CHAPTER 2

Unsupervised Learning

(Probabilistic clustering via EM algorithm)

Content

- Introduction to Unsupervised Learning
- K-means clustering
- **Probabilistic clustering via EM algorithm**
- Hierarchical clustering
- Unsupervised Learning with Python
- Unsupervised Learning with SAS EM
- Determine Number of Clusters with Python
- Density-based Spatial Clustering of Applications with Noise (DBSCAN)

PROBABILISTIC CLUSTERING

Probabilistic Clustering

- Formulate the clustering problem via probability distribution.
- Introduce the concept called maximum likelihood estimator (MLE).
- Introduce the concept mixture model.

Probabilistic Clustering

- To introduce the probabilistic approach, we have to first review K-means clustering.
- The probabilistic approach is closely related to K-means clustering.

Review: K-means clustering

Minimize the following objective function:

$$\min_{I_{ik},\boldsymbol{c}_k} J(I_{ik},\boldsymbol{c}_k), where\ J(I_{ik},\boldsymbol{c}_k) = \sum_{i=1}^{n}\sum_{k=1}^{c} I_{ik}\left|\left|\boldsymbol{x}_i - \boldsymbol{c}_k\right|\right|^2$$

Step 1: Updating Assignment

- Assign each sample to the closest centroid
- That is,

$$I_{ik} = 1 \text{ if } \left|\left|\boldsymbol{x}_i - \boldsymbol{c}_k\right|\right|^2 \leq \left|\left|\boldsymbol{x}_i - \boldsymbol{c}_j\right|\right|^2, \text{ for j=1,…c}$$
$$I_{ik} = 0 \quad \text{otherwise} \qquad .$$

Step 2: Updating Centroid

- Compute the centroids by the following formula

$$\boldsymbol{c}_k = \frac{\sum_{i=1}^{n} I_{ik}\boldsymbol{x}_i}{\sum_{i=1}^{n} I_{ik}}$$

Probabilistic Clustering

- In K-means clustering, the binary indicator $I_{ik} = \{0,1\}$ is used. It also indicates which sample point belongs to which cluster.
- The probabilistic approach changes the binary indicator $I_{ik} = \{0,1\}$ to be a probability z_{ik} with $\sum_{k=1} z_{ik} = 1$.

- **Questions**: Why we have to do this? What is the advantage of employing probabilistic approach?
- **Answer**: It can introduce a way to guess the number of clusters, which is not solvable by K-means clustering.

Probabilistic Clustering – Example 1

- **Question**: Given the following data

1	2	3	11	12	13

- Partition the data into two groups using the probabilistic clustering method with initial guesses $c_1 = 1$ and $c_2 = 11$.

- Remark: We can clearly see that there are two groups in the data. The first group is $\{1,2,3\}$ while the second group is $\{11,12,13\}$. Surely, we can use K-means clustering to solve the problem. The purpose of this example is to illustrate the use of a probabilistic clustering method.

Probabilistic Clustering – Example 1

- Answer: First, we solve the problem using K-means clustering algorithm. Then, we will introduce the probabilistic clustering approach.
- Initial guess: $c_1 = 1$ and $c_2 = 11$
- Step 1: Update the assignment

Original Data	1	2	3	11	12	13
Distance to cluster c_1	0	1	2^2	10^2	11^2	12^2
Distance to cluster c_2	10^2	9^2	8^2	0	1	2^2
Assignment	c_1	c_1	c_1	c_2	c_2	c_2

- Step 2: Compute the Centroid:

$$c_1 = \frac{1+2+3}{3} = 2 \ \ and\ c_2 = \frac{11+12+13}{3} = 12$$

Probabilistic Clustering – Example 1

- Since the cluster centers are not the same, we have to apply the whole procedure again
- Previous cluster centers: $c_1 = 2$ and $c_2 = 12$
- Step 1: Update the assignment

Original Data	1	2	3	11	12	13
Distance to cluster c_1	1	0	1^2	9^2	10^2	10^2
Distance to cluster c_2	11^2	10^2	9^2	1	0	1^2
Assignment	c_1	c_1	c_1	c_2	c_2	c_2

- Step 2: Compute the Centroid:

$$c_1 = \frac{1+2+3}{3} \quad and\ c_2 = \frac{11+12+13}{3} = 12$$

**No change in c_1 and c_2. We can output the results.

Probabilistic Clustering – Example 1

- Now, we introduce the probabilistic approach.
- We assume that the two groups follow two normal distributions with means c_1 and c_2. Both have unit standard derivation.
- The followings are the key steps:
  - Step 1: Compute the probabilities (Assignment step)
  - Step 2: Compute the centroid.
- Similar to K-means clustering, this approach still has the assignment step and the centroid step.

Probabilistic Clustering – Example 1

- Initial guess: $c_1 = 1$ and $c_2 = 11$
- Step 1: Compute the probabilities
- Since we assume that both clusters follow normal distributions with means c_1 and c_2, we can compute the probabilistic assignments z_{i1} and z_{i2} as below

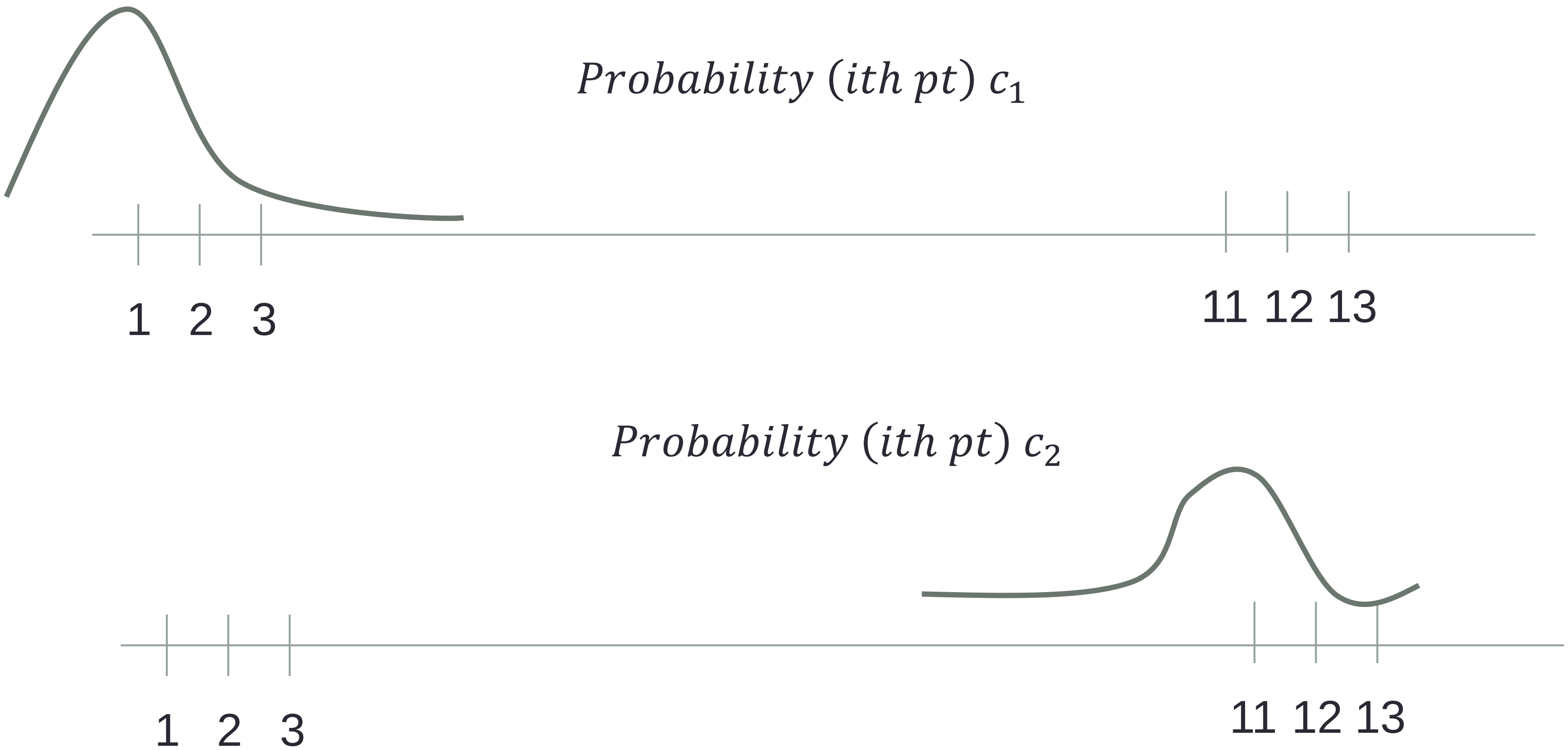
$$z_{i1} = \frac{Probability\ (ith\ pt)\ c_1}{Probability\ (ith\ pt)\ c_1 + Probability\ (ith\ pt)\ c_2}$$

$$z_{i2} = \frac{Probability\ (ith\ pt)\ c_2}{Probability\ (ith\ pt)\ c_1 + Probability\ (ith\ pt)\ c_2}$$

- If $z_{i1} > z_{i2}$, it means the ith point is more likely to belong to c_1. Otherwise, it is c_2.

Probabilistic Clustering – Example 1

- Graphically illustration



Probabilistic Clustering – Example 1

- To find these two probabilities, we need to compute the following:

$$Probability\ (ith\ pt)\ c_1 = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_1)^2}{2}\right)$$

$$Probability\ (ith\ pt)\ c_2 = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_2)^2}{2}\right)$$

Probabilistic Clustering – Example 1

- So, the two probabilistic labels are

$$z_{i1} = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_1)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_1)^2}{2}\right) + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_2)^2}{2}\right)}$$

$$z_{i2} = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_2)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_1)^2}{2}\right) + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_2)^2}{2}\right)}$$

Probabilistic Clustering – Example 1

- Step 1: Compute the two probabilities

Original Data	1	2	3	11	12	13
z_{i1}	1	1	1	0	0	0
z_{i2}	0	0	0	1	1	1
Assignment	c_1	c_1	c_1	c_2	c_2	c_2

- Step 2: Compute the Centroids:

$$c_1 = \frac{\sum_{i=1}^{n} z_{i1} x_i}{\sum_{i=1}^{n} z_{i1}} = 2 \ \ and \ \ c_2 = \frac{\sum_{i=1}^{n} z_{i2} x_i}{\sum_{i=1}^{n} z_{i2}} = 12$$

Since the cluster centers are not the same, we have to apply the whole procedure again

Probabilistic Clustering – Example 1

- $c_1 = 2 \;\; and \;\; c_2 = 12$
- Step 1: Compute the two probabilities
- By applying the two formulae:

$$z_{i1} = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_1)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_1)^2}{2}\right) + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_2)^2}{2}\right)}$$

$$z_{i2} = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_2)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_1)^2}{2}\right) + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_2)^2}{2}\right)}$$

Probabilistic Clustering – Example 1

- Step 1: Compute the two probabilities

Original Data	1	2	3	11	12	13
z_{i1}	1	1	1	0	0	0
z_{i2}	0	0	0	1	1	1
Assignment	c_1	c_1	c_1	c_2	c_2	c_2

- Step 2: Compute the Centroids:

$$c_1 = \frac{\sum_{i=1}^{n} z_{i1} x_i}{\sum_{i=1}^{n} z_{i1}} = 2 \;\; and \;\; c_2 = \frac{\sum_{i=1}^{n} z_{i2} x_i}{\sum_{i=1}^{n} z_{i2}} = 12$$

Since the cluster centers are the same as the previous ones, we can stop and output the solutions.

Principle of Probabilistic Clustering

- Step 1: Compute the two probabilities

$$z_{i1} = \frac{Probability\ (ith\ pt)\ c_1}{Probability\ (ith\ pt)\ c_1 + Probability\ (ith\ pt)\ c_2}$$

$$z_{i2} = \frac{Probability\ (ith\ pt)\ c_2}{Probability\ (ith\ pt)\ c_1 + Probability\ (ith\ pt)\ c_2}$$

- Step 2: Compute the Centroids:

$$c_1 = \frac{\sum_{i=1}^{n} z_{i1} x_i}{\sum_{i=1}^{n} z_{i1}} \quad and \quad c_2 = \frac{\sum_{i=1}^{n} z_{i2} x_i}{\sum_{i=1}^{n} z_{i2}}$$

- Remark: It is noted that we may assume different distributions for the data samples. We will discuss this later.

Probabilistic Clustering – Example 2

- **Question**: Given the following data

1	2	3	3	4	5

- Partition the data into two groups using the probabilistic clustering method with initial guesses $c_1 = 1$ and $c_2 = 5$.
- Assume the two groups follow normal distributions with means c_1 and c_2. Both have unit standard deviation.

Probabilistic Clustering – Example 2

- Answer:
- Step 1: Compute the two probabilities
- By applying the two formulae:

$$z_{i1} = \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i - c_1)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i - c_1)^2}{2}\right) + \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i - c_2)^2}{2}\right)}$$

$$z_{i2} = \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i - c_2)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i - c_1)^2}{2}\right) + \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i - c_2)^2}{2}\right)}$$

Probabilistic Clustering – Example 2

- initial guesses $c_1 = 1$ and $c_2 = 5$
- Step 1: Compute the two probabilities

Original Data	1	2	3	3	4	5
z_{i1}	0.9997	0.9820	0.5	0.5	0.018	0.0003
z_{i2}	0.0003	0.0180	0.5	0.5	0.9820	0.9997
Assignment	c_1	c_1	c_1	c_2	c_2	c_2

- Step 2: Compute the Centroids:

$$c_1 = \frac{\sum_{i=1}^{n} z_{i1} x_i}{\sum_{i=1}^{n} z_{i1}} = 2.0124 \;\; and \;\; c_2 = \frac{\sum_{i=1}^{n} z_{i2} x_i}{\sum_{i=1}^{n} z_{i2}} = 3.9876$$

Since the cluster centers are not the same as the previous ones, we have to apply the whole procedure again.

Probabilistic Clustering – Example 2

- initial guesses $c_1 = 2.0124$and $c_2 = 3.9876$
- Step 1: Compute the two probabilities

Original Data	1	2	3	3	4	5
z_{i1}	0.9811	0.8782	0.5	0.5	0.1218	0.0189
z_{i2}	0.0189	0.1218	0.5	0.5	0.8782	0.9811
Assignment	c_1	c_1	c_1	c_2	c_2	c_2

- Step 2: Compute the Centroids:

$$c_1 = \frac{\sum_{i=1}^{n} z_{i1} x_i}{\sum_{i=1}^{n} z_{i1}} = 2.10641 \ \ and \ \ c_2 = \frac{\sum_{i=1}^{n} z_{i2} x_i}{\sum_{i=1}^{n} z_{i2}} = 3.8936$$

Since the cluster centers are not the same as the previous ones, we have to apply the two steps again. However, we have to repeat many times.

Probabilistic Clustering – Example 2

- In this example, we can see the difference between K-means clustering and probabilistic clustering approaches.
- For K-means clustering, we must assign a data sample to a cluster center.
- However, a data sample may belong to more than one cluster centers.
- In this example, we can see that the data samples $\{3,3\}$ belong to both cluster centers.

Remark on Probabilistic Clustering

- Assume each cluster follows a normal distribution. Thus, each has two distribution parameters, namely mean c and variance σ^2.
- So, the probabilistic clustering has the following settings:

Setting A:

Clusters have different means and difference variances. That is, Cluster 1 c_1, σ_1^2; Cluster 2 c_2, σ_2^2;

Setting B:

Clusters have different means but same variances. That is, Cluster 1 c_1, σ^2; Cluster 2 c_2, σ^2;

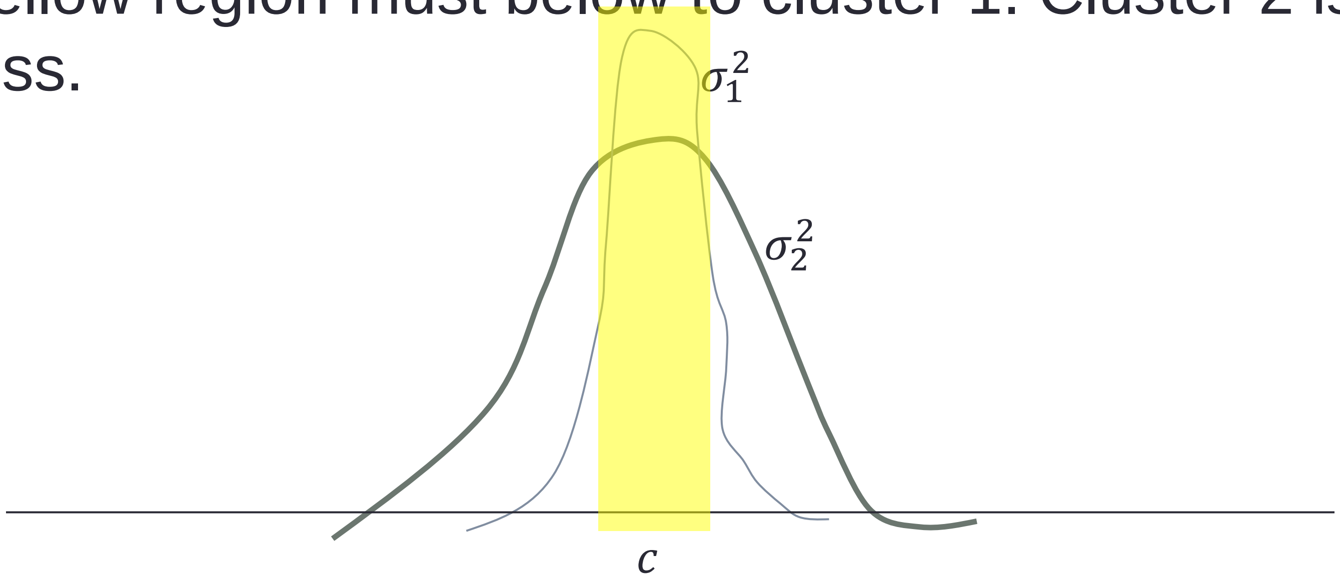
Remark on Probabilistic Clustering

- However, the following setting is not useful.

Setting C:

Clusters have same mean but different variances. That is, Cluster 1 c, σ_1^2; Cluster 2 c, σ_2^2;

Because the two clusters are overlapped, the samples in the yellow region must below to cluster 1. Cluster 2 is then useless.



Covariance Matrix

- Covariance setting in sklearn

covariance_type : *{'full', 'tied', 'diag', 'spherical'}, default='full'*

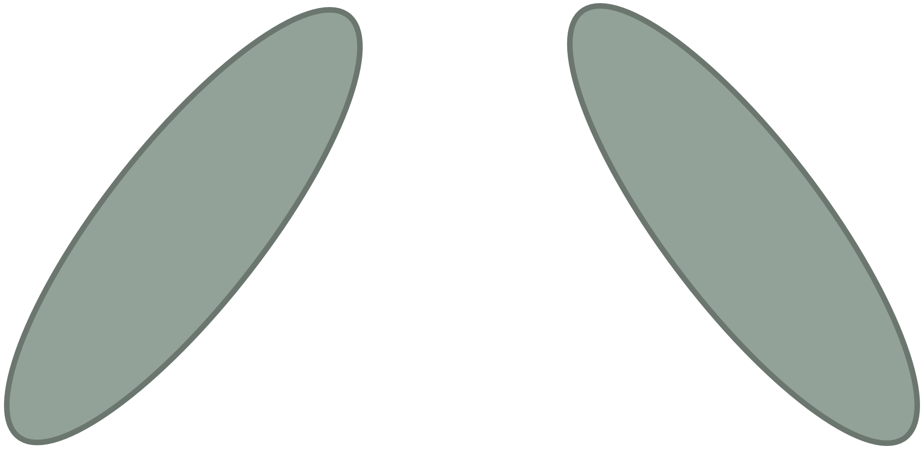
String describing the type of covariance parameters to use. Must be one of:

- 'full': each component has its own general covariance matrix.
- 'tied': all components share the same general covariance matrix.
- 'diag': each component has its own diagonal covariance matrix.
- 'spherical': each component has its own single variance.

Covariance Matrix

- **Full**

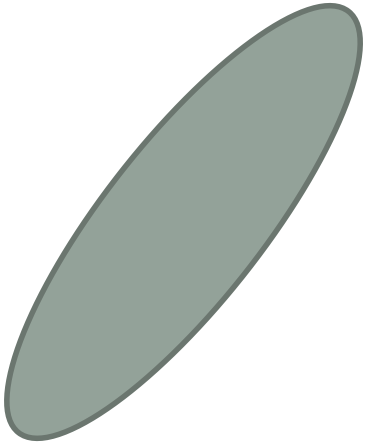
Cluster 1 Cluster 2



Shapes of the two clusters can be very different.

- **Tier**

Cluster 1 and Cluster 2



Both clusters have the same shape.

Covariance Matrix

- **diag**

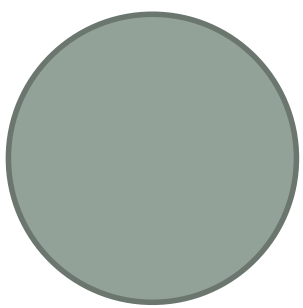
The two clusters can have different variances along the major directions.

Covariance matrix: $\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$

Two different clusters can have different σ_{11} and σ_{22}.
σ_{11} is the variance in the x-direction.
σ_{22} is the variance in the y-direction.

- **Spherical**

Cluster 1 and Cluster 2



Both clusters have the spherical shape.

Origin of Probabilistic Clustering

- In this probabilistic approach, we actually employ a method to solve the problem.
- It is called **Expectation and Maximization** (EM) algorithm.
- It consists of two steps: Expectation step (E-step) and Maximization step (M-step).

Principle of Probabilistic Clustering

- E-Step (Step 1): Compute the two probabilities

$$z_{i1} = \frac{Probability\ (ith\ pt)\ c_1}{Probability\ (ith\ pt)\ c_1 + Probability\ (ith\ pt)\ c_2}$$

$$z_{i2} = \frac{Probability\ (ith\ pt)\ c_2}{Probability\ (ith\ pt)\ c_1 + Probability\ (ith\ pt)\ c_2}$$

- M-Step (Step 2): Compute the Centroids:

$$c_1 = \frac{\sum_{i=1}^{n} z_{i1} x_i}{\sum_{i=1}^{n} z_{i1}} \quad and \quad c_2 = \frac{\sum_{i=1}^{n} z_{i2} x_i}{\sum_{i=1}^{n} z_{i2}}$$

E-step

- In EM algorithm, the E-step is to partition the data into groups.
- The general formula is

$$z_{ik} = \frac{Probability\ (ith\ pt)\ c_k}{\sum_k Probability\ (ith\ pt)\ c_k}$$

- Obviously,

$$\sum_{k=1} z_{ik} = 1 \quad \text{with } 0 \le \mathrm{z_{ik}} \le 1$$

- For each point, the largest value of z_{ik} indicates which group the point belongs to.
- Remark: the probability density function is defined by user. If it is assumed to be a normal, we have to use normal density. If assumed to be an exponential distribution, we have to use exponential density.

M-step (objective function)

- In the above clustering problems, the M-step is to solve the following log probability function.

$$\max_{c_k} \sum_{i,k} z_{ik} \log(p_{ik})$$

where $p_{ik} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - c_k)^2}{2}\right)$

- After some simplifications, the above model becomes

$$\max_{c_k} \sum_{i,k} z_{ik} \left(-\frac{1}{2}\log(2\pi) - \frac{(x_i - c_k)^2}{2}\right)$$

M-step

- Since $\sum_k z_{ik} = 1$ and$-\frac{1}{2}\log(2\pi)$ is a constant, the optimization problem can be simplified to

$$\min_{\mathrm{c_k}} \sum_{i,k} z_{ik}(x_i - c_k)^2$$

- This is very similar to the K-means clustering problem except that the assignment term z_{ik}.
- Of course, the above situation only consider the normality case. Different probability function gives different objective function.

M-step and Likelihood Function

- The role of M-step is to guess the unknown parameters, say the centers c_1 and c_2 by a maximization procedure called maximum likelihood estimation (MLE).
- We first introduce the likelihood function.
- Then, we explain the procedure that can maximize the likelihood function.

What is likelihood? Example 1

Example: Suppose we have the following data

0,1,1,0,0,1,1,0

- In this case it is reasonable to guess the data follow the Bernoulli distribution.
- The remaining question is how can we find the parameter p? i.e.,

$$P(X = x) = p^x(1-p)^{1-x}$$

- This can be achieved by

$$\mathrm{argmax}_p P(Data|B(p))$$

What is likelihood? Example 2

Example: Suppose the following are marks in a course

55.5, 67, 87, 48, 63

- We may have to do some guess works. Marks may follow a Normal distribution.
- The remaining question is how can we find the parameter μ,σ i.e.

$$N(\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2\sigma}(x-\mu)^2}$$

- This can be achieved by

$$argmax_{\mu,\sigma}p(Data|\mu,\sigma)$$

Maximum Likelihood Estimation

- A likelihood function is usually denoted as $L(p)$. p is the parameter to be estimated.
- We want to select a parameter p which will **maximize** the probability that the data was generated from the model with the parameter p plugged-in.
- The parameter **p** is called the maximum likelihood estimator.
- The maximum of the function can be obtained by setting the derivative of the function ==0 and solving for p.

Two Important Facts

- **Fact 1**: If $A_1,\ldots,A_n$ are independent then

$$P(A_1,\ldots,A_n) = \prod_{i=1}^{n} P(A_i)$$

- **Fact 2**: The log function is monotonically increasing. x · y ! Log(x) · Log(y)

- Therefore if a function f(x) >= 0, achieves a maximum at x1, then log(f(x)) also achieves a maximum at x1.

Example of MLE (Recall: Example 1)

Example: Suppose we have the following data

0,1,1,0,0,1,1,0

- In this case it is reasonable to guess the data follow the Bernoulli distribution.
- The remaining question is how can we find the parameter p? i.e.,

$$P(X = x) = p^x (1-p)^{1-x}$$

- This can be achieved by

$$\text{argmax}_p P(Data | B(p))$$

Example of MLE (Recall: Example 1)

- **Answer**:

$$\begin{aligned} L(p) &= P(0,1,1,0,0,1,0,1|p) \\ &= P(0|p)P(1|p)\ldots P(1|p) \\ &= (1-p)p\ldots p \\ &= p^4(1-p)^4 \end{aligned}$$

- Now, choose p which maximizes L(p). Instead we will maximize l(p)= LogL(p)

$$\begin{aligned} \ell(p) &= logL(p) = 4log(p) + 4log(1-p) \\ \frac{d\ell(p)}{dp} &= \frac{4}{p} - \frac{4}{1-p} \equiv 0 \\ \rightarrow p &= \frac{1}{2} \end{aligned}$$

Example of MLE (Recall: Example 1)

- If we replace the Bernoulli variables by $x_i = \{0,1\}$, we have that the the proportion p is the sample proportion. That is,

$$\mathrm{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where $x_i = \{0,1\}$.

- In this example, $x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0, x_5 = 0, x_6 = 1, x_7 = 0, x_8 = 1.$

$$p = \frac{0+1+1+0+0+1+0+1}{8} = \frac{4}{8}$$

Example 2

- Suppose the weights of randomly selected American female college students are normally distributed with unknown mean μ and standard deviation σ. A random sample of 10 American female college students yielded the following weights (in pounds):

  115 122 130 127 149 160 152 138 149 180

- Assume the data follow a normal distribution. Identify the likelihood function and the maximum likelihood estimator of μ of all American female college students. Using the given sample, find a maximum likelihood estimate of μ as well.

Example 2

Answer:

- The probability density function is

$$f(x_i;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

- The likelihood function is

$$L(\mu,\sigma) = \sigma^{-n}(2\pi)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

- Now, we take logarithm to the likelihood function and it is given as below

$$l(\mu,\sigma) = \log(L(\mu,\sigma) = -nlog(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Example 2

- By taking the first derivative of $l(\mu, \sigma)$ with respect to μ and setting it to be zero, we have the estimated $\hat{\mu}$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- That is, the maximum likelihood estimator of μ is the sample mean of the data.
- So,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{10}(115 + \cdots + 180) = 142.2$$

Example 3

- Let X_1, X_2,..., X_n be a random sample from a normal distribution with unknown mean μ and variance σ^2.
- Find maximum likelihood estimators of mean μ and variance σ^2.

Example 3

Answer:

- Again, we have the likelihood function as below

$$L(\mu, \sigma) = \sigma^{-n}(2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

- Taking logarithm to the likelihood function, we have

$$l(\mu, \sigma) = \log(L(\mu, \sigma) = -nlog(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

- By taking the first derivative of $l(\mu, \sigma)$ with respect to μ and setting it to be zero, we have the estimated $\hat{\mu}$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Example 3

- By taking the first derivative of $l(\mu, \sigma)$ with respect to σ and setting it to be zero, we have

$$-\frac{n}{\sigma} + \frac{2}{2\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

- the estimated $\hat{\sigma}$ is

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Which is the sample standard derivation.

Reference Notes

- The following slides are for reference only.
- It introduces another probabilistic clustering technique that can cluster textural data.
- The method assumes each cluster follows a Bernoulli distribution. Then, use EM algorithm to update the probabilities and the distribution parameter.

Example 4. Mixture of Coin Tosses

- We are given two set of coin tosses. Each coin was tosses 10 times.
- Their mixture is modelled by the following equation

$$\max_{\theta_{\mathrm{k}}} \sum_{k=1}^{2} \sum_{i=1}^{5} z_{ik} \log(p_{ik})$$

where $p_{ik} = \Pi_{\mathrm{j}=1}^{10} \theta_k^{x_{\mathrm{i}}^{j}} (1-\theta_k)^{\left(1-x_{\mathrm{i}}^{j}\right)}$. Assume z_{ik} are some known functions.

Example 4. Mixture of Coin Tosses

- Index i: ith set, $i = 1,2, \ldots 5$
- Index j: jth toss , $j = 1,2, \ldots 10$
- Index k: kth coin, $k = 1,2$

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

$If\, i = 1\ and\ k = 1,$
$p_{ik} = \Pi_{j=1}^{10} \theta_k^{x_i^j} (1 - \theta_k)^{(1 - x_i^j)}$ represents the 1st set and assumed to be 1st coin.

Example 4. Mixture of Coin Tosses

- Objective function:

$$\max_{\theta_{\mathrm{k}}} \sum_{k=1}^{2} \sum_{i=1}^{5} z_{ik} \log(p_{ik})$$

- This objective function partitions the data into two clusters.
- Each cluster is labelled by z_{ik} with centroid θ_k
- Find θ_k

Example 4. Mixture of Coin Tosses

- Answer: By taking differentiation w.r.t. θ_k and setting to zero, the MLE estimate of θ_k is

$$\theta_k = \frac{\sum_{i=1}^{5} z_{ik} \left(\frac{1}{10} \sum_{j=1}^{10} x_i^j \right)}{\sum_{i=1}^{5} z_{ik}}$$

This estimate will used in the M-step of the EM algorithm. This is shown in next section.

Main Results of MLE

- The likelihood function for the normal variables is

$$\sigma^{-n}(2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

If σ is given, the MLE estimate of μ is

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

If σ is not known, the MLE estimates of μ and σ are

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \;\; and \;\; \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Main Results of MLE

- The likelihood function for the Bernoulli variables is

$$\Pi_{i=1}^{n} p^{x_{\mathrm{i}}}(1-p)^{(1-x_i)}$$

The MLE estimate of p is

$$\mathrm{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- The likelihood function for the mixture of Bernoulli variables is

$$\max_{\theta_{\mathrm{k}}} \sum_{k=1}^{2} \sum_{i=1}^{n} z_{ik} \log(p_{ik})$$

The MLE estimate of θ_k is

$$\theta_k = \frac{\sum_{i=1}^{n} z_{ik} \left(\frac{1}{m} \sum_{j=1}^{m} x_i^j\right)}{\sum_{i=1}^{n} z_{ik}}$$

Summary of MLE

- Steps to perform the estimate

1. Take the product of the probability density functions.
2. Take logarithm to the product.
3. Take the first derivative of the log function and set it to be zero.

Example 1

- Assume that we have two coins, C1 and C2.
- Assume the bias of C1 is θ_1 (i.e. probability of getting heads with C1, it may not be a fair coin)
- Assume the bias of C2 is θ_2 (i.e. probability of getting heads with C2, it may not be a fair coin)

Example 1

- We have two coins. We toss each of them 10 times.

Question: Find θ_1 and θ_2 if we know the identities of the coins. (i.e. We can find these two parameters separately!)

MLE Problem - Example 1

- **Answer**: Again, we discuss the MLE problem first and then EM problem.
- For a coin tossing problem, it follows a Bernoulli distribution. So, the probability distribution function is

$$\theta_k^x (1 - \theta_k)^{1-x}$$

- Here $x = \{0,1\}$ is a Bernoulli variable.
- By "Example of MLE (Recall: Example 1)", we know that the estimates θ_1 and θ_2 can be obtained by the following formula

$$\theta_1 = \frac{number\ of\ heads\ using\ C1}{total\ number\ of\ flips\ using\ C1}$$

and

$$\theta_2 = \frac{number\ of\ heads\ using\ C2}{total\ number\ of\ flips\ using\ C2}$$

MLE Problem - Example 1

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

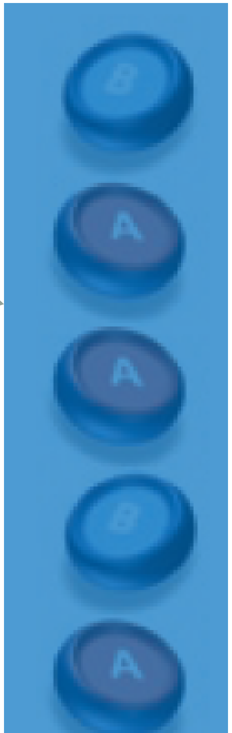
	Coin A	Coin B
		5H, 5T
	9H, 1T	
	8H, 2T	
		4H, 6T
	7H, 3T	
Total:	24H, 6T	9H, 11T

$$\theta_1 = \frac{24}{24+6} = 0.8 \text{ and } \theta_2 = \frac{9}{9+11} = 0.45$$

EM Algorithm - Example 1

- Now, we modify the problem.
- We do not know the identities of the coins used for each set of tosses (we treat them as hidden variables) Assume the two unknown identities as $z_1(i)$ and $z_2(i)$
- We have to cluster the data into two groups.

Assume not to be known.



H T T T H H T H T H   Label: $z_1(1), z_2(1)$

H H H H T H H H H H   Label: $z_1(2), z_2(2)$

H T H H H H H T H H   …

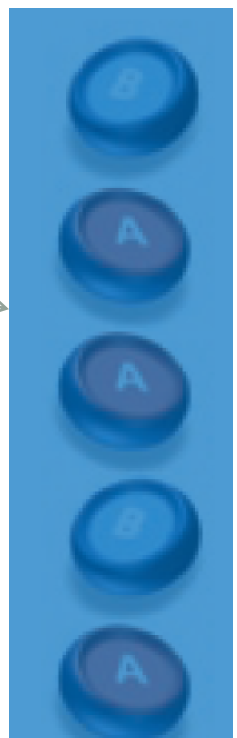
H T H T T T H H T T

T H H H T H H H T H

EM Algorithm - Example 1

- Assume the two probability assignments as $z_1(i)$ and $z_2(i)$ with $z_1(i) + z_2(i) = 1$
- The index i in $z_1(i)$ and $z_2(i)$ refers to the ith sample point (i.e. ith set in this example).
- If $z_1(i) > z_2(i)$, the ith set is more likely to belong set 1. Otherwise, it belongs to set 2.

Assume not to be known.



H T T T H H T H T H   Label: $z_1(1), z_2(1)$

H H H H T H H H H H   Label: $z_1(2), z_2(2)$

H T H H H H H T H H $\cdots$

H T H T T T H H T T

T H H H T H H H T H

EM Algorithm - Example 1

- **Question**: There are two objectives in this single question.
- Use EM Algorithm to find the probability assignment "labels" $z_1(\mathrm{i})$ and $z_2(i)$ of the set of tosses.
- Estimate the Bernoulli parameters (i.e. θ_1 and θ_2) of the two distributions.

EM Algorithm - Example 1

Answer: Following the principle, we have the following two steps:

- E-Step (Step 1): Compute the two probabilities

$$z_{i1} = \frac{Probability\ (ith\ pt)\ c_1}{Probability\ (ith\ pt)\ c_1 + Probability\ (ith\ pt)\ c_2}$$

$$z_{i2} = \frac{Probability\ (ith\ pt)\ c_2}{Probability\ (ith\ pt)\ c_1 + Probability\ (ith\ pt)\ c_2}$$

- M-Step (Step 2): Compute the Centroids:

Example 1 (EM Algorithm) – E step

- Initial guess for the two Bernoulli parameters θ_1 and θ_2 as

$$\theta_1 = 0.6 \text{ and } \theta_2 = 0.5$$

Step 1: (E-Step)

$$z_1(i) = \frac{Probability\ (ith\ set)\ \theta_1}{Probability(ith\ set)\ \theta_1 + Probability(ith\ set)\ \theta_2}$$

$$z_2(i) = \frac{Probability(ith\ set)\theta_2}{Probability(ith\ set)\ \theta_1 + Probability(ith\ set)\ \theta_2}$$

Example 1 (EM Algorithm) – E step

- Probability (ith set) θ_1: $\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j}$;
- Probability (ith set) θ_2: $\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}$;

$$z_1(i) = \frac{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j}}{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j} + \Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}$$

$$z_2(i) = \frac{\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j} + \Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}$$

It is noted that $z_1(i) + z_2(i) = 1$.

Example 1 (EM Algorithm) – E step

- Probability (ith set) for θ_1:

Note: $x_i^j = 1$ if the toss j at set i is a head.
$x_i^j = 0$ if the toss j at set i is a tail.

ith set	Prob	
1	$\Pi_{j=1}^{10} \theta_1^{x_1^j} (1-\theta_1)^{1-x_1^j}$	H T T T H H T H T H
2	$\Pi_{j=1}^{10} \theta_1^{x_2^j} (1-\theta_1)^{1-x_2^j}$	H H H H T H H H H H
3	$\Pi_{j=1}^{10} \theta_1^{x_3^j} (1-\theta_1)^{1-x_3^j}$	H T H H H H H T H H
4	$\Pi_{j=1}^{10} \theta_1^{x_4^j} (1-\theta_1)^{1-x_4^j}$	H T H T T T H H T T
5	$\Pi_{j=1}^{10} \theta_1^{x_5^j} (1-\theta_1)^{1-x_5^j}$	T H H H T H H H T H

Example 1 (EM Algorithm) – E step

- Computing $z_1(i)$, we have

Note: $x_i^j = 1$ if the toss j at set i is a head.
$x_i^j = 0$ if the toss j at set i is a tail.

$H \rightarrow 1$ and $T \rightarrow 0$

Recall: $\theta_1 = 0.6$

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

$$\Pi_{j=1}^{10} \theta_1^{x_i^j} (1-\theta_1)^{1-x_i^j}$$
$$= (0.6^1 \times 0.4^0)(0.6^0 \times 0.4^1) \ldots (0.6^1 \times 0.4^0)$$

$$\Pi_{j=1}^{10} \theta_1^{x_i^j} (1-\theta_1)^{1-x_i^j}$$
$$= (0.6^0 \times 0.4^1)(0.6^1 \times 0.4^0) \ldots (0.6^1 \times 0.4^0)$$

Example 1 (EM Algorithm) – E step

- Probability (ith set) for θ_2:

Note: $x_i^j = 1$ if the toss j at set i is a head.
$x_i^j = 0$ if the toss j at set i is a tail.

ith set	Prob	
1	$\Pi_{j=1}^{10} \theta_2^{x_1^j} (1-\theta_2)^{1-x_1^j}$	H T T T H H T H T H
2	$\Pi_{j=1}^{10} \theta_2^{x_2^j} (1-\theta_2)^{1-x_2^j}$	H H H H T H H H H H
3	$\Pi_{j=1}^{10} \theta_2^{x_3^j} (1-\theta_2)^{1-x_3^j}$	H T H H H H H T H H
4	$\Pi_{j=1}^{10} \theta_2^{x_4^j} (1-\theta_2)^{1-x_4^j}$	H T H T T T H H T T
5	$\Pi_{j=1}^{10} \theta_2^{x_5^j} (1-\theta_2)^{1-x_5^j}$	T H H H T H H H T H

Example 1 (EM Algorithm) – E step

- Computing $z_2(i)$, we have

Note: $x_i^j = 1$ if the toss j at set i is a head.
$x_i^j = 0$ if the toss j at set i is a tail.

$H \rightarrow 1$ and $T \rightarrow 0$

Recall: $\theta_2 = 0.5$

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

$$\Pi_{j=1}^{10} \theta_2^{x_i^j} (1-\theta_2)^{1-x_i^j}$$
$$= (0.5^1 \times 0.5^0)(0.5^0 \times 0.5^1) \ldots (0.5^1 \times 0.5^0)$$

$$\Pi_{j=1}^{10} \theta_2^{x_i^j} (1-\theta_2)^{1-x_i^j}$$
$$= (0.5^0 \times 0.5^1)(0.5^1 \times 0.5^0) \ldots (0.5^1 \times 0.5^0)$$

Example 1 (EM Algorithm) – E step

- The results are

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

	z_1(i)	z_2(i)
1	0.4491	0.5509
2	0.8050	0.1950
3	0.7335	0.2665
4	0.3522	0.6478
5	0.6472	0.3528

Example 1 (EM Algorithm) – E step

- The results are

	z_1(i)	z_2(i)	
H T T T H H T H T H	0.4491	0.5509	More likely to be coin 2
H H H H T H H H H H	0.8050	0.1950	More likely to be coin 1
H T H H H H H T H H	0.7335	0.2665	More likely to be coin 1
H T H T T T H H T T	0.3522	0.6478	More likely to be coin 2
T H H H T H H H T H	0.6472	0.3528	More likely to be coin 1

EM Algorithm - Example 1

- M-Step (Step 2): Compute the Centroids:

- In this example, x_i is a sequence (e.g. HTTHHTT..), which follows a Bernoulli distribution. We have to use the following strategy.

Refer to "Main Results of MLE".
The objective function is

$$\max_{\theta_\mathrm{k}} \sum_{i,k} z_{ik} \log(p_{ik})$$

p_{ik} is the probability of the ith sample and it is

$$p_{ik} = \Pi_{\mathrm{j}=1}^{10} {\theta_k}^{x_\mathrm{i}^j} (1-\theta_k)^{\left(1-x_\mathrm{i}^j\right)}$$

By taking differentiation w.r.t. θ_k and setting to zero, the MLE estimate of θ_k is

$$\theta_k = \frac{\sum_{i=1}^{5} z_{ik} \left(\frac{1}{10} \sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{ik}}$$

Example 1 (EM Algorithm) – M step

- Conclusion of M-step: Updating the Bernoulli parameters θ_1 and θ_2.

$$\theta_1 = \frac{\sum_{i=1}^{5} z_{i1}\left(\frac{1}{10}\sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{i1}} \text{ and } \theta_2 = \frac{\sum_{i=1}^{5} z_{i2}\left(\frac{1}{10}\sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{i2}}$$

- So, we have

$$\theta_1 = 0.7130 \text{ and } \theta_2 = 0.5813$$

Key Steps of EM Algorithm

- Review of the two steps:
- E-step:

$$z_1(i) = \frac{Probability\ (ith\ set)\ \theta_1}{Probability(ith\ set)\ \theta_1 + Probability(ith\ set)\ \theta_2}$$

$$z_2(i) = \frac{Probability(ith\ set)\theta_2}{Probability(ith\ set)\ \theta_1 + Probability(ith\ set)\ \theta_2}$$

- M-step:

$$\theta_1 = \frac{\sum_{i=1}^{5} z_{i1}\left(\frac{1}{10}\sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{i1}} \text{ and } \theta_2 = \frac{\sum_{i=1}^{5} z_{i2}\left(\frac{1}{10}\sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{i2}}$$

Example 1 (EM Algorithm)

- In the mid-term test, if you are asked a question about EM algorithm, it will only be about Bernoulli variables with two hidden set of variables (i.e. z_{i1} and z_{i2}).
- Next, we apply the two steps of EM algorithm and find the solution.

Example 1 (EM Algorithm) – E step

- Given $\theta_1 = 0.7130$ and $\theta_2 = 0.5813$
- E-step:

$$z_1(i) = \frac{Probability\ (ith\ set)\ \theta_1}{Probability(ith\ set)\ \theta_1 + Probability(ith\ set)\ \theta_2}$$

$$z_2(i) = \frac{Probability(ith\ set)\theta_2}{Probability(ith\ set)\ \theta_1 + Probability(ith\ set)\ \theta_2}$$

Example 1 (EM Algorithm) – E step

- Probability (ith set) for $\theta_1 = 0.7130$:

Note: $x_i^j = 1$ if the toss j at set i is a head.

$x_i^j = 0$ if the toss j at set i is a tail.

ith set	Prob	
1	$\Pi_{j=1}^{10} \theta_1^{x_1^j}(1-\theta_1)^{1-x_1^j}$	H T T T H H T H T H
2	$\Pi_{j=1}^{10} \theta_1^{x_2^j}(1-\theta_1)^{1-x_2^j}$	H H H H T H H H H H
3	$\Pi_{j=1}^{10} \theta_1^{x_3^j}(1-\theta_1)^{1-x_3^j}$	H T H H H H H T H H
4	$\Pi_{j=1}^{10} \theta_1^{x_4^j}(1-\theta_1)^{1-x_4^j}$	H T H T T T H H T T
5	$\Pi_{j=1}^{10} \theta_1^{x_5^j}(1-\theta_1)^{1-x_5^j}$	T H H H T H H H T H

Example 1 (EM Algorithm) – E step

- Probability (ith set) for θ_2=0.5813:

Note: $x_i^j = 1$ if the toss j at set i is a head.
$x_i^j = 0$ if the toss j at set i is a tail.

ith set	Prob
1	$\Pi_{j=1}^{10} \theta_2^{x_1^j}(1-\theta_2)^{1-x_1^j}$
2	$\Pi_{j=1}^{10} \theta_2^{x_2^j}(1-\theta_2)^{1-x_2^j}$
3	$\Pi_{j=1}^{10} \theta_2^{x_3^j}(1-\theta_2)^{1-x_3^j}$
4	$\Pi_{j=1}^{10} \theta_2^{x_4^j}(1-\theta_2)^{1-x_4^j}$
5	$\Pi_{j=1}^{10} \theta_2^{x_5^j}(1-\theta_2)^{1-x_5^j}$

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

Example 1 (EM Algorithm) – E step

- Probability (ith set) θ_1: $\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j}$;
- Probability (ith set) θ_2: $\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}$;

$$z_1(i)=\frac{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j}}{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j}+\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}$$

$$z_2(i)=\frac{\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j}+\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}$$

It is noted that $z_1(i)+z_2(i)=1$.

Example 1 (EM Algorithm) – E step

- The results are

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

	z_1(i)	z_2(i)
1	0.2958	0.7042
2	0.8115	0.1885
3	0.7064	0.2936
4	0.1901	0.8099
5	0.5735	0.4265

Example 1 (EM Algorithm) – M step

- Updating the Bernoulli parameters θ_1 and θ_2.
- We have to use the weighted mean of the MLE of the estimate

$$\theta_1 = \frac{\sum_{i=1}^{5} z_{i1}\left(\frac{1}{10}\sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{i1}} \text{ and } \theta_2 = \frac{\sum_{i=1}^{5} z_{i2}\left(\frac{1}{10}\sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{i2}}$$

- So, we have

$$\theta_1 = 0.7453 \text{ and } \theta_2 = 0.5693$$

Example 1 (EM Algorithm)

Compare the two unknown parameters before and after the iteration

- Previous iteration: $\theta_1 = 0.7130$ and $\theta_2 = 0.5813$
- This iteration: $\theta_1 = 0.7453$ and $\theta_2 = 0.5693$

They are not the same. So, we have to apply the two updates again.

Example 1 (EM Algorithm) – E step

- Given $\theta_1 = 0.7453$ and $\theta_2 = 0.5693$
- E-step:

$$z_1(i) = \frac{Probability\ (ith\ set)\ \theta_1}{Probability(ith\ set)\ \theta_1 + Probability(ith\ set)\ \theta_2}$$

$$z_2(i) = \frac{Probability(ith\ set)\theta_2}{Probability(ith\ set)\ \theta_1 + Probability(ith\ set)\ \theta_2}$$

Example 1 (EM Algorithm) – E step

- Probability (ith set) for $\theta_1 = 0.7453$:

Note: $x_i^j = 1$ if the toss j at set i is a head.
$x_i^j = 0$ if the toss j at set i is a tail.

ith set	Prob
1	$\Pi_{j=1}^{10}\theta_1^{x_1^j}(1-\theta_1)^{1-x_1^j}$
2	$\Pi_{j=1}^{10}\theta_1^{x_2^j}(1-\theta_1)^{1-x_2^j}$
3	$\Pi_{j=1}^{10}\theta_1^{x_3^j}(1-\theta_1)^{1-x_3^j}$
4	$\Pi_{j=1}^{10}\theta_1^{x_4^j}(1-\theta_1)^{1-x_4^j}$
5	$\Pi_{j=1}^{10}\theta_1^{x_5^j}(1-\theta_1)^{1-x_5^j}$

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

Example 1 (EM Algorithm) – E step

- Probability (ith set) for θ_2=0.5693:

Note: $x_i^j = 1$ if the toss j at set i is a head.
$x_i^j = 0$ if the toss j at set i is a tail.

ith set	Prob	
1	$\Pi_{j=1}^{10} \theta_2^{x_1^j} (1-\theta_2)^{1-x_1^j}$	H T T T H H T H T H
2	$\Pi_{j=1}^{10} \theta_2^{x_2^j} (1-\theta_2)^{1-x_2^j}$	H H H H T H H H H H
3	$\Pi_{j=1}^{10} \theta_2^{x_3^j} (1-\theta_2)^{1-x_3^j}$	H T H H H H H T H H
4	$\Pi_{j=1}^{10} \theta_2^{x_4^j} (1-\theta_2)^{1-x_4^j}$	H T H T T T H H T T
5	$\Pi_{j=1}^{10} \theta_2^{x_5^j} (1-\theta_2)^{1-x_5^j}$	T H H H T H H H T H

Example 1 (EM Algorithm) – E step

- Probability (ith set) θ_1: $\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j}$;
- Probability (ith set) θ_2: $\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}$;

$$z_1(i) = \frac{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j}}{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j} + \Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}$$

$$z_2(i) = \frac{\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j} + \Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}$$

It is noted that $z_1(i) + z_2(i) = 1$.

Example 1 (EM Algorithm) – E step

- The results are

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

	z_1(i)	z_2(i)
1	0.2176	0.7824
2	0.8698	0.1302
3	0.7512	0.2488
4	0.1116	0.8884
5	0.5769	0.4231

Example 1 (EM Algorithm) – M step

- Updating the Bernoulli parameters θ_1 and θ_2.
- We have to use the weighted mean of the MLE of the estimate

$$\theta_1 = \frac{\sum_{i=1}^{5} z_{i1}\left(\frac{1}{10}\sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{i1}} \text{ and } \theta_2 = \frac{\sum_{i=1}^{5} z_{i2}\left(\frac{1}{10}\sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{i2}}$$

- So, we have

$$\theta_1 = 0.7681 \text{ and } \theta_2 = 0.5495$$

Example 1 (EM Algorithm)

Compare the two unknown parameters before and after the iteration

- Previous iteration: $\theta_1 = 0.7453$ and $\theta_2 = 0.5693$
- This iteration: $\theta_1 = 0.7681$ and $\theta_2 = 0.5495$

They are not the same. So, we have to apply the two updates again.

Example 1 (EM Algorithm) – E step

- Given $\theta_1 = 0.7681$ and $\theta_2 = 0.5495$
- E-step:

$$z_1(i) = \frac{Probability\ (ith\ set)\ \theta_1}{Probability(ith\ set)\ \theta_1 + Probability(ith\ set)\ \theta_2}$$

$$z_2(i) = \frac{Probability(ith\ set)\theta_2}{Probability(ith\ set)\ \theta_1 + Probability(ith\ set)\ \theta_2}$$

Example 1 (EM Algorithm) – E step

- Probability (ith set) for $\theta_1 = 0.7681$:

Note: $x_i^j = 1$ if the toss j at set i is a head.
$x_i^j = 0$ if the toss j at set i is a tail.

ith set	Prob	
1	$\Pi_{j=1}^{10} \theta_1^{x_1^j} (1-\theta_1)^{1-x_1^j}$	← H T T T H H T H T H
2	$\Pi_{j=1}^{10} \theta_1^{x_2^j} (1-\theta_1)^{1-x_2^j}$	← H H H H T H H H H H
3	$\Pi_{j=1}^{10} \theta_1^{x_3^j} (1-\theta_1)^{1-x_3^j}$	← H T H H H H H T H H
4	$\Pi_{j=1}^{10} \theta_1^{x_4^j} (1-\theta_1)^{1-x_4^j}$	← H T H T T T H H T T
5	$\Pi_{j=1}^{10} \theta_1^{x_5^j} (1-\theta_1)^{1-x_5^j}$	← T H H H T H H H T H

Example 1 (EM Algorithm) – E step

- Probability (ith set) for θ_2=0.5495:

Note: $x_i^j = 1$ if the toss j at set i is a head.
$x_i^j = 0$ if the toss j at set i is a tail.

ith set	Prob	
1	$\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_1^j}(1-\theta_2)^{1-x_1^j}$	H T T T H H T H T H
2	$\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_2^j}(1-\theta_2)^{1-x_2^j}$	H H H H T H H H H H
3	$\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_3^j}(1-\theta_2)^{1-x_3^j}$	H T H H H H H T H H
4	$\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_4^j}(1-\theta_2)^{1-x_4^j}$	H T H T T T H H T T
5	$\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_5^j}(1-\theta_2)^{1-x_5^j}$	T H H H T H H H T H

Example 1 (EM Algorithm) – E step

- Probability (ith set) θ_1: $\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j}$;
- Probability (ith set) θ_2: $\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}$;

$$z_1(i) = \frac{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j}}{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j} + \Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}$$

$$z_2(i) = \frac{\Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}{\Pi_{\mathrm{j}=1}^{10}\theta_1^{x_i^j}(1-\theta_1)^{1-x_i^j} + \Pi_{\mathrm{j}=1}^{10}\theta_2^{x_i^j}(1-\theta_2)^{1-x_i^j}}$$

It is noted that $z_1(i) + z_2(i) = 1$.

Example 1 (EM Algorithm) – E step

- The results are

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

	z_1(i)	z_2(i)
1	0.1617	0.8303
2	0.9129	0.0871
3	0.7943	0.2057
4	0.0663	0.9337
5	0.5871	0.4129

Example 1 (EM Algorithm) – M step

- Updating the Bernoulli parameters θ_1 and θ_2.
- We have to use the weighted mean of the MLE of the estimate

$$\theta_1 = \frac{\sum_{i=1}^{5} z_{i1}\left(\frac{1}{10}\sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{i1}} \text{ and } \theta_2 = \frac{\sum_{i=1}^{5} z_{i2}\left(\frac{1}{10}\sum_{j=1}^{10} x_i^j\right)}{\sum_{i=1}^{5} z_{i2}}$$

- So, we have

$$\theta_1 = 0.7832 \text{ and } \theta_2 = 0.5346$$

Example 1 (EM Algorithm)

Compare the two unknown parameters before and after the iteration

- Previous iteration: $\theta_1 = 0.7681$ and $\theta_2 = 0.5495$
- This iteration: $\theta_1 = 0.7832$ and $\theta_2 = 0.5346$

They are not the same.

We can keep applying the procedure.

Example 1 (EM Algorithm)

Eventually, we will get the solution as

$$\theta_1 = 0.7968 \text{ and } \theta_2 = 0.5196$$

The results show that the set seems to be classified correctly.

Sequence		z_1(i)	z_2(i)
H T T T H H T H T H (Coin 2)	1	0.1031	0.8969
H H H H T H H H H H (Coin 1)	2	0.9519	0.0481
H T H H H H H T H H (Coin 1)	3	0.8454	0.1546
H T H T T T H H T T (Coin 2)	4	0.0307	0.9693
T H H H T H H H T H (Coin 1)	5	0.6014	0.3986

Remark for EM Algorithm

- In the mid-term test, if you are asked a question, you only need to write down the following two steps.
- E-step:

	z_1(i)	z_2(i)
1	0.1617	0.8303
2	0.9129	0.0871
3	0.7943	0.2057
4	0.0663	0.9337
5	0.5871	0.4129

- M-step: $\theta_1 = 0.7832$ and $\theta_2 = 0.5346$
- (You may add some equations in your calculations.)

Remark for EM Algorithm

- In this module, you only need to know the following two different centroids:

$$\theta_k = \frac{\sum_{i=1}^{n} z_{ik} x_i}{\sum_{i=1}^{n} z_{ik}}$$

(use when the distribution is normal)

and

$$\theta_k = \frac{\sum_{i=1}^{n} z_{ik} \left(\frac{1}{m} \sum_{j=1}^{m} x_i^j\right)}{\sum_{i=1}^{n} z_{ik}}$$

(use when the distribution is Bernoulli)

n is the number of samples and m is the length of the sequence.

EM ALGORITHM FOR MISSING DATA PROBLEM

EM Algorithm - Example 1

- In this example, we first consider the MLE problem. Then, we will modify it to be EM problem.
- Let events be “grades in a class”

Guess Distribution	Collected Sample
a = Gets an A	$P(A) = \frac{1}{2}$
b = Gets a B	$P(B) = \mu$
c = Gets a C	$P(C) = 2\mu$
d = Gets a D	$P(D) = \frac{1}{2} - 3\mu$

- What is the maximum likelihood estimate of μ? Express the estimate of μ in terms of a,b,c & d

EM Algorithm - Example 1

- It is noted that the guess distribution is similar to the assuming Gaussian distribution for the data in Examples 2 and 3.
- It is also noted that the collected samples are the $X_1, \ldots X_n$ in Examples 2 and 3.
- Remark: $P(A) + P(B) + P(C) + P(D) = 1$.

EM Algorithm - Example 1

- Answer:
- The likelihood function is

$$L(\mu) = \left(\frac{1}{2}\right)^{a} (\mu)^{b}(2\mu)^{c}\left(\frac{1}{2} - 3\mu\right)^{d}$$

The log-likelihood function is

$$l(\mu) = \log\big(L(\mu)\big)$$

$$= alog\left(\frac{1}{2}\right) + blog(\mu) + clog(2\mu) + dlog\left(\frac{1}{2} - 3\mu\right)$$

EM Algorithm - Example 1

- By taking the first derivative of the log function with respect to μ, we have

$$\frac{b}{\mu} + \frac{2c}{2\mu} - \frac{3d}{\frac{1}{2} - 3\mu} = 0$$

- So, the estimate is

$$\mu = \frac{b + c}{6(b + c + d)}$$

Example 1
Same Problem with Hidden Information

- Let events be "grades in a class"

Guess Distribution	Collected Sample
a = Gets an A	$P(A) = \frac{1}{2}$
b = Gets a B	$P(B) = \mu$
c = Gets a C	$P(C) = 2\mu$
d = Gets a D	$P(D) = \frac{1}{2} - 3\mu$

- Somehow, the statistical forms of a and b are not known. They can be treated as "missing data". But we know that

$$a + b = h$$

- We also know the statistical forms of c and d
- What is the maximum likelihood estimate of μ?

Example 1
Same Problem with Hidden Information

- So, we have the missing data problem.
- The forms of a and b are not known. But their relationship is known.
- Intuitively, the problem can be solved as below

Expectation:

The expected value of a and b can be computed as

$$a = \frac{P(A)}{P(A)+P(B)} h = \frac{\frac{1}{2}}{\frac{1}{2}+\mu} h \text{ and } b = \frac{P(B)}{P(A)+P(B)} = \frac{\mu}{\frac{1}{2}+\mu} h$$

The uses of $\frac{P(A)}{P(A)+P(B)}$ and $\frac{P(B)}{P(A)+P(B)}$ are intuitive guesses.

Example 1
Same Problem with Hidden Information

- As μ is not known, we have to use the MLE as in "EM Algorithm - Example 1"
- We have the following alternatively updating scheme:

Expectation:

The expected value of a and b can be computed as

$$a = \frac{\frac{1}{2}}{\frac{1}{2}+\mu} h \text{ and } b = \frac{\mu}{\frac{1}{2}+\mu} h$$

We must have $a + b = h$

However, μ is not known.

Maximization:

As the value of μ is not known, we can compute the value of μ by

$$\mu = \frac{b+c}{6(b+c+d)}$$

Example 1
Same Problem with Hidden Information

- The technique to solve this type of problem is called **Expectation and Maximization (EM) Algorithm** .
- Similar to K-means clustering algorithm, we alternatively apply the two updating equations until there is no change in the value of μ.

Example 1
Same Problem with Hidden Information

Go back to our problem.

We begin with a guess for μ

- Define $\mu(t)$ the estimate of μ at the t-th iteration

  $b(t)$ the estimate of b at the t-th iteration

At start, we have $\mu(0)$ = initial guess

Expectation step (E-step): $b(t) = \frac{\mu(t)h}{\frac{1}{2}+\mu(t)}$ and $a(t) = h - b(t)$

Maximization step (M-step): $\mu(t+1) = \frac{b(t)+c}{6(b(t)+c+d)}$

Alternatively applying these two steps until no change (converge)

Example 1
Same Problem with Hidden Information

In our example,
suppose we have

$$h = 20$$

$$c = 10$$

$$d = 10$$

$$\mu(0) = 0$$

t	$\boldsymbol{\mu(t)}$	$\boldsymbol{b(t)}$
0	0	0
1	0.0833	2.857
2	0.0937	3.158
3	0.0947	3.185
4	0.0948	3.187
5	0.0948	3.187
6	0.0948	3.187

Stop as no change

Example 1: Summary of EM Algorithm

- Expectation Step (E-Step):

As we require $a + b = h$, we have

$$a = \frac{P(A)}{P(A)+P(B)} h \text{ and } b = \frac{P(B)}{P(A)+P(B)}$$

- Maximization Step (M-Step):

We use MLE to estimate the unknown parameter μ. This is given by

$$\mu = \frac{b + c}{6(b + c + d)}$$

AIC and BIC

- Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used to infer the total number of clusters.
- The number of clusters is the one with minimum values of AIC and BIC.
- AIC:

$$AIC = 2k - 2\ln(L)$$

- BIC:

$$BIC = \ln(n)\,k - 2\ln(L)$$

where L is the maximized value of the likelihood function of the model M

n is the sample size and k is the number of clusters.

AIC and BIC

- That is, we have to apply the EM algorithm with different number of clusters $k = 1,2,3$ ...
- The one with minimum values of AIC and BIC is the number of clusters.