364041]]),
'pi': 0.57, 'lambda': 0.88486685305626767}
```

Random 1 start:

```
{'phi': 0.4475736376195681, 'sigma_1': matrix([[ 0.87820589,  0.47300381],
[ 0.47300381,  0.40278968]]), 'sigma_0': matrix([[ 1.98996833,  1.45905695],
[ 1.45905695,  1.32032964]]), 'mu_1': matrix([[ 0.40618372,  1.03386316]]),
'mu_0': matrix([[-0.03985263, -0.46872801]]), 'pi': 0.41325060979256356, 'lambda': 0.67641
```

Random 2 start:

```
{'phi': 0.9999999480057111, 'sigma_1': matrix([[ 1.57698609,  1.23182323],
[ 1.23182323,  1.5197048 ]]), 'sigma_0': matrix([[ 1.43657067,  0.35393284],
[ 0.35393284,  0.08890104]]), 'mu_1': matrix([[ 0.14888773,  0.16911763]]),
'mu_0': matrix([[ 0.06237175, -0.20027122]]), 'pi': 0.362784982492773,
'lambda': 0.9740310467641029}
```

2B5

0	1.0	✓
1	1.0	✓
2	5.2910748229e-10	
3	0.999999999922	✓
4	4.10560135051e-13	
5	0.999999999107	✓
6	0.999999999999	✓
7	0.999999999999	✓
8	1.0	✓
9	1.0	✓
10	1.28149768981e-09	
11	9.54600552997e-09	
12	1.76548240647e-12	
13	0.999999999993	✓
14	0.999999963276	✓
15	1.37480619424e-10	
16	3.81772639608e-12	
17	1.0	✓
18	1.0	✓
19	0.999999999999	✓
20	1.83143619726e-13	
21	1.0	✓
22	0.999999972651	✓
23	1.14540895386e-09	
24	0.999999999983	✓
25	6.77826060315e-09	
26	5.0424198135e-13	
27	3.71326354912e-12	
28	1.0	✓
29	5.644032623e-11	
30	1.0	✓
31	1.0	✓
32	0.999999999998	✓
33	6.01396355796e-10	
34	0.999999999994	✓
35	7.64259227773e-11	
36	1.0	✓
37	1.7347678872e-11	
38	2.20629791596e-11	
39	0.999999999839	✓
40	2.97420189651e-12	
41	0.999999999934	✓
42	1.0	✓
43	3.23941817644e-12	
44	0.999999999985	✓
45	5.57364314051e-09	
46	0.999999989351	✓
47	1.71714723668e-11	
48	0.999999990859	✓
49	5.53677858465e-05	

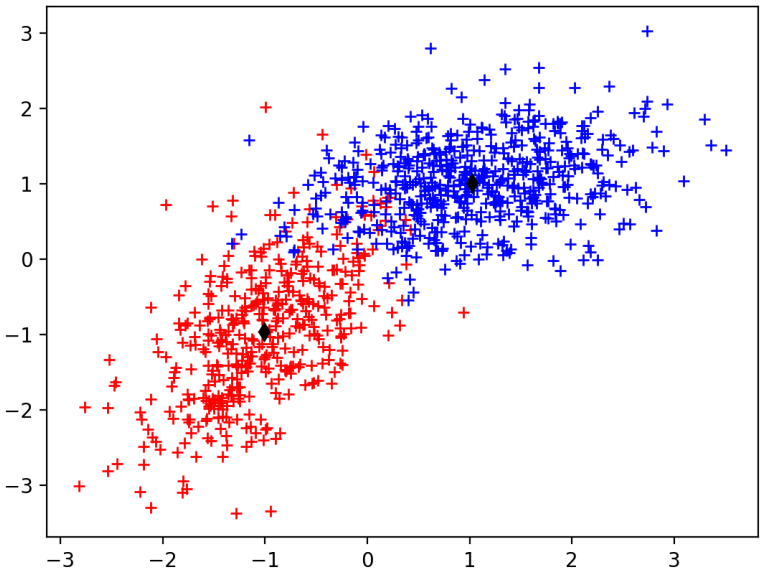

Figure 4: Scatter plot after EM