MACHINE LEARNING COMS 4771, HOMEWORK 4

Name: Yang Bai UNI: yb2356

Problem 1 (10 points): EM Derivation

E-step.

According to Bayesian rules,

$$\tau_{nj} = p(z_n = j|x_n, \theta) = \frac{p(x_n|z_n = j, \theta)p(z_n = j|\theta)}{p(x_n|\theta)}$$

$$\tau_{nj} = \frac{\pi_j \prod_{i=1}^{M} \mu_j(i)^{x_n(i)}}{\sum_{l=1}^{K} \pi_l \prod_{i=1}^{M} \mu_l(i)^{x_n(i)}}$$

M-step.

Set $\theta = \underset{\theta}{\arg\ \max} \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log \frac{p(x_n, z_n = j|\theta)}{\tau_{nj}}$

$$\sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log \frac{p(x_n, z_n = j|\theta)}{\tau_{nj}}$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log p(x_n, z_n = j|\theta) - \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log \tau_{nj}$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log p(x_n, z_n = j|\theta) - const$$

There are some restrictions:

$$\sum_{m=1}^{M} \mu_j(m) = 1, \quad \sum_{j=1}^{K} \pi_j = 1\,, \quad \sum_{i=1}^{M} x(i) = 1$$

Using Lagrange method:

$$let\ Q(\theta) = \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log p(x_n, z_n = j|\theta) - \sum_{j=1}^{K} \lambda_{1j} \left(\sum_{i=1}^{M} \mu_j(i) - 1 \right) - \lambda_2 \left(\sum_{j=1}^{K} \pi_j - 1 \right)$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log \quad \pi_j \prod_{i=1}^{M} \mu_j(i)^{x_n(i)} \quad - \sum_{j=1}^{K} \lambda_{1j} \left(\sum_{i=1}^{M} \mu_j(i) - 1 \right) - \lambda_2 \left(\sum_{j=1}^{K} \pi_j - 1 \right)$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \left[\log(\pi_j) + \sum_{i=1}^{M} x_n(i) \log\left(\mu_j(i)\right) \right] - \sum_{j=1}^{K} \lambda_{1j} \left(\sum_{i=1}^{M} \mu_j(i) - 1 \right) - \lambda_2 \left(\sum_{j=1}^{K} \pi_j - 1 \right)$$

$$\frac{\partial Q(\theta)}{\partial \mu_j(i)} = \frac{\sum_{n=1}^{N} \tau_{nj}\, x_n(i)}{\mu_j(i)} - \lambda_{1j} = 0$$

$$\Longrightarrow \mu_j(i) = \frac{\sum_{n=1}^{N} \tau_{nj}\, x_n(i)}{\lambda_{1j}}$$

$$\sum_{i=1}^{M} \mu_j(i) = 1, so \quad \frac{\sum_{i=1}^{M} \sum_{n=1}^{N} \tau_{nj}\, x_n(i)}{\lambda_{1j}} = 1, \quad \lambda_{1j} = \sum_{i=1}^{M} \sum_{n=1}^{N} \tau_{nj}\, x_n(i)$$

$$so \quad \mu_j(i)^{(t+1)} = \frac{\sum_{n=1}^{N} \tau_{nj}{}^{(t)}\, x_n(i)}{\sum_{n=1}^{N} \tau_{nj}{}^{(t)}}$$

$$\frac{\partial Q(\theta)}{\partial \pi_j} = \frac{\sum_{n=1}^{N} \tau_{nj}}{\pi_j} - \lambda_2 = 0$$

$$\pi_j = \frac{\sum_{n=1}^{N} \tau_{nj}}{\lambda_2}$$

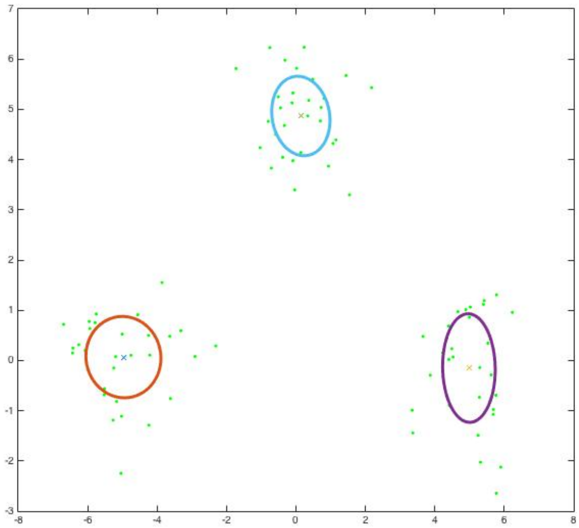
$$\sum_{j=1}^{K} \pi_j = 1, so \sum_{j=1}^{K} \frac{\sum_{n=1}^{N} \tau_{nj}}{\lambda_2} = 1, \qquad \lambda_2 = \sum_{j=1}^{K} \sum_{n=1}^{N} \tau_{nj}$$

$$so \quad \pi_j{}^{(t+1)} = \frac{\sum_{n=1}^{N} \tau_{nj}{}^{(t)}}{\sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj}{}^{(t)}} = \frac{\sum_{n=1}^{N} \tau_{nj}{}^{(t)}}{\sum_{n=1}^{N} 1} = \frac{\sum_{n=1}^{N} \tau_{nj}{}^{(t)}}{N}$$

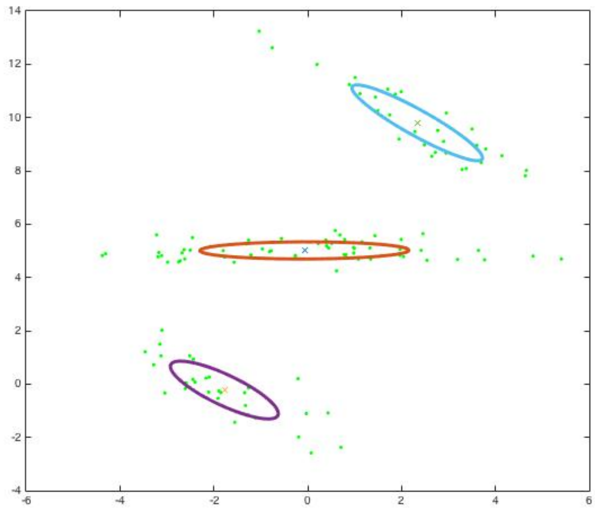
Problem 2 (20 points): EM for Bernoulli Mixtures

Part A:

The result for dataA



Result for dataB:



Part B:

In this problem, X is a vector of 50 dimension. Xn represents the n th data, and Xn(i) represents the i th dimension of n th data. All Xn(i) (i=1,2,…,50) are iid

Bernoulli(μ_n). $p(X_n|z_n = j, \theta) = \prod_{i=1}^{50} \mu_j{}^{X_n(i)} (1 - \mu_j)^{1 - X_n(i)}$

Treat Bernoulli distribution as multinomial distribution. So $p(X_n|z_n = j, \theta) = \prod_{i=1}^{50} \prod_{m=1}^{2} \ \mu_j(m)^{X_n(i)(m)}$, where $\mu_j(1) = \mu_j, X_n(i)(1) = X_n(i) \ \ (m = 1) \ and \ \ \mu_j(2) = 1 - \mu_j \ , \ X_n(i)(2) = 1 - X_n(i)(m = 2)$.

E-step.

According to Bayesian rules,

$$\tau_{nj} = p(z_n = j|x_n, \theta) = \frac{p(x_n|z_n = j, \theta)p(z_n = j|\theta)}{p(x_n|\theta)}$$

$$\tau_{nj} = \frac{\pi_j \prod_{i=1}^{50} \mu_j{}^{x_n(i)} (1-\mu_j)^{1-x_n(i)}}{\sum_{l=1}^{K} \pi_l \prod_{i=1}^{50} \mu_l{}^{x_n(i)} (1-\mu_l)^{1-x_n(i)}}$$

$$\tau_{nj} = \frac{\pi_j \mu_j{}^{\sum_{i=1}^{50} x_n(i)} (1-\mu_j)^{\sum_{i=1}^{50} (1-x_n(i))}}{\sum_{l=1}^{K} \pi_l \, \mu_j{}^{\sum_{i=1}^{50} x_n(i)} (1-\mu_j)^{\sum_{i=1}^{50} (1-x_n(i))}}$$

M-step.

Set $\theta = \underset{\theta}{\arg\ \max} \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log \frac{p(x_n, z_n = j|\theta)}{\tau_{nj}}$

$$\sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log \frac{p(x_n, z_n = j|\theta)}{\tau_{nj}}$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log p(x_n, z_n = j|\theta) - \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log \tau_{nj}$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log p(x_n, z_n = j|\theta) - const$$

There are some restrictions:

$$\sum_{m=1}^{2} \mu_j(m) = 1, \quad \sum_{j=1}^{K} \pi_j = 1\,, \quad \sum_{m=1}^{2} Xn(m) = 1$$

Using Lagrange method:

$$let\ Q(\theta) =$$

$$\sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log p(x_n, z_n = j|\theta) - \sum_{j=1}^{K} \lambda_{1j} \left(\sum_{i=1}^{M} \mu_j(i) - 1 \right) - \lambda_2 \left(\sum_{j=1}^{K} \pi_j - 1 \right)$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log \left[\pi_j \prod_{i=1}^{50} \prod_{m=1}^{2} \mu_j(m)^{X_n(i)(m)} \right] - \sum_{j=1}^{K} \lambda_{1j} \left(\sum_{m=1}^{M} \mu_j(m) - 1 \right)$$

$$- \lambda_2 \left(\sum_{j=1}^{K} \pi_j - 1 \right)$$

$$= \sum_{n=1}^{N}\sum_{j=1}^{K}\tau_{nj}\left[\log(\pi_j) + \sum_{i=1}^{50}\sum_{m=1}^{M} x_n(i)(m)\log\left(\mu_j(m)\right)\right] - \lambda_2\left(\sum_{j=1}^{K}\pi_j - 1\right) - \sum_{j=1}^{K}\lambda_{1j}\left(\sum_{m=1}^{M}\mu_j(m) - 1\right)$$

$$\frac{\partial Q(\theta)}{\partial \mu_j(m)} = \frac{\sum_{n=1}^{N}\tau_{nj}\sum_{i=1}^{50} x_n(i)(m)}{\mu_j(m)} - \lambda_{1j} = 0$$

$$\Longrightarrow \mu_j(m) = \frac{\sum_{n=1}^{N}\tau_{nj}\sum_{i=1}^{50} x_n(i)(m)}{\lambda_{1j}}$$

$$\sum_{i=1}^{M}\mu_j(i) = 1, \qquad so \ \frac{\sum_{i=1}^{M}\sum_{n=1}^{N}\tau_{nj}\sum_{i=1}^{50} x_n(i)(m)}{\lambda_{1j}} = 1,$$

$$\lambda_{1j} = \sum_{i=1}^{M}\sum_{n=1}^{N}\tau_{nj}\sum_{i=1}^{50} x_n(i)(m)$$

$$\mu_j(m)^{(t+1)} = \frac{\sum_{n=1}^{N}\tau_{nj}\sum_{i=1}^{50} x_n(i)(m)}{\sum_{i=1}^{M}\sum_{n=1}^{N}\tau_{nj}\sum_{i=1}^{50} x_n(i)(m)}$$

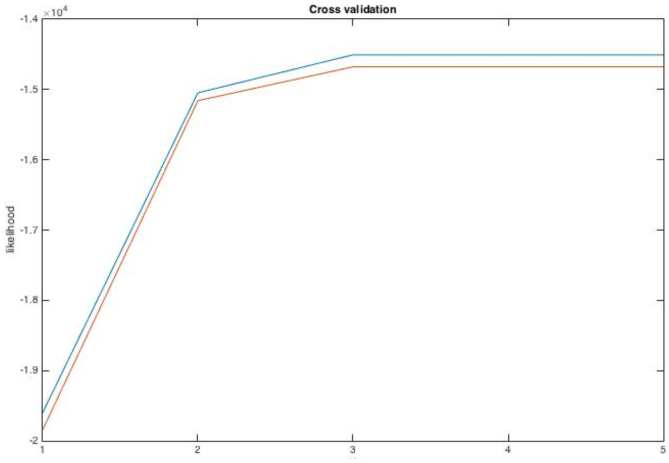
because $\sum_{m=1}^{M} x_n(i)(m) = 1$, so

$$\mu_j(m)^{(t+1)} = \frac{\sum_{n=1}^{N}\tau_{nj}\sum_{i=1}^{50} x_n(i)(m)}{50\sum_{n=1}^{N}\tau_{nj}}$$

${\pi_j}^{(t+1)}$ is the same as problem 1.

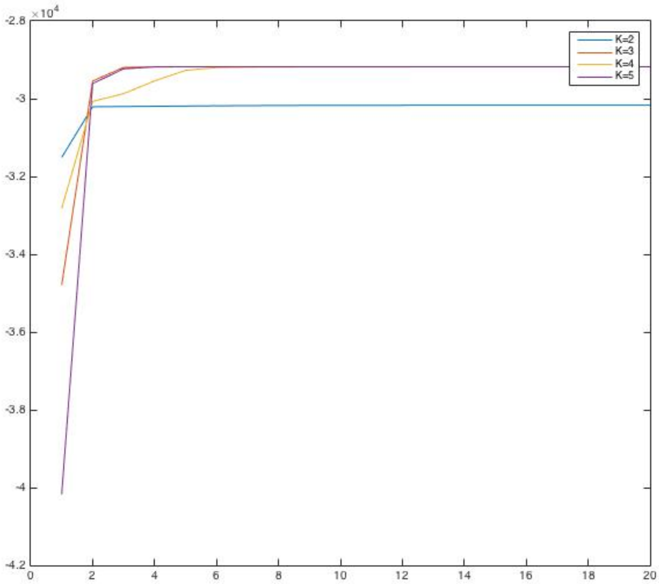
$${\pi_j}^{(t+1)} = \frac{\sum_{n=1}^{N}{\tau_{nj}}^{(t)}}{\sum_{n=1}^{N}\sum_{j=1}^{K}{\tau_{nj}}^{(t)}} = \frac{\sum_{n=1}^{N}{\tau_{nj}}^{(t)}}{\sum_{n=1}^{N} 1} = \frac{\sum_{n=1}^{N}{\tau_{nj}}^{(t)}}{N}$$

Cross validation:



K =3 is the best fit.

The convergence process of likelihood of different K.



Log likelihood:

1) Training

	K=1		K=2		K=3		K=4		K=5	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
1	-18823	-15224	-15124	-15065	-14632	-14551	-14632	-14550	-14632	-14550
2	-14565	-17753	-15106	-15120	-14584	-14598	-14583	-14599	-14584	-14598
3	-19838	-19605	-15041	-15131	-14563	-14619	-14563	-14618	-14563	-14619
4	-19778	-19654	-15131	-15047	-14622	-14565	-14622	-14565	-14622	-14565
5	-20636	-19323	-15161	-15012	-14647	-14534	-14647	-14534	-14646	-14535
6	-17486	-14770	-15134	-15039	-14660	-14522	-14660	-14522	-14660	-14522
7	-19221	-20749	-15068	-15108	-14537	-14646	-14537	-14646	-14537	-14646
8	-17422	-14797	-15193	-14982	-14751	-14432	-14751	-14432	-14750	-14434
9	-14569	-17666	-14981	-15191	-14543	-14641	-14542	-14642	-14543	-14641

10	-20466	-19372	-15122	-15059	-14669	-14513	-14669	-14513	-14669	-14513
mean	-18280	-17891	-15106	-15075	-14621	-14562	-14621	-14562	-14621	-14562
std	22.39	22.34	61.33	62.04	66.12	66.29	66.26	66.49	65.87	65.99

Mean μ when K=3:
0.203 0.504 0.801
Standard deviation of μ when K=3
0.0029 0.0047 0.0027

Mean π when K=3:
0.3118 0.3282 0.3599
Standard deviation of π when K=3
0.0128 0.0099 0.0123

Here is a result table of one random initialization.

	$\pi 1$	$\pi 2$	$\pi 3$	$\pi 4$	$\pi 5$
K=1	1				
K=2	0.580	0.420			
K=3	0.316	0.359	0.325		
K=4	0.359	0.316	0.032	0.293	
K=5	0.316	0.325	0.087	0.128	0.144

	$\mu 1$	$\mu 2$	$\mu 3$	$\mu 4$	$\mu 5$
K=1	0.105				
K=2	0.679	0.233			
K=3	0.509	0.203	0.794		
K=4	0.203	0.509	0.797	0.793	
K=5	0.509	0.794	0.203	0.205	0.203

Problem 3 (10 points): K-Means for image segmentation

K=2

$$\mu = \begin{matrix} 0.2760 & 0.2654 & 0.1632 \\ 0.7275 & 0.8108 & 0.8699 \end{matrix}$$

K=3

$$\mu = \begin{matrix} 0.4672 & 0.4272 & 0.3521 \\ 0.1957 & 0.2010 & 0.0925 \\ 0.7469 & 0.8436 & 0.9104 \end{matrix}$$

K=4

$$\mu = \begin{matrix} 0.4654 & 0.4128 & 0.3198 \\ 0.1957 & 0.2010 & 0.0925 \\ 0.9135 & 0.9556 & 0.9712 \\ 0.5375 & 0.6867 & 0.8100 \end{matrix}$$

K=5

$$\mu = \begin{matrix} 0.9091 & 0.9481 & 0.9604 \\ 0.5089 & 0.6940 & 0.8367 \\ 0.3936 & 0.4294 & 0.3967 \\ 0.6045 & 0.3298 & 0.1795 \\ 0.1738 & 0.2070 & 0.0936 \end{matrix}$$

K=6

$$\mu = \begin{matrix} 0.6594 & 0.5483 & 0.4825 \\ 0.1726 & 0.2074 & 0.0925 \\ 0.6264 & 0.7839 & 0.8948 \\ 0.9245 & 0.9634 & 0.9801 \\ 0.3943 & 0.6507 & 0.8275 \\ 0.4366 & 0.3469 & 0.2681 \end{matrix}$$

As we randomly initialize the K means, we might have some numerical inconsistencies. For example, sometimes one of the K different values of μ becomes NaN, because there is no data point near this μ and z(i) becomes 0. Then μ_i becomes NaN.

$$\vec{\mu}_i = \sum_{n=1}^{N} \vec{x}_n \vec{z}_n(i) \Big/ \sum_{n=1}^{N} \vec{z}_n(i)$$

This problem appears when some of the initialized μ are very far away from data points so no point is classified into their clusters. To avoid this mistake, we can initialize μ by choosing K points randomly from the dataset, so there is at least one point in every cluster(the μ_i itself). Under this circumstance, no z(i) will become zero and no μ_i will become NaN.

Problem 4 (10 points): Jensen's inequality

a) The arithmetic mean of non-negative numbers is at least their geometric mean.

$$\text{arithmetic mean : } f_1(x) = \sum_{i=1}^{N} p_i x_i$$

$$\text{geometric mean: } f_2(x) = \prod_{i=1}^{N} x_i^{p_i}$$

(1) When there is at least one xi that equals zero: arithmetic mean is the sum of xi so it is non-negative; geometric mean is the multiplication of xi so it equals zero. Thus, the arithmetic mean of non-negative numbers is at least their geometric mean.

(2) When all xi are positive:

$$\ln\ f_1(x) = \ln\left(\sum_{i=1}^{N} p_i x_i\right)$$

$$\ln\ f_2(x) = \sum_{i=1}^{N} p_i \ln(\ x_i)$$

Because ln(x) is concave,

$$\ln\left(\sum_{i=1}^{N} p_i x_i\right) \geq \sum_{i=1}^{N} p_i \ln(\ x_i)$$

$$\ln\ f_1(x)\ \geq\ \ln\ f_2(x)$$

As ln(x) monotonically increases when x>0,

$$f_1(x) \geq f_2(x)$$

which means the arithmetic mean of non-negative numbers is at least their geometric mean.

b)

$$LHS = \sum_{i=1}^{m} \frac{\alpha_i \exp{(\theta^t f_i)}}{\alpha_i} = \sum_{i=1}^{m} \alpha_i \exp{(\theta^t f_i - \ln \alpha_i)}$$

e^x is convex, so

$$\sum_{i=1}^{m} p_i \exp{(x_i)} \geq \exp\left(\sum_{i=1}^{m} p_i x_i\right) \quad when \sum_{i=1}^{m} p_i = 1$$

$$we\ know\ that \sum_{i=1}^{m} \alpha_i = 1, so$$

$$LHS = \sum_{i=1}^{m} \alpha_i \exp(\theta^t f_i - \ln \alpha_i)$$

$$\geq \exp\left(\sum_{i=1}^{m} \alpha_i (\theta^t f_i - ln\, \alpha_i)\right)$$

$$= \exp\left(\theta^t \sum_{i=1}^{m} \alpha_i f_i - \sum_{i=1}^{m} \alpha_i\, ln\, \alpha_i\right) = RHS$$

Appendix

code about K means clustering:

```
% K means clastering
% end loop when no data is classified to a different cluster from
last loop
while (updateNum~=0)

  for i=1:K
   miu2= repmat(miu(i,:),40000,1);
   diff(i,:)=sum( ((im_1D-miu2).^2)');
  end
z=z_new;
z_new=zeros(K,40000);
z_new(find((diff-repmat(min(diff),K,1))==0))=1;
  for i=1:K
    new_miu(i,:)= sum(im_1D.*(repmat(z_new(i,:),3,1)'));
  end

new_miu=new_miu./repmat(sum(z_new')',1,3);
updateNum=nnz(z_new-z);
miu=new_miu;
iteration=iteration+1;
end
```

code for EM algorithm for problem 2 part B:

```
for t=1:Nloop
XX= sum(XTrain');
% E step
Tnj = repmat(pai_old, 1, N).*(repmat(miu_old,1,N).^repmat(XX,
K,1)).*(repmat(1.-miu_old,1,N).^repmat(50-XX, K,1));
T = Tnj./ repmat(sum(Tnj),K,1);

% M step
miu_new = sum( T'.* repmat(XX', 1,K))./sum(T')./50;
pai_new= sum(T')./N;

%update parameters
miu_old=miu_new';
pai_old = pai_new';

end
```