1.

a. since marginal ditribution of X is obtained by summing up the joing distribtion over all possible states of Z $\rightarrow$ $P(X) = \sum_z P(x,z) = \sum_z P(z|x)\,P(z)$

$P(x|z) = \prod_{k=1}^{K} N(x|\mu_k,\Sigma_k)^{z_k}$ $P(z) = \prod_{k=1}^{K} \pi_k^{z_k}$ (given)

$\rightarrow$ $P(x) = \sum_z \left\{ \prod_{k=1}^{K} \pi_k^{z_k} * \prod_{k=1}^{K} N(x|\mu_k,\Sigma_k)^{z_k} \right\} = \sum_z \left\{ \prod_{k=1}^{K} \left[\pi_k N(x|\mu_k,\Sigma_k)\right]^{z_k} \right\}$ (*)

Because $Z^{(k)}$ is a binary vector of size k, with 1 only in kth element and 0 in all others.

In other words, with 0, $\left[\pi_k N(x|\mu_k,\Sigma_k)\right]^{z_k}$ = 1

Thus, we can express the (*) as: $P(x) = \pi_1 N(x|\mu_1,\Sigma_1) + \pi_2 N(x|\mu_2,\Sigma_2) + ... + \pi_k N(x|\mu_k,\Sigma_k)$

$\rightarrow P(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k,\Sigma_k)$.

b. $z_k^n \in \{0,1\}$, its expectation w.r.t $P(z_k^n|x^n) \rightarrow E_{P(z_k^n|x^n)}\left[z_k^n\right] = P(z_k^n|x^n)$

for Bayes Rule, $P(z_k^n|x^n) = \frac{P(z_k^n=1|)P(x^n|z_k^n=1)}{\sum_{j=1}^{K} P(z_j^n=1|)P(x^n|z_j^n=1)}$(*)

$Z^{(k)}$ is a binary vector of size k, with 1 only in kth element and 0 in all others.

$P(z_k^n = 1) = \prod_{k=1}^{K} \pi_k^{z_k} = \pi_k$

$P(x^n|z_k^n) = \prod_{k=1}^{K} N(x^n|\mu_k,\Sigma_k)^{z_k^n} = N(x^n|\mu_k,\Sigma_k)^{z_k^n}$

Thus, we can express (*) as:

(*) = $\frac{\pi_k N(x^n|\mu_k,\Sigma_k)^{z_k^n}}{\sum_{j=1}^{K} \pi_j N(x^n|\mu_j,\Sigma_j)^{z_j^n}}$ with fixed value of $\pi_k, \mu_k, \Sigma_k$.

c. *we know that* $P(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k,\Sigma_k)$, *given* $N$, *iid data samples, the log likelihood function*

$L = E_{P(z|x)} log \prod_{i=1}^{N} p(x^i,z^i|\pi,\mu,\Sigma) = \sum_{i=1}^{N} E_{P(z|x)} log\left[\pi_k N(x^i|\mu_z,\Sigma_z)\right]$

since $N(x^i|\mu_z,\Sigma_z) = \frac{1}{(2\pi)^{n/2}|\Sigma_z|^{1/2}} exp[-\frac{1}{2}(x^i-\mu_z)^T \Sigma_z^{-1}(x^i-\mu_z)]$, then

L = $\sum_{i=1}^{N}\sum_{k=1}^{K} \tau_k^i [log\pi_k - n/2log(2\pi) - 1/2log|\Sigma_k| - 1/2(x^i-\mu_k)^T \Sigma_k^{-1}(x^i-\mu_k)]$

= $\sum_{i=1}^{N}\sum_{k=1}^{K} \tau_k^i [log\pi_k - n/2log(2\pi) - 1/2log|\Sigma_k| - 1/2\ trace[(x^i-\mu_k)^T \Sigma_k^{-1}(x^i-\mu_k)]]$

we hope to maximize L subject to $\sum_{k=1}^{K} \pi_k = 1$, using lagrange multiplier, we can transform to unconstrained optimization problem:

$L' = \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_k^i [log\pi_k - n/2log(2\pi) - 1/2log|\Sigma_k| - 1/2\ trace[(x^i-\mu_k)^T \Sigma_k^{-1}(x^i-\mu_k)]] + \lambda(1 - \sum_{k=1}^{K} \pi_k)$

Firstly, set the derivative of L' w.r.t $\mu_k$ to zero

$\sum_{i=1}^{N} \tau_k^i \Sigma_k^{-1}(x^i-\mu_k) = 0 \rightarrow \mu_k = \frac{\sum_{i=1}^{N} \tau_k^i x^i}{\sum_{i=1}^{N} \tau_k^i}$

Then, we set the derivative of L' w.r.t $\Sigma_k$ to zero,

$\sum_{i=1}^{N} \tau_k^i \Sigma_k^T - \sum_{i=1}^{N} \tau_k^i (x^i-\mu_k)^T (x^i-\mu_k) = 0$

( $\frac{\delta L'}{\delta \Sigma_k^{-1}} log|\Sigma_k| = -\Sigma_k^T$, $\frac{\delta L'}{\delta \Sigma_k^{-1}} trace[(x^i-\mu_k)^T \Sigma_k^{-1}(x^i-\mu_k)] = \frac{\delta L'}{\delta \Sigma_k^{-1}} trace[\Sigma_k^{-1}(x^i-\mu_k)(x^i-\mu_k)^T] = (x^i-\mu_k)(x^i-\mu_k)^T$ )

$\rightarrow \Sigma_k = \frac{\sum_{i=1}^{N} \tau_k^i (x^i-\mu_k)(x^i-\mu_k)^T}{\sum_{i=1}^{N} \tau_k^i}$

Finally, we set the derivative of L' w.r.t $\pi_k$ to 0:

$\sum_{i=1}^{N} \frac{N(x^i|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x^i|\mu_j,\Sigma_j)} - \lambda = 0$

here we multiply $\pi_k$ to both sides

$\sum_{i=1}^{N} \frac{\pi_k N(x^i|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x^i|\mu_j,\Sigma_j)} - \lambda\pi_k = 0$ (*)

and then sum over k

$\sum_{i=1}^{N} \frac{\sum_{k=1}^{K} \pi_k N(x^i|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x^i|\mu_j,\Sigma_j)} - \lambda \sum_{k=1}^{K} \pi_k = 0 \rightarrow \lambda = N$

replace $\lambda$ with N into (*) we can get $\pi_k = \frac{\sum_{i=1}^{N} \tau_k^i}{N}$, since $\tau_k^i = \frac{\pi_k N(x^i|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x^i|\mu_j,\Sigma_j)}$.

d.

since all components have covariance $\varepsilon I$, $N(x^i|\mu_k,\Sigma_k) = \frac{1}{(2\pi)^{m/2}\varepsilon^{n/2}} exp[-\frac{1}{2\varepsilon}(x^i-\mu_k)^T(x^i-\mu_k)]$

$\tau_k^i = \frac{\pi_k N(x^i|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x^i|\mu_j,\Sigma_j)} = \frac{\pi_k exp[-\frac{1}{2\varepsilon}(x^i-\mu_k)^T(x^i-\mu_k)]}{\sum_{j=1}^{K} \pi_j exp[-\frac{1}{2\varepsilon}(x^i-\mu_j)^T(x^i-\mu_j)]}$

when taking limit $\varepsilon \rightarrow 0$, in denominator, the term for which $(x^i-\mu_k)^T(x^i-\mu_k)$ (closest to $\mu_k$) is smallest will go to 0 most slowly, hence the posterior possibility $\tau_k^i$ for $x^i$ will go to 0 except for j which $\tau_k^j$ will go to 1. This is independent of $\pi_j$. So, $\tau_k^i$ can be considered as $r_{ik} \in \{0,1\}$.

Log likelihood function:

$$L = \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_k^i [log\pi_k - n/2log(2\pi) - 1/2log\varepsilon - 1/(2\varepsilon)(x^i-\mu_k)^T(x^i-\mu_k)]$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} [log\pi_k - n/2log(2\pi) - 1/2log\varepsilon - 1/(2\varepsilon)(x^i-\mu_k)^T(x^i-\mu_k)]$$

here $\sum_{k=1}^{K} r_{ik}[log\pi_k - n/2log(2\pi) - 1/2log\varepsilon]$ is constant, and $\tau_k^i$ is not dependent on $\pi_k$, thus

L = $\sum_{i=1}^{N} \sum_{k=1}^{K} [-\frac{1}{(2\varepsilon)} r_{ik}(x^i-\mu_k)^T(x^i-\mu_k)] + constant$

Clearly, maximizing the expected complete data log likelihood is similar to minimizing the objective function of K-means: J = $\sum_{i=1}^{N} \sum_{k=1}^{K} [r_{ik}(x^i-\mu_k)^T(x^i-\mu_k)]$

e.

if x is discrete, we can sum over N with corresponding possibility to get expectation:

E[x] = $\sum_{n=1}^{N} x^n p(x^n) = \sum_{n=1}^{N} x^n \sum_{k=1}^{K} \pi_k p(x^n|k) = \sum_{k=1}^{K} \pi_k \sum_{n=1}^{N} x^n p(x^n|k) = \sum_{k=1}^{K} \pi_k \mu_k$

if x is continous, we need to take culumative

E[x] = $\int_x xp(x)dx = \int_x x \sum_{k=1}^{K} \pi_k p(x|k)dx = \sum_{k=1}^{K} \pi_k \int_x xp(x|k)dx = \sum_{k=1}^{K} \pi_k \mu_k$ *since* $\int_x xp(x|k)dx = \mu_k$

Thus E[x] = $\sum_{k=1}^{K} \pi_k \mu_k$

as for covariance:

cov(x) = $E[(x - E[x])(x - E[x])^T] = E[xx^T] - E[x]E[x]^T$ …….(*)

discrete: $E[xx^T] = \sum_x xx^T p(x) = \sum_x xx^T \sum_{k=1}^{K} \pi_k p(x|k) = \sum_{k=1}^{K} \pi_k \sum_x xx^T p(x|k) = \sum_{k=1}^{K} \pi_k E_k[xx^T]$

continuous: $E[xx^T] = \int_x xx^T p(x)dx = \int_x xx^T \sum_{k=1}^{K} \pi_k p(xx^T|k)dx = \sum_{k=1}^{K} \pi_k \int_x xx^T p(xx^T|k)dx = \sum_{k=1}^{K} \pi_k E_k[xx^T]$

Thus, $E[xx^T] = \sum_{k=1}^{K} \pi_k E_k[xx^T]$ for both cases.

For a single component,

$\Sigma_k = E_k[xx^T] - E_k[x]E_k[x]^T \rightarrow E_k[xx^T] = \Sigma_k + E_k[x]E_k[x]^T = \Sigma_k + \mu_k\mu_k^T$

substitute all of them into equation (*)

cov(x) = $\sum_{k=1}^{K} \pi_k[\Sigma_k + \mu_k\mu_k^T] - E[x]E[x]^T = \sum_{k=1}^{K} \pi_k[\Sigma_k + \mu_k\mu_k^T] - \sum_{k=1}^{K} \pi_k\mu_k \sum_{k=1}^{K} \pi_k\mu_k^T$

2.

a. likelihood function: J = $\prod_i^{region} h_i^{n_i}$

log likelihood function: L = logJ = $log \prod_i^{region} h_i^{n_i} = \sum_i^{region} n_i log h_i$

b. considered constraint $\sum_i^{region} h_i\Delta_i = 1$, use Lagrange multiplier and transform to unconstraint problem:

L' = $\sum_i^{region} n_i log h_i + \lambda(\sum_i^{region} h_i\Delta_i - 1)$

$\frac{\delta L'}{\delta h_i} = \frac{n_i}{h_i} + \lambda\,\Delta_i = 0 \rightarrow h_i = -\frac{n_i}{\lambda\Delta_i}$

since $\sum_i^{region} h_i\Delta_i = 1$, $\lambda$ = -N

Thus MLE for $h_i = \frac{n_i}{N\Delta_i}$

c.

1) False

Non-parametric density estimation are the models which can't be decribed by a fixed number of parameters, in fact, there are many, many parameters in a mean square error sense.

2) False

Whether a kernel function is optimal depends on the property of data set.

3) False

if we devide each variable in a D-dimensional space into M bins, the total number of bins will be $M^D$. It's curse of dimensionality because of exponential scaling with D. Thus it's not appropriate wit high dimension. Also, if $n^D >> M$, n samples, most bins are empty.

4) True

Under the assumption of some distribution, parametric method is to estimate the fixed number of parameter under certain probability distribution.

3

a. from definition

H(X,Y) =

$$-\sum_{x\in X}\sum_{y\in Y} p(x,y)logp(x,y) = -\sum_{x\in X}\sum_{y\in Y} p(x)p(y|x)logp(x)p(y|x) = -\sum_{x\in X}\sum_{y\in Y} p(x)p(y|x)(logp(x)+logp(y|x))$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x)p(y|x)logp(x) - \sum_{x\in X}\sum_{y\in Y} p(x)p(y|x)\ logp(y|x) = -\sum_{x\in X} p(x)logp(x) - \sum_{x\in X}\sum_{y\in Y} p(x)p(y|x)logp(y|x)$$

H(X) = $-\sum_{x\in X} p(x)logp(x)$ H(Y|X) = $-\sum_{x\in X}\sum_{y\in Y} p(x)p(y|x)logp(y|x) \rightarrow$ H(X,Y) = H(X) + H(Y|X)

in order to prove H(X,Y) $\leq$ H(X) + H(Y), we need to prove H(Y|X) $\leq$ H(Y)

From the definition of mutual information and make use of Bayes rule

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)log\frac{p(x,y)}{p(x)p(y)} = \sum_{x\in X}\sum_{y\in Y} p(x,y)log\frac{p(y|x)}{p(y)} = \sum_{x\in X}\sum_{y\in Y} p(x,y)logp(y|x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)logp(y)$$

$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)logp(y|x) - \sum_{x\in X}\sum_{y\in Y} p(x|y)p(y)logp(y)$$

$\sum_{x\in X} p(x|y) = 1 \rightarrow$ I(X,Y) = $\sum_{x\in X}\sum_{y\in Y} p(x,y)logp(y|x) - \sum_{y\in Y} p(y)logp(y) = H(Y) - H(Y|X)$

Note the KL divergence of p(x,y) and p(x)p(y) :

KL(p(x,y) ||p(x)p(y)) = $-\sum_{x\in X}\sum_{y\in Y} p(x,y)log\frac{p(x)p(y)}{p(x,y)} = I(X,Y)$

since KL has non-negative property, we can have I(X,Y) $\geq 0$

we have $H(Y) \geq H(Y|X)$

So H(X) + H(Y) $\geq H(Y|X)$ +H(X)

we have shown that H(X,Y) = H(X) + H(Y|X)

Consequently, H(X,Y) $\leq$ H(X) + H(Y)

b.

From (a) : H(X,Y) = H(X) + H(Y|X)

= H(X) $-\sum_{x\in X}\sum_{y\in Y} p(x)p(y|x)logp(y|x)$ (*)

since X and Y are independent, p(x,y) = p(x)p(y) = p(x)p(y|x) $\rightarrow$ p(y) = p(y|x)

replace $p(y|x)$ with p(y) into (*)

H(X,Y) = H(X) $-\sum_{x\in X}\sum_{y\in Y} p(x)p(y)logp(y) = H(X) - \sum_{x\in X} p(x) \sum_{y\in Y} p(y)logp(y) = H(X)+H(Y)$

c.

From a we have already proved that

$I(X,Y) = H(Y) - H(Y|X)$ and H(X,Y) = H(X) + H(Y|X)

Thus I(X,Y) = H(Y) - (H(X,Y)-H(X)) = H(Y)+H(X)-H(X,Y)

d.

Z = X +Y

H(Z|X) =

$$-\sum_{x\in X}\sum_{z\in X+Y} p(x)p(z|x)logp(z|x) = -\sum_{x\in X}\sum_{y\in Y} p(x)p(x+y|x)logp(x+y|x) = -\sum_{x\in X}\sum_{y\in Y} p(x)p(y|x)logp(y|x) = H(Y|X)$$

similarly, we can have H(Z|Y) = H(X|Y)

from (a) we have proved that I(Y,X) = $H(Y) - H(Y|X)$ without independent constraint
Thus, I(Z,X) = H(Z) - H(Z|X), I(Z,Y) = H(Z) - H(Z|Y)
since mutual information is non-negative ( from a)
we can have H(Z) $\geq$ H(Z|X) and H(Z) $\geq$ H(Z|Y)
since H(Z|Y) = H(X|Y) and H(Z|X) = $H(Y|X)$
Thus H(Z) $\geq H(Y|X)$ and H(Z) $\geq$ H(X|Y) …..(*)

when X and Y are independent, from (b) we proved that H(X,Y) = H(X) + H(Y)
combine
(c) I(X,Y) = H(X)+H(Y) - H(X,Y)
(b) H(X,Y) = H(X) + H(Y) and
(a) I(Y,X) = $H(Y) - H(Y|X)$
$\rightarrow H(Y) = H(Y|X)$ and similarily we can get H(X) = H(X|Y)
plug into (*)
we get H(Z) $\geq H(Y)$ and H(Z) $\geq H(X)$
In another words, the addition of independent random variables add uncertainty.

e.
From b we know that when X and Y are independent, H(X,Y) = H(X) + H(Y)
here we want H(Z) = H(X) + H(Y)
So it needs to know the assumption to hold H(X,Y) = H(Z)

H(Z) = $-\sum_{z\in Z} p(z)logp(z)$

H(X,Y) = $-\sum_{x\in X}\sum_{y\in Y} p(x,y)logp(x,y)$

in order to make these two entropy equal to each other, we need a one-to-one mapping from (x,y)

to z. Then H(Z) = $-\sum_{z\in Z} p(z)logp(z)$ = $-\sum_{x\in X}\sum_{y\in Y} p(x,y)logp(x,y)$ =H(X,Y)

4
4.1
use f(x) to classifier result, and y as true label
denote loss function as L(y=-1,f(x) = 1) = q and L(y=1,f(x) = -1) = p

Since bayes classifier means to take the label which has the less cost, it needs to compare expectation of loss function for different judgment.

if Bayes classifier classify x as 1,
loss(1) = p(y=1|x)L(y=1,f(x)=1) + p(y=-1|x)L(y=-1,f(x)=1) = 0+p(y=-1|x)L(y=-1,f(x)=1) = p(y=-1|x)*q
if Bayes classifier classify x as -1
loss(-1) = p(y=1|x)L(y=1,f(x)=-1) + p(y=-1|x)L(y=-1,f(x) = -1) = p*p(y=1|x)

classify x as 1 when loss(1) < loss(-1)

p(y=-1|x)*q < p*p(y=1|x)
(1-p(y=1|x))*q < p * p(y=1|x)
p(y=1|x) > $\frac{q}{p+q}$
so the Bayes classifier will be f(x) = $\left\{1 \; if \; p(y=1|x) \; > \frac{q}{p+q};\; 0 \; otherwise\right\}$

4.2
from 4.1 we know x will be classified as 1 when

p(y=-1|x)*q < p*p(y=1|x)
→ p(y=-1|x)p(x)*q < p*p(y=1|x)p(x)
$\frac{p(y=-1)p(x|y=-1)*q}{p(y=1)p(x|y=1)*p} < 1$
we use $\mu_0, \Sigma_0$ for class -1 and $\mu_1, \Sigma_1$ for class 1, d as dimension and plug in Gaussian distribution :

$\frac{p(y=-1)\frac{1}{(2\pi)^{d/2}|\Sigma_0|^{1/2}}exp[-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)]*q}{p(y=1)\frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}}exp[-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)]*p}$ <1

→ $\frac{p(y=-1)\frac{1}{|\Sigma_0|^{1/2}}exp[-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)]*q}{p(y=1)\frac{1}{|\Sigma_1|^{1/2}}exp[-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)]*p}$ <1

take log on both sides:

$log[\frac{p(y=-1)\frac{1}{|\Sigma_0|^{1/2}}exp[-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)]*q}{p(y=1)\frac{1}{|\Sigma_1|^{1/2}}exp[-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)]*p}] \; < \; 0$

Let h(x) = $log[\frac{p(y=-1)\frac{1}{|\Sigma_0|^{1/2}}exp[-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)]*q}{p(y=1)\frac{1}{|\Sigma_1|^{1/2}}exp[-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)]*p}]$

=

$$logp(y=-1)-\tfrac{1}{2}|\Sigma_0| \; -\tfrac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)+logq-logp(y=1)+\tfrac{1}{2}|\Sigma_1| \; +\tfrac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)-logp$$

=

$$[-\tfrac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) \; +\tfrac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)]+[logp(y=-1)-\tfrac{1}{2}|\Sigma_0|+logq-logp(y=1)+\tfrac{1}{2}|\Sigma_1|-logp]$$

because the term in the second bracket is constant ( determined by data set)
h(x) = $[-(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) \; +(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)]$ +constant < 0
Thus the classifier will be
f(x) = {1 if sign(h(x)) <0; -1 otherwise}
since h(x) is a quadratic function, the decision boundary is not linear, it will be hyperquadratic.
b.
if $\Sigma_0 = \Sigma_1 = \Sigma$, x will be classified as 1 class if
$[-(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) \; +(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]$ +constant < 0
$-x^T\Sigma^{-1}x \; +x^T\Sigma^{-1}\mu_0+\mu_0{}^T\Sigma^{-1}x-\mu_0{}^T\Sigma^{-1}\mu_0+x^T\Sigma^{-1}x \; -x^T\Sigma^{-1}\mu_1 \; -\mu_1{}^T\Sigma^{-1}x+\mu_1{}^T\Sigma^{-1}\mu_1$ +constant<0
$x^T\Sigma^{-1}\mu_0+\mu_0{}^T\Sigma^{-1}x-x^T\Sigma^{-1}\mu_1 \; -\mu_1{}^T\Sigma^{-1}x \; + \; constant' < 0$ (1)
Consequently, the decision boundary is a hyperplane since it's linear w.r.t. x.

c.
if $\Sigma_0 = \Sigma_1 = I,$
we can express (1) as
$x^T\mu_0+\mu_0{}^Tx-x^T\mu_1 \; -\mu_1{}^Tx \; + \; constant' < 0$
$\mu_1{}^Tx \; = \; x^T\mu_1, \mu_0{}^Tx \; = x^T\mu_0$

thus, $x^T(\mu_0 - \mu_1) + constant < 0$
the decision boundary is still linear w.r.t x, and this hyperplane normal to $\mu_0 - \mu_1$ .

5.
basic task
Here, the implementation of EM initialize mixture component equally and initialize $\mu_{jc}$ with uniform distribution randomly plus normalization in order to guarantee $\sum_{j}^{n_w} \mu_{jc} = 1$
To avoid the impact of random initialization, it runs 10 times to get average accuracy for differenct maximum iteration configuration. As shown in figure 1, accuracy become stable after 300 and it's reasonable to accept 80% correct classification.

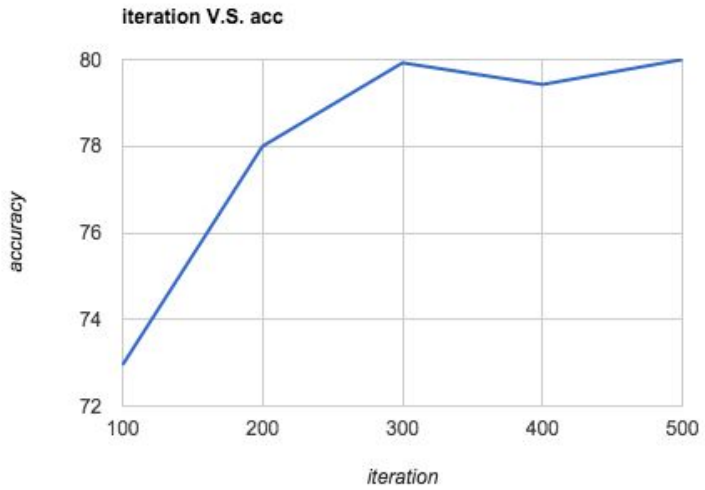


figure 1. accuracy vs. iteration
Besides, I have tried another initialization for $\mu_{jc}$ , to make it the same as mixture component, equal probability to start algorithm. However, its accuracy is pretty bad as shown in figure 2. EM is unable to make progress under this configuration. Also, just with a very sparse matrix for count, it's difficult to update posterior probability and mixture component.

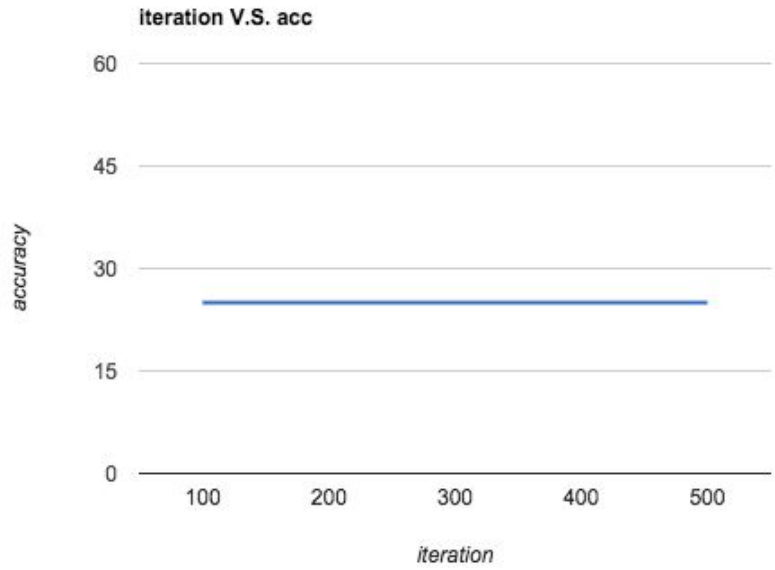


figure 2. accuracy vs. iteration

Compared to normal distribution, randomly initialization with uniform distribution makes the normalization step afterwards easier.

Additionally, classification has been compared with original label in terms of the number of documents in each class as shown in the histogram figure 3.

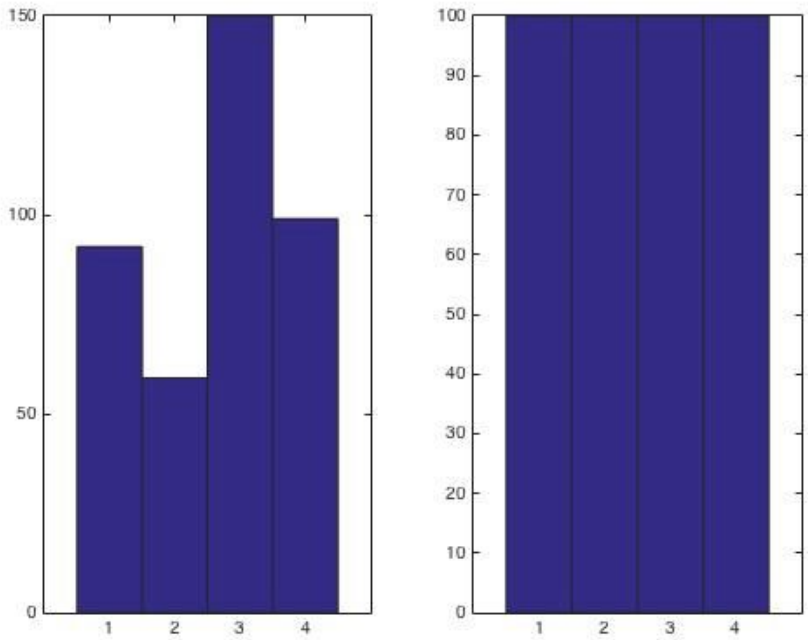


figure 3 classification ( left) v.s true label (right)
(count of objects in each cluster)

Finally, 5 top words for each topic has been output under 500 iteration.

| | word index | | | | |
|---|---|---|---|---|---|
| k=1 | 27 | 85 | 47 | 63 | 68 |
| k=2 | 49 | 87 | 92 | 93 | 41 |
| k=3 | 27 | 85 | 47 | 39 | 63 |
| k=4 | 30 | 57 | 64 | 86 | 41 |

table 1. 5 top words for each topic