CS228 Homework 5 Solutions

Instructor: Stefano Ermon – ermon@stanford.edu

Available: 03/3/2017; Due: 03/18/2017

---

1. **[25 points]** (**Bayesian inference**) Let $X \in \{x^1, ..., x^K\}$ be a multinomial variable and let θ be a parametrization of the distribution of X, i.e. $P(X = x^k|\theta) = \theta_k$. Let $\mathcal{D} = \{x[1], ..., x[M]\}$ be a dataset consisting of M realizations of X. We would like to infer something interesting about θ based on $\mathcal{D}$.

   Our strategy so far has been to infer a true set of parameters θ^* from which the data was generated; we found θ^* using the principle of maximum likelihood; this was an example of the so-called *frequentist* approach to statistics. In *Bayesian* statistics, we instead construct a *posterior* distribution $P(\theta|\mathcal{D})$ that can be used to describe our uncertainty over the parameters given the evidence we observed. Recall that we will construct $P(\theta|\mathcal{D})$ using Bayes theorem:

   $$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}.$$

   This approach has the advantage of better modeling the full uncertainty over the parameters. However, the probabilities no longer correspond to limiting frequencies within a data-generating process described by $P(\mathcal{D}|\theta)$. Instead, they can only be interpreted as "beliefs". Most crucially, these probabilities can be arguably called "subjective" because they depend on a set of arbitrary initial beliefs specified by $P(\theta)$. This may raise objections, since we might want our inferences about the world to be independent of any subjective choices by the statistician. It is also not always clear how to specify $P(\theta)$ and how to translate our prior beliefs into probabilities. Arguments like these are part of the great frequentist vs. Bayesian debate in statistics. Here, we will see a concrete example of how the Bayesian approach can be useful.

   (a) **[8 points]** Let's use the posterior to make predictions on new samples. Suppose that the likelihood $P(\mathcal{D}|\theta) = \prod_{j=1}^{M} P(x[j]|\theta)$ is a product of categorical distributions (i.e. we assume that the observations are independent of each other given θ) and let's choose a Dirichlet prior over θ, i.e. $\theta \sim Dirichlet(\alpha_1, ..., \alpha_K)$. Recall that $\theta \sim Dirichlet(\alpha_1, ..., \alpha_K)$ if $P(\theta) \propto \prod_k \theta_k^{\alpha_k - 1}$.
   Show that the Bayesian predictive probability using a Dirichlet prior is

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

   where $M[i]$ is the number of times $x[m] = x^i$ appears in the dataset and $\alpha = \sum_i \alpha_i$. $X[M+1]$ is assumed conditionally independent of $\mathcal{D}$ given θ.
   Note that this probability has a very neat interpretation: $M[i]/M$ by itself is simply the frequency of class i in our dataset. By adding a Dirichlet prior, we effectively augment our dataset with $\alpha[i]$ "virtual" data points of class i: the predictive probability is the same we would've had it there were $\alpha[i]$ extra points of class i in the actual dataset!
   Hint: Recall from Lecture 15 that the posterior $P(\theta \mid x[1], \cdots, x[M])$ is given by $Dirichlet(\alpha'_1, ..., \alpha'_K)$, where

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

Answers

$$\begin{aligned} P(X[M+1]|\mathcal{D}) &= \int P(X[M+1]|\mathcal{D},\theta)P(\theta|\mathcal{D})\mathrm{d}\theta \\ &= \int P(X[M+1]|\theta)P(\theta|\mathcal{D})\mathrm{d}\theta \\ &= E_{P(\theta|\mathcal{D})}[P(X[M+1]|\theta)] \end{aligned}$$

Using proposition 17.4 from the textbook, we have
$E[\theta_i] = \frac{\alpha_i}{\alpha}$
Our posterior is $Dirichlet(\alpha_1 + M[1], \alpha_2 + M[2], ..., \alpha_k + M[k])$
Hence $P(X[M+1] = x^i|\mathcal{D}) = \frac{\alpha_i + M[i]}{\alpha + M}$

(b) **[8 points]** Now we want to compute the Bayesian predictive probability over two samples. Show how to compute

$$P(X[M+1] = x^i, X[M+2] = x^j|\mathcal{D})$$

Answers

$$\begin{aligned} &P(X[M+1] = x^i, X[M+2] = x^j|\mathcal{D}) \\ &= P(X[M+2] = x^j|X[M+1] = x^i, \mathcal{D})P(X[M+1] = x^i|\mathcal{D}) \\ &= \frac{M[i]+\alpha_i}{M+\alpha} \cdot \frac{M[j] + \mathbb{1}\{x^i = x^j\} + \alpha_j}{M+1+\alpha} \end{aligned}$$

The first step uses chain rule. For the second step, the second term is given by the Bayesian predictive probability above, and the first term is equivalent to adding $X[M+1]$ to the dataset $\mathcal{D}$ and applying the Bayesian predictive probability of the same form.

(c) **[9 points]** Suppose we decide to use the approximation

$$P(X[M+1] = x^i, X[M+2] = x^j|\mathcal{D}) \approx P(X[M+1] = x^i|\mathcal{D}) \cdot P(X[M+2] = x^j|\mathcal{D})$$

That is, we ignore the dependencies between $X[M+1]$ and $X[M+2]$. Analyze the error in this approximation (the ratio between the approximation and the correct probability). What is the quality of this approximation for small M? What is the asymptotic behavior of the approximation when $M \to \infty$.

In general, Bayesian inference may not always be tractable, and often requires approximations such as the one in this question. Finding such approximations is the topic of a large subfield of machine learning which studies the problem of "approximate inference".

Answers

- When $i \neq j$, the ratio is:

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

  Thus, for small M, the independent approximation can yield a bad estimate and overestimates the probability. As $M \to \infty$, the approximation converges to the accurate estimate.

- When $i = j$, the ratio is:

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

  Again, for small M, the independent approximation can yield a bad estimate and as $M \to \infty$, since this implies that $M[j] \to \infty$, the approximation converges to the accurate estimate.
  We gave 1 point of extra credit to those who noted that with large M but small $M[j]$, the approximation is not good in the latter case though it is good in the former case.

2. **[75 points]** Programming Assignment [1]

This homework explores parameter learning in latent variable graphical models in the context of a hypothetical problem involving voter registration.

Suppose you are working as a volunteer on behalf of one of the two major presidential candidates in an evenly divided city in a swing state – lets say, Cleveland, Ohio. Your goal is to register as many voters as possible who are likely to support your candidate. However, your party has a limited number of volunteers, and needs to be wise about how it spends scarce resources in canvassing for voters. Fortunately, a local university recently conducted an extensive survey in which the city was partitioned into fifty precincts, and twenty citizens per precinct were surveyed on their political views. A table has been made publicly available that lists for each respondent:

- The precinct $i = 1, ..., N$. In our city, $N = 50$.
- The index $j = 1, ..., M$ of the respondent within its precinct; in our case, $M = 20$.
- Two summary statistics $X_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)})^T$ indicating respectively the overall social conservatism/liberalism and economic conservatism/liberalism of the j-th respondent in district i.
- In five of the fifty precincts, we have explicit respondent party preferences $Z_{ij} \in \{0, 1\}$.

Your goal will be to use the summary statistics to obtain probabilistic information about party preference for the respondents that have not been explicitly surveyed (missing data).

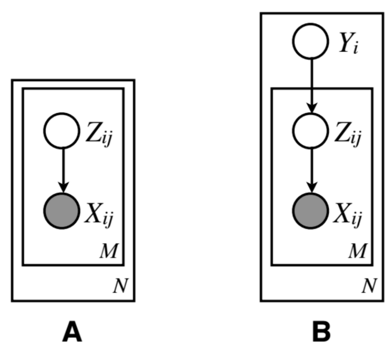


Figure 1: Graphical representation of the models used in this problem.

We will use two models shown in Figure 1 and described below.

(A) **Gaussian mixture model**

We model the X_{ij} as a mixture of two Gaussians $P(X) = (1 - \pi)\mathcal{N}(X|\mu_0, \Sigma_0) + \pi\mathcal{N}(X|\mu_1, \Sigma_1)$. This model tries to capture the following generative story: first, each person independently samples a party preference Z_{ij} from a Bernoulli distribution with parameter π; then, their sample summary statistics are sampled as $X_{ij} \mid z_{ij} \sim \mathcal{N}(\mu_{z_{ij}}, \Sigma_{z_{ij}})$, where $\mu_l = (\mu_l^{(1)}, \mu_l^{(2)})^T$ is the class-conditional mean for party l, and Σ_l is the class-conditional variance/covariance matrix for party l. Note that precinct membership does not factor into this model, despite its use in indexing the variables.

With regards to the above model, answer the following questions.

i **[10 points]** Estimate the parameters π, μ_0, μ_1, Σ_0, and Σ_1 by maximum likelihood using just data from the five precincts in which respondents have reported their party preferences. The data is provided in 'survey-labeled.dat'. Provide an explicit estimator formula for each parameter.
Answers: Estimations from the data (note that there are several ways of writing the sufficient statistics):

[1] Assignment adapted from Cornell's BTRY 6790, instructed by Adam Siepel

Sufficient statistic: $\sum_{i,j} z_{ij}$
$\hat{\pi} = 0.57$
Sufficient statistics: $\sum_{i,j} 1\{z_{ij} = 0\}$, $\sum_{i,j} 1\{z_{ij} = 0\}x_{ij}$, $\sum_{i,j} 1\{z_{ij} = 0\}x_{ij}x_{ij}^\top$
$\hat{\mu}_0 = [-0.99437209 \quad -1.11730233]^\top$
$$\hat{\Sigma}_0 = \begin{bmatrix} 0.30811884 & 0.28553768 \\ 0.28553768 & 0.81346635 \end{bmatrix}$$
Sufficient statistics: $\sum_{i,j} 1\{z_{ij} = 1\}$, $\sum_{i,j} 1\{z_{ij} = 1\}x_{ij}$, $\sum_{i,j} 1\{z_{ij} = 1\}x_{ij}x_{ij}^\top$
$\hat{\mu}_1 = [1.04922807 \quad 0.98085965]^\top$,
$$\hat{\Sigma}_1 = \begin{bmatrix} 0.77827888 & 0.19683566 \\ 0.19683566 & 0.24996938 \end{bmatrix}$$

ii **[15 points]** In this part, we will use the unlabelled dataset 'survey-unlabeled.dat' to learn the model parameters. Implement a Gaussian-Mixtures EM algorithm for model A and use it to estimate $\theta = \{\pi, \mu_0, \mu_1, \sigma_0, \sigma_1\}$.

- The E-step computes $P(z_{ij}|x_{ij}, \theta)$, which is the expected "counts" for every respondent given a fixed value of the model parameters.
- The M-step re-estimates the model parameters, θ based on the expected values calculated in the E-step.

The algorithm should compute and output the log likelihood on every iteration, and should terminate when this quantity increases by less than a value of 0.01 between iterations. Run the algorithm with three different initializations: one equal to your estimates from part 2(A)i and two other (poorer) initializations of your choice. Plot the log likelihood as a function of algorithm iteration for all three cases. Comment on any differences in the local maxima that are found. Report your parameter estimates.

Answers: The E and M step for part (i)
E - step:
$P(z_{ij}|x_{ij}, \theta) \propto P(x_{ij}|z_{ij}, \theta)P(z_{ij}|\theta)$
Normalizing gives the actual value.
M - step:
$$\pi = \frac{\sum_i^N \sum_j^M P(z_{ij}=1|x_{ij}, \theta)}{\sum_i^N \sum_j 1}$$
$$\mu_0 = \frac{\sum_i^N \sum_j^M x_{ij} P(z_{ij}=0|x_{ij}, \theta)}{\sum_i^N \sum_j P(z_{ij}=0|x_{ij}, \theta)}$$
$$\mu_1 = \frac{\sum_i^N \sum_j^M x_{ij} P(z_{ij}=1|x_{ij}, \theta)}{\sum_i^N \sum_j P(z_{ij}=1|x_{ij}, \theta)}$$
$$\Sigma_0 = \frac{\sum_i^N \sum_j^M (x_{ij} - mu_1)^T (x_{ij} - mu_1) P(z_{ij}=0|x_{ij}, \theta)}{\sum_i^N \sum_j P(z_{ij}=0|x_{ij}, \theta)}$$
$$\Sigma_1 = \frac{\sum_i^N \sum_j^M (x_{ij} - mu_1)^T (x_{ij} - mu_1) P(z_{ij}=1|x_{ij}, \theta)}{\sum_i^N \sum_j P(z_{ij}=1|x_{ij}, \theta)}$$
$\hat{\pi} = 0.586$
$\hat{\mu}_0 = [-1.04691 \quad -1.02906]^\top$
$\hat{\mu}_1 = [0.98700 \quad 0.99639]^\top$
$$\hat{\Sigma}_0 = \begin{bmatrix} 0.35868 & 0.30649 \\ 0.30649 & 0.75075 \end{bmatrix} \quad \hat{\Sigma}_1 = \begin{bmatrix} 0.71964 & 0.14373 \\ 0.14373 & 0.30861 \end{bmatrix}$$

(B) **Geography-aware mixture model**
The second model attempts to capture the fact that respondents in the city tend to be geographically separated by party preference, with some precincts showing strong preferences for one party and others showing strong preferences for the other party. In this model, an additional variable $Y_i \in \{0, 1\}$ is introduced for each precinct i, representing that precinct's preferred party. Our new model has the following generative story: first, the Y_i variables are drawn i.i.d. from a Bernoulli distribution with parameter ϕ. Then, the party preferences Z_{ij} as sampled according to

$$p(z_{ij}|y_i) = \begin{cases} \lambda & \text{if } z_{ij} = y_i \\ (1-\lambda) & \text{otherwise} \end{cases}$$

Here, λ is a new parameter that we introduce. Note also that the Z_{ij} variables are conditionally independent given the Y_i variables. Finally, the X_{ij} are sampled as in the previous model.

With regards to the above model, answer the following questions.

i. **[5 points]** We will first estimate the parameters using data from the five precincts in which respondents have reported their party preferences provided in 'survey-labeled.dat'. Specifically, estimate the parameters ϕ and λ by the approximate method of setting each y_i to the consensus (majority) of the corresponding z_{ij} values ($y_i = I(\sum_{j=1}^{M} z_{ij} \geq \frac{M}{2})$), then acting as as if the y_is were observed. Explicitly write out the log likelihood function in terms of ϕ, λ, μ_0, μ_1, Σ_0, and Σ_1. Derive maximum likelihood estimators for ϕ and λ in terms of completely observed x, y, and z variables. Note that the estimates for μ_0, μ_1, Σ_0, and Σ_1 will remain unchanged from the ones estimated for part 2(A)i.

Answers: $\hat{\phi} = \frac{|\{i:y_i=1\}|}{N} = \frac{\sum_i 1\{\sum_j z_{ij} \geq \frac{M}{2}\}}{N}$

$\hat{\lambda} = \frac{\sum_{i,j} 1\{\hat{y}_i = z_{ij}\}}{MN}$

$= \frac{\sum_{i,j} 1\{1\{\sum_j z_{ij} \geq \frac{M}{2}\} = z_{ij}\}}{MN}$

new sufficient statistic:

$\sum_{i,j} 1\{1\{\sum_j z_{ij} \geq \frac{M}{2}\} = z_{ij}\}$

Estimates:

$\hat{\phi} = 0.6$

$\hat{\lambda} = 0.93$

Log likelihood function: The log-likelihood is given by

$$\begin{aligned} \mathcal{L} &= \log p(\mathcal{D}) \\ &= \sum_i \log\left(p(y_i) + \sum_j (\log p(x_{ij}|z_{ij}) + \log(z_{ij}|y_i)) \right) \end{aligned}$$

But, note that we can explicitly write out each factor as follows:

$$\begin{aligned} p(x|z) &= \frac{1}{2\pi|\Sigma_z|^{1/2}} \exp(-(x-\mu_z)^\top \Sigma_z^{-1}(x-\mu_z)/2) \\ p(z|y) &= \lambda^{1\{z=y\}}(1-\lambda)^{1\{z\neq y\}} \\ p(y) &= \phi^{1\{y=1\}}(1-\phi)^{1\{y=0\}}. \end{aligned}$$

We substitute these into the original log-likelihood function, simplify the expression and maximize with respect to each parameter.

ii. **[10 points]** Having estimated your parameters for model B by "supervised" training, use them to analyze the unlabeled data set ('survey-unlabeled.dat') by identifying precincts to be targeted by party 1. Specifically, compute $p(y_i|x_{i,1:M})$ for each precinct i, where $x_{i,1:M} = \{x_{ij} : 1 \leq j \leq M\}$, and identify those precincts for which this probability exceeds 0.5. Summarize your results by presenting a table with one row per precinct, in ascending order by index, a column with the quantity $p(y_i|x_{i,1:M})$, and a mark indicating the precincts that exceed a threshold of 0.5. In addition, compute $p(z_{ij}|x_{i,1:M})$ for every respondent (i,j), and summarize the results by plotting the data points on a two-dimensional plane and coloring them blue if $p(z_{ij} = 1|x_{i,1:M}) > 0.5$ or red otherwise. Also indicate the positions of the two means.

Hint: From the factorized joint distribution of the precinct,

$$p(y_i = 1|x_{i,1:M}) = \frac{p(y_i = 1)\sum_{z_{i,1:M}} \prod_{j=1}^{M} p(x_{ij}|z_{ij})p(z_{ij}|y_i = 1)}{p(x_{i,1:M})}.$$

Answers:

A. $p(y_i = 1|x_{i\cdot})$
$= \frac{p(y_i=1)p(x_{i\cdot}|y_i=1)}{p(x_{i\cdot})}$
$= \frac{p(y_i=1)\sum_{z_{i\cdot}} p(x_{i\cdot}, z_{i\cdot}|y_i=1)}{p(x_{i\cdot})}$
$= \frac{p(y_i=1)\sum_{z_{i\cdot}} \prod_{j=1}^{M} p(x_{ij}|z_{ij})p(z_{ij}|y_i=1)}{p(x_{i\cdot})}$ [The structure of the BN yields this factorization]
$= \frac{p(y_i=1)\prod_{j=1}^{M} \sum_{z_{ij}} p(x_{ij}|z_{ij})p(z_{ij}|y_i=1)}{p(x_{i\cdot})}$ [can push the summation inside due to separability]

B. **table:** 28 precincts out of 50 support party 1.
Computing $p(z_{ij} = 1|x_{i\cdot})$: We marginalize everything but z_{ij}:

$$\begin{aligned}
p(z_{ij}|x_{i\cdot}) &\propto p(x_{i\cdot}, z_{ij}) \\
&= \sum_{y, z_{ij'}: j' \neq j} p(x_{i\cdot}, z_{i\cdot}, y) \\
&= \sum_{y} p(y) \sum_{z_{ij'}: j' \neq j} p(x_{i\cdot}, z_{i\cdot}|y) \\
&= \sum_{y} p(y) \sum_{z_{ij'}: j' \neq j} \prod_{k=1}^{M} p(x_{ik}|z_{ik})p(z_{ik}|y) \\
&= \sum_{y} p(y)p(x_{ij}|z_{ij})p(z_{ij}|y) \prod_{j' \neq j} \sum_{z_{ij'}} p(x_{ij'}|z_{ij'})p(z_{ij'}|y)
\end{aligned}$$

C. **plot:**

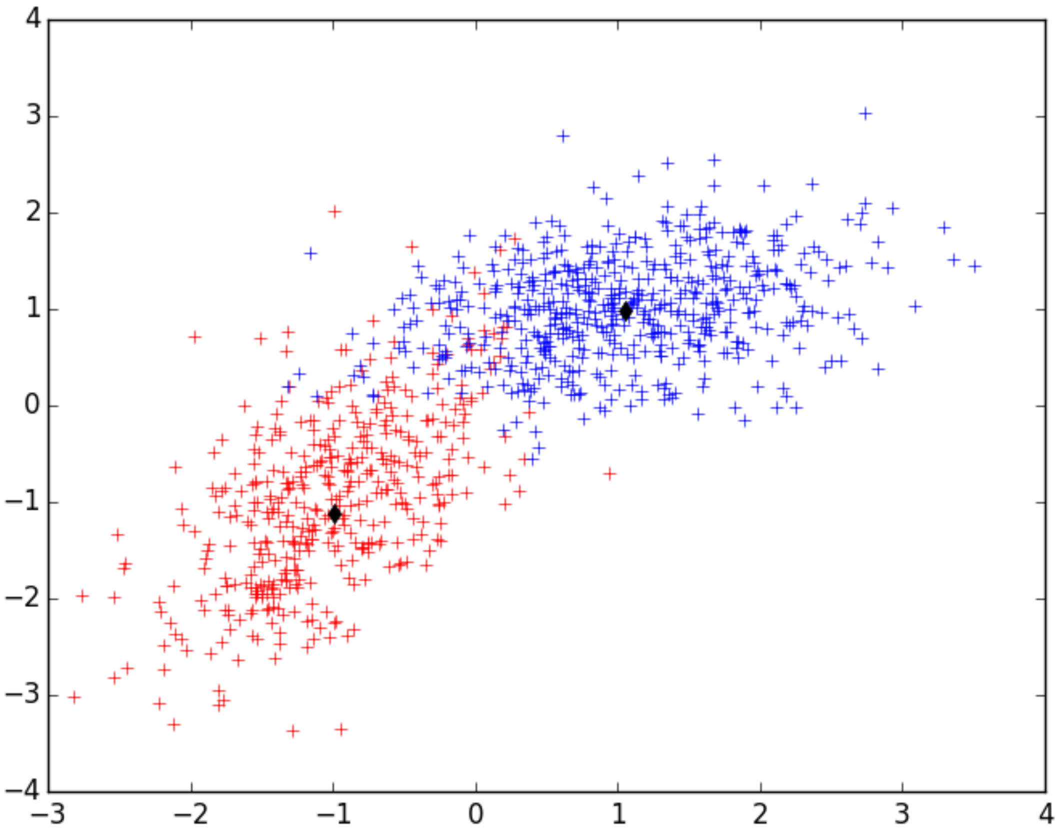


Figure 2: means are marked by diamonds

Precinct	$p(y=1 \mid x_{i\cdot})$	$p \geq 0.5$
0	1.00e+00	1
1	1.00e+00	1
2	1.11e-11	0
3	1.00e+00	1
4	1.80e-16	0
5	1.00e+00	1
6	1.00e+00	1
7	1.00e+00	1
8	1.00e+00	1
9	1.00e+00	1
10	4.11e-11	0
11	2.27e-09	0
12	4.74e-15	0
13	1.00e+00	1
14	1.00e+00	1
15	6.63e-12	0
16	1.95e-14	0
17	1.00e+00	1
18	1.00e+00	1
19	1.00e+00	1
20	2.59e-16	0
21	1.00e+00	1
22	1.00e+00	1
23	2.66e-11	0
24	1.00e+00	1
25	5.72e-10	0
26	8.62e-15	0
27	3.68e-15	0
28	1.00e+00	1
29	1.04e-12	0
30	1.00e+00	1
31	1.00e+00	1
32	1.00e+00	1
33	1.17e-11	0
34	1.00e+00	1
35	2.03e-12	0
36	1.00e+00	1
37	9.35e-14	0
38	1.80e-13	0
39	1.00e+00	1
40	1.14e-14	0
41	1.00e+00	1
42	1.00e+00	1
43	1.45e-14	0
44	1.00e+00	1
45	8.28e-10	0
46	1.00e+00	1
47	4.29e-14	0
48	1.00e+00	1
49	9.22e-05	0

iii. **[15 points]** Now, we will estimate the parameters for model B using the unlabelled dataset 'survey-unlabeled.dat'. Derive E- and M-step updates for model B. Start by writing down the complete log likelihood from part 2(B)i. Let $\theta = \{\phi, \lambda, \mu_0, \mu_1, \Sigma_0, \Sigma_1\}$ denote the model parameters.

- In the E-step, derive an expression for $p(y_i, z_{i,1:M}|x_{i,1:M}, \theta)$, which is the expected value of the relevant "counts" for a fixed value of the parameters. Can this be represented compactly (in a factored form)?
- The M-step re-estimate the model parameters, θ based on the expected values computed in the E-step. As a hint, note that the parameter update equations can be easily computed once we have $p(y_i, z_{ij}|x_{i,1:M}, \theta), p(z_{ij}|x_{i,1:M}, \theta)$, and $p(y_{ij}|x_{i,1:M}, \theta)$.

Hint: You calculations in the M-step should roughly mirror the ones in part 2(B)i. In addition, the E-step will resemble that for model A.

Answers: The E and M step

E - step:

Remember the idea of EM is just finding a lower bound on the log-likelihood (using Jensen's inequality) with a distribution over latent variables and then optimizing over the parameter space of the lower bound. In the below x_i denotes ith row of x.

$$\begin{aligned}
l(D) &= \log p(\mathbf{X}) \\
&= \log \prod_i p(x_i) \\
&= \sum_i \log p(x_i) \\
&= \sum_i \log \sum_{z_{ij}} \sum_{yi} p(x_i, z_{ij}, y_i) \\
&= \sum_i \log \sum_{z_{ij}} \sum_{yi} p(x_i, z_{ij}, y_i) \frac{Q(z_{ij}, y_i|x_i)}{Q(z_{ij}, y_i|x_i)} \\
&= \sum_i \log \mathbb{E}_{[z_{ij}, y_i] \sim Q} \frac{p(x_i, z_{ij}, y_i)}{Q(z_{ij}, y_i|x_i)} \\
&\geq \sum_i \mathbb{E}_{[z_{ij}, y_i] \sim Q} \log \frac{p(x_i, z_{ij}, y_i)}{Q(z_{ij}, y_i|x_i)}
\end{aligned} \tag{1}$$

Now the last inequality (that is, Jensen's inequality) holds with equality when $Q(z_{ij}, y_i|x_i) = p(y_i, z_{ij}|x_i)$. The E-step consists of setting Q to this value.

M - step:

Now in the M step, we maximize the lower bound over the parameter space, let $\theta = [\phi, \lambda, \mu_0, \mu_1, \Sigma_0, \Sigma_1]$

$$\hat{\theta} = \text{argmax}_\theta \sum_i \mathbb{E}_{[z_{ij}, y_i] \sim Q} \log p(x_i, z_{ij}, y_i),$$

where we have dropper the Q in the denominator because it does not affect the value of the bound.

Since p is a Bayesian network, we know that this likelihood is maximized by setting the parameters to their empirical estimates, where examples are weighted according to Q.

In our case, this will require assigning weights according to three probabilities: $p(y_i, z_{ij}|x_i)$,

$p(z_{ij}|x_{i\cdot})$ and $p(y_i|x_{i\cdot})$. We may calculate the first one as follows:

$$\begin{aligned}
p(y_i, z_{ij}|x_{i\cdot}) &\propto p(x_{i\cdot}, z_{ij}, y_i) \\
&= \sum_{z_{ij'}:j'\neq j} p(x_{i\cdot}, z_{i\cdot}, y_i) \\
&= p(y_i) \sum_{z_{ij'}:j'\neq j} p(x_{i\cdot}, z_{i\cdot}|y_i) \\
&= p(y_i) \sum_{z_{ij'}:j'\neq j} \prod_{k=1}^{M} p(x_{ik}|z_{ik})p(z_{ik}|y_i) \\
&= p(y_i)p(x_{ij}|z_{ij})p(z_{ij}|y_i) \prod_{j'\neq j} \sum_{z_{ij'}} p(x_{ij'}|z_{ij'})p(z_{ij'}|y_i)
\end{aligned}$$

The other two probabilities can be easily calculated from the one above.
The estimates of model parameters in terms of these probabilities have the following form:

$\phi = \frac{\sum_i^N P(y_i=1|x_i,\theta)}{\sum_i^N 1}$

$\lambda = \frac{\sum_i^N \sum_j^M P(z_{ij}=1,y_i=1|x_i,\theta)+P(z_{ij}=0,y_i=0|x_i,\theta)}{\sum_i^N \sum_j^M 1}$

$\mu_0 = \frac{\sum_i^N \sum_j^M x_{ij}(P(z_{ij}=0,y_i=1|x_i,\theta)+P(z_{ij}=0,y_i=0|x_i,\theta))}{\sum_i^N \sum_j^M (P(z_{ij}=0,y_i=1|x_i,\theta)+P(z_{ij}=0,y_i=0|x_i,\theta))}$

$\mu_1 = \frac{\sum_i^N \sum_j^M x_{ij}(P(z_{ij}=1,y_i=1|x_i,\theta)+P(z_{ij}=1,y_i=0|x_i,\theta))}{\sum_i^N \sum_j^M (P(z_{ij}=1,y_i=1|x_i,\theta)+P(z_{ij}=1,y_i=0|x_i,\theta))}$

$\Sigma_0 = \frac{\sum_i^N \sum_j^M (x_{ij}-\mu_0)^T(x_{ij}-\mu_0)(P(z_{ij}=0,y_i=1|x_i,\theta)+P(z_{ij}=0,y_i=0|x_i,\theta))}{\sum_i^N \sum_j^M (P(z_{ij}=0,y_i=1|x_i,\theta)+P(z_{ij}=0,y_i=0|x_i,\theta))}$

$\Sigma_1 = \frac{\sum_i^N \sum_j^M (x_{ij}-\mu_1)^T(x_{ij}-\mu_1)(P(z_{ij}=1,y_i=1|x_i,\theta)+P(z_{ij}=1,y_i=0|x_i,\theta))}{\sum_i^N \sum_j^M (P(z_{ij}=1,y_i=1|x_i,\theta)+P(z_{ij}=1,y_i=0|x_i,\theta))}$

iv. **[10 points]** Based on the above updates, implement an EM algorithm for model B. Run the algorithm with three different initializations, plot the log likelihood and report the parameter estimates.
Answers:

v. **[10 points]** Using your best set of parameter estimates (those yielding the highest log likelihood) again compute $p(y_i = 1|x_{i,1:M})$ for each precinct i, and identify those precincts for which this probability exceeds 0.5. Report a new version of the table from part 2(B)ii. In addition, again compute $p(z_{ij}|x_{i,1:M})$ and prepare a new plot with red and blue data points.

Answers:

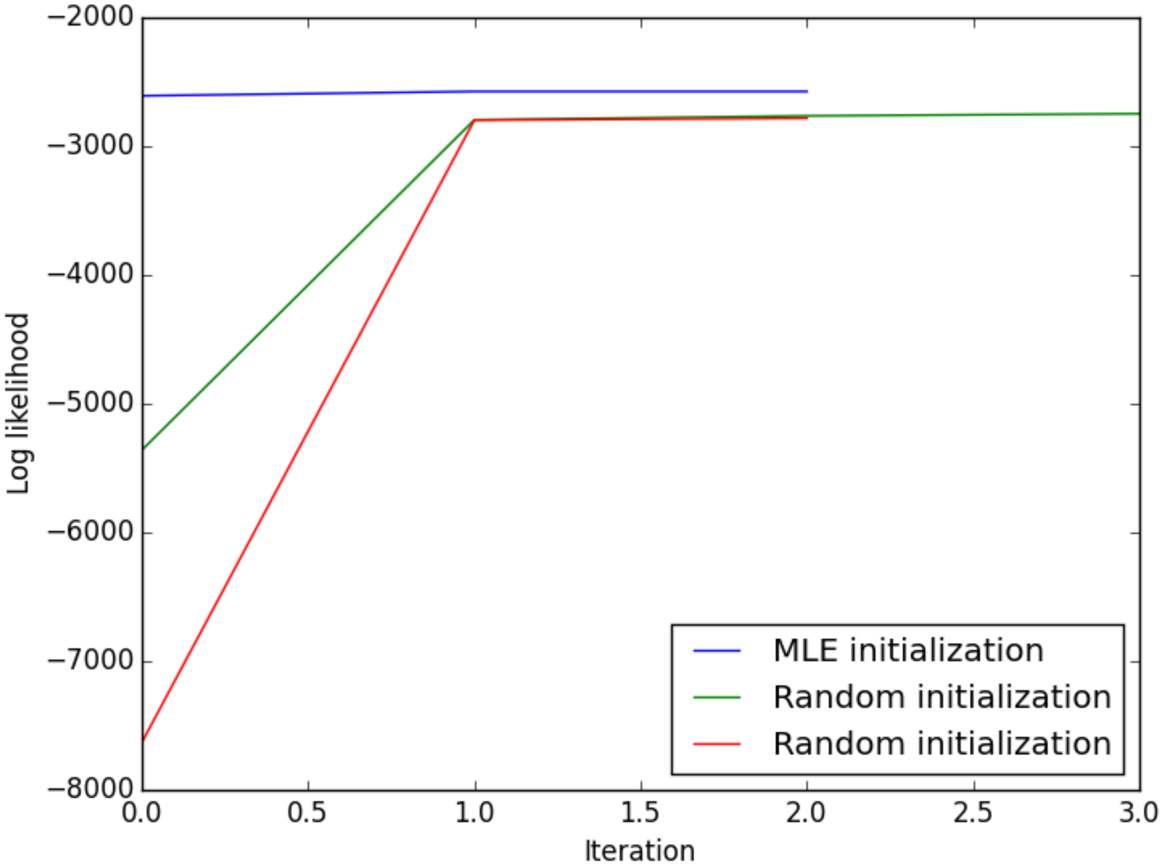
A. $\hat{\phi} = 0.56$
$\hat{\lambda} = 0.8899$
$\hat{\mu}_0 = [-0.98785594 \quad -0.93048333]^\top$
$\hat{\mu}_1 = [1.03634252 \quad 1.01433742]^\top$
$\hat{\Sigma}_0 = \begin{bmatrix} 0.42052449 & 0.38219896 \\ 0.38219896 & 0.87710661 \end{bmatrix}$ $\hat{\Sigma}_1 = \begin{bmatrix} 0.67681216 & 0.12857611 \\ 0.12857611 & 0.30105244 \end{bmatrix}$

B. **table:** 28 precincts out of 50 support party 1.

C. **plot:**

Hints:

(a) It is important that you understand EM algorithm well in order to do this programming assignment. One good note on EM algorithm is : `https://people.csail.mit.edu/rameshvs/content/gmm-em.pdf`.



(b) One common source of mistake in part b is how you calculate the log of product-sum. Note that:

$$log(\prod_i \sum_j f(x_{ij})) = \sum_i log(\sum_j f(x_{ij}))$$

This is not equivalent to:

$$log(\prod_i \sum_j f(x_{ij})) = \sum_i \sum_j log(f(x_{ij}))$$

(c) Be careful on how you use numpy's reshape and transpose function when you write code for log-likelihood.

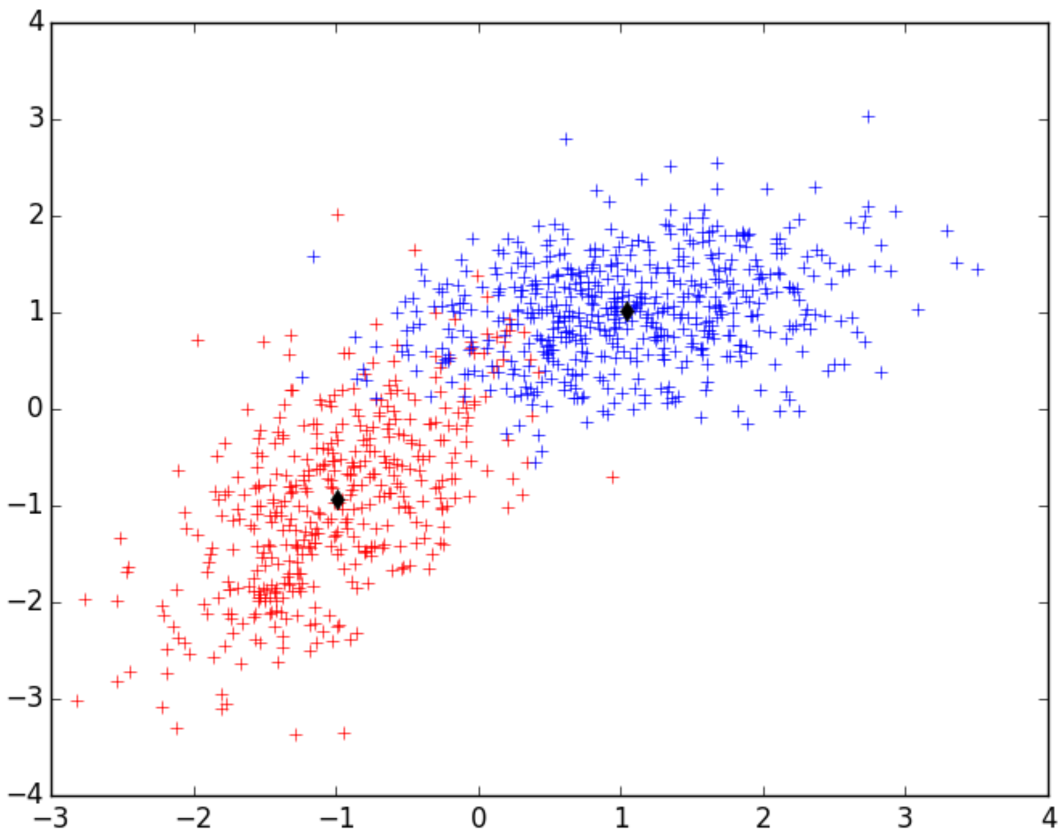


Figure 3: means are marked by diamonds

Precinct	$p(y=1\|x_{i\cdot})$	$p \geq 0.5$
0	1.00e+00	1
1	1.00e+00	1
2	2.03e-10	0
3	1.00e+00	1
4	1.51e-13	0
5	1.00e+00	1
6	1.00e+00	1
7	1.00e+00	1
8	1.00e+00	1
9	1.00e+00	1
10	5.26e-10	0
11	2.94e-09	0
12	6.37e-13	0
13	1.00e+00	1
14	1.00e+00	1
15	4.16e-11	0
16	1.13e-12	0
17	1.00e+00	1
18	1.00e+00	1
19	1.00e+00	1
20	6.84e-14	0
21	1.00e+00	1
22	1.00e+00	1
23	4.16e-10	0
24	1.00e+00	1
25	2.27e-09	0
26	1.33e-13	0
27	1.43e-12	0
28	1.00e+00	1
29	1.98e-11	0
30	1.00e+00	1
31	1.00e+00	1
32	1.00e+00	1
33	2.33e-10	0
34	1.00e+00	1
35	2.64e-11	0
36	1.00e+00	1
37	5.93e-12	0
38	8.14e-12	0
39	1.00e+00	1
40	1.16e-12	0
41	1.00e+00	1
42	1.00e+00	1
43	1.30e-12	0
44	1.00e+00	1
45	1.45e-09	0
46	1.00e+00	1
47	6.71e-12	0
48	1.00e+00	1
49	1.96e-05	0