CS/ENGR M148 L13:
Hidden Markov Models and Expectation Maximization

Sandra Batista

Administrative News

This week in discussion section:

No lab this week.

Project check-in this week on unsupervised learning.

Tas will share guidelines for data pre-processing during discussion

Midterm grades posted. Grade distribution on piazza. Regrade requests due by 6 pm 11/20/24.

Thanks for your helpful response to survey!

PS3 data posted.

Survey Results

Attempts: 128 out of 129

How are the pace and level of lecture content?

Too fast or too much content	31 respondents	**24 %**
Just about right	80 respondents	62 %
Too slow or want more content	17 respondents	13 %
No Answer	1 respondent	1 %

Attempts: 129 out of 129

How are you finding the lab and discussion section pace?

Too fast or too much content	27 respondents	**21 %**
Just right	96 respondents	74 %
Too slow or want more content	6 respondents	5 %

Survey Results

Attempts: 129 out of 129

How are you and your group finding the nearly weekly project check-in?

Very useful	35 respondents	**27 %**
Neutral	73 respondents	57 %
Not useful	21 respondents	16 %

Attempts: 129 out of 129

What is your experience and perception of the difficulty level of the problem sets (coding and written problems)?

Too difficult	32 respondents	**25 %**
Just right	95 respondents	74 %
Want more challenge	2 respondents	2 %

Survey Results

Attempts: 129 out of 129

What is your perception of the difficulty of midterm?

Too difficult	78 respondents	**60 %**
Fair	51 respondents	40 %
Not challenging enough		0 %

Attempts: 129 out of 129

Do you feel that you had sufficient background in mathematics (especially linear algebra), statistics, and computer science for the course?

Extremely comfortable with knowledge applied in course	22 respondents	**17 %**
comfortable with knowledge applied in course	86 respondents	67 %
Wish there were more prerequisites	21 respondents	16 %

Survey Response Anonymous comments

- Paraphrasing because I did not ask permission to share anonymous comments
- **Review before lecture**: Some students find it useful, some only want new material
- **Sincere apology**: Someone expressed concern that they did not feel I was helpful or respectful to them during office hours. Please accept this apology. I would like to improve so no students feel that way.

Survey Response Anonymous comments

- **Generally positive**
- **Midterm too challenging**:

1. Shared solutions
2. Shared from where in course materials problems came
3. Considering alternative grading scheme
4. Advocating for P/NP changes, but told unlikely to be approved in Engineering

- **Not enough practice problems**:

1. Previous instructor explicitly asked us not to share previous exams.
2. Extra credit opportunity: We will credit a question bank from which you can write and submit practice problems
3. Course staff will screen problems and give a set as practice problems for the final

Join our slido for the week…

https://app.sli.do/event/nCV57u4mC7eUMit9euSBr2

Today's Learning Objectives

Students will be able to:

- Review: Clustering
- Hidden Markov Models (HMM)
- Inference with HMM: Viterbi Algorithm
- Expectation-Maximization (EM) Algorithm
- EM and Clustering: Gaussian Mixture Models

Agglomerative clustering

1. Use any computable cluster similarity measure $sim(C_i, C_j)$ e.g., Euclidean distance,
2. For n objects $v_1, \ldots, v_n$, assign each to a singleton cluster $C_i = \{v_i\}$
3. Repeat {

   - identify two most similar clusters C_j and C_k (could be ties-chose one pair)

   - delete C_j and C_k and add $(C_j \cup C_k)$ to the set of clusters.

   } until just one cluster.
4. Dendrograms diagram the sequence of cluster merges.

[Shah 2020]

Divisive clustering

1. Put all objects in one cluster
2. Repeat until all clusters are singletons {

- choose a cluster to split based on some criterion.
- replace the chosen cluster with sub-clusters.

}

[Shah 2020]

K-means algorithm

1. Begin with a decision on the value of K = number of clusters.
2. Put any initial partition that classifies the data into K clusters. You may assign the training samples randomly or systematically.
3. Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

Repeat the above three steps until convergence is achieved.

[Shah 2020]

How good is clustering?

Within-Cluster Sum of Squares:

$$WSS = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{i,j} - c_{k,j})^2$$

For a single data point:

Let a_i be average distance from point i to other points in same cluster

Let b_i be average distance from point i to points in nearest cluster

Silhouette score:

$$\text{silhouette score}_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

[Yu, Barter 2024]

Rand Index

How to compare different sets of clusters?

For each pair of data points, consider if the clusters put the Same pair of points in the same cluster or different clusters.

Rand Index = # pairs in agreement/ total # of pairs

[Yu, Barter 2024]

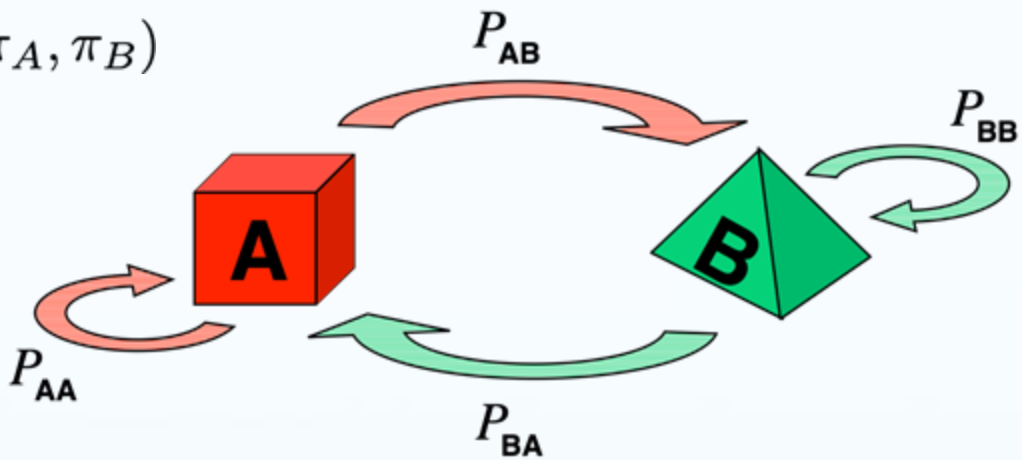
Today's Learning Objectives

Students will be able to:

- ✓ Review: Clustering
- Hidden Markov Models (HMM)
- Inference with HMM: Viterbi Algorithm
- Expectation-Maximization (EM) Algorithm
- EM and Clustering: Gaussian Mixture Models

A two state Markov chain

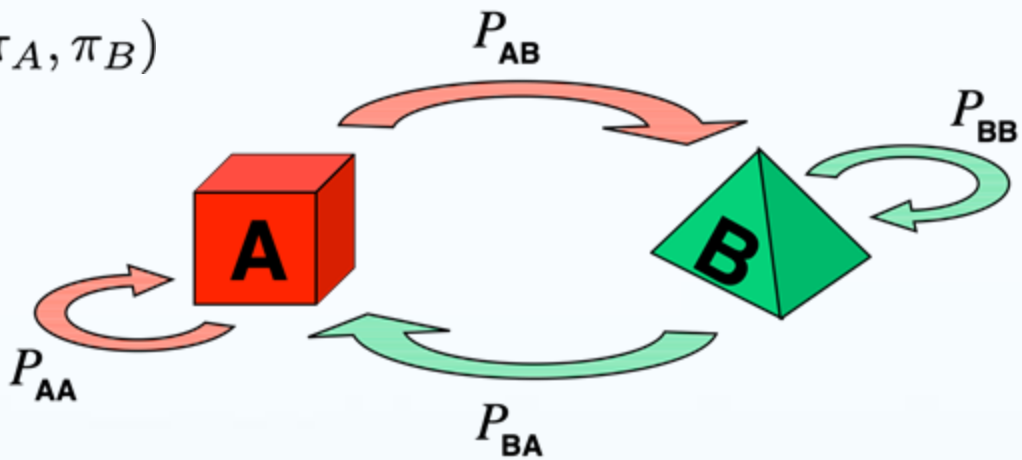
$$\pi = (\pi_A, \pi_B)$$



$$\Pr(X_0 = A) = \pi_A \qquad \Pr(X_{n+1} = A | X_n = A) = p_{AA}$$
$$\Pr(X_0 = B) = \pi_B \qquad \Pr(X_{n+1} = B | X_n = A) = p_{AB}$$
$$\Pr(X_{n+1} = A | X_n = B) = p_{BA}$$
$$\Pr(X_{n+1} = B | X_n = B) = p_{BB}$$

A two state Markov chain

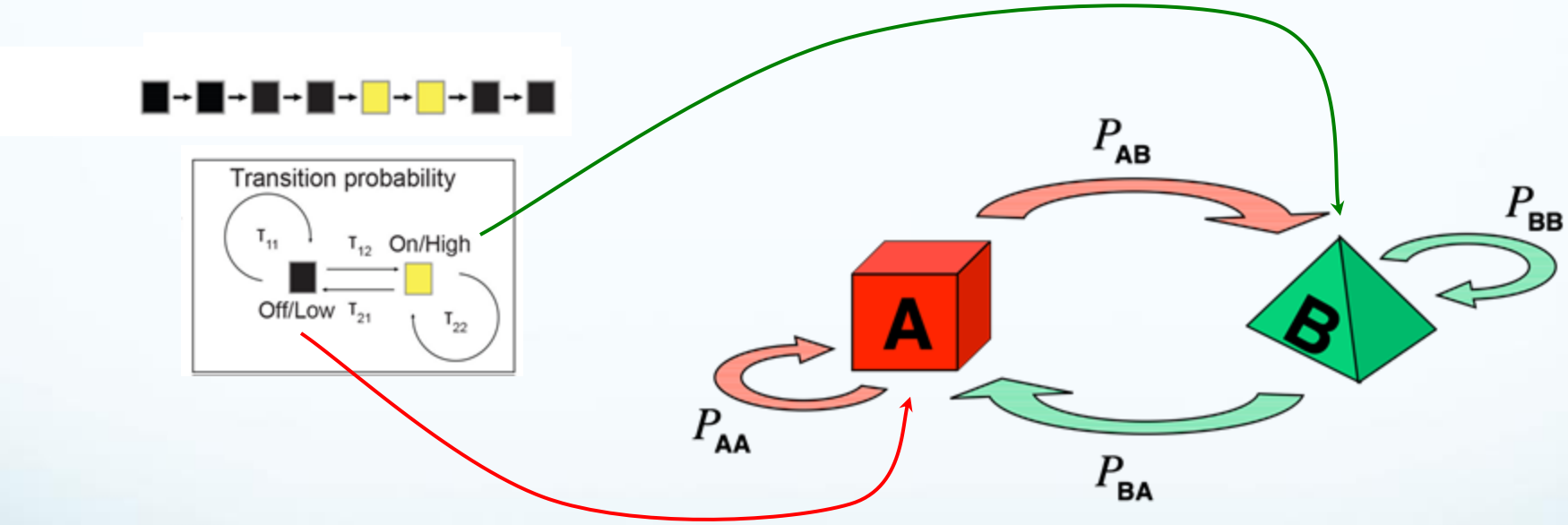
$$\pi = (\pi_A, \pi_B)$$



$$\Pr(X_0 = B, X_1 = A, X_2 = A) = \pi_B \cdot p_{BA} \cdot p_{AA}$$

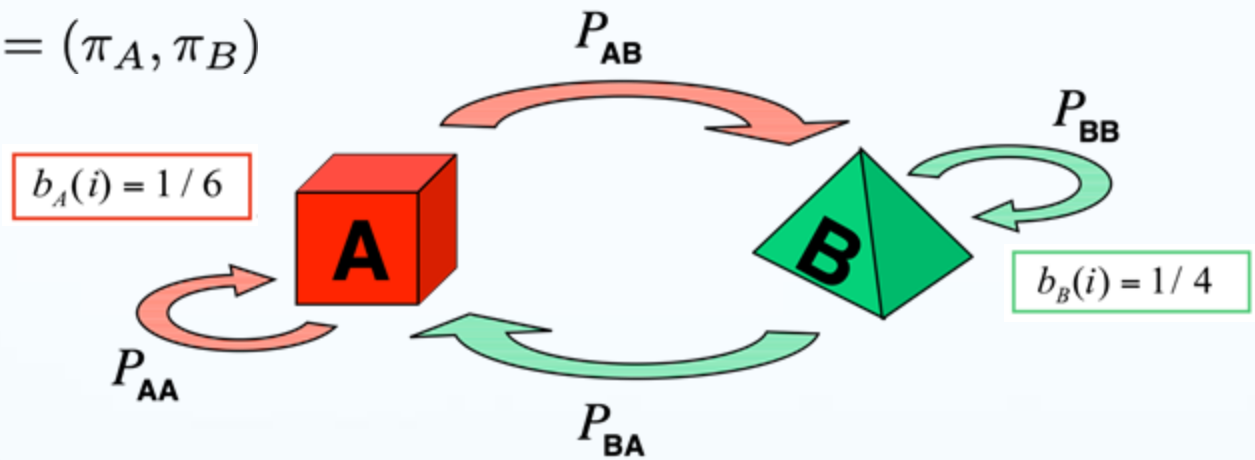
$$\Pr(X_0 = B, X_1 = A, X_2 = A, X_3 = B, X_4 = A, X_5 = A) = \pi_B \cdot p_{BA}^2 \cdot p_{AA}^2 \cdot p_{AB}$$

A two-state gene expression model



A hidden Markov model

$$\pi = (\pi_A, \pi_B)$$



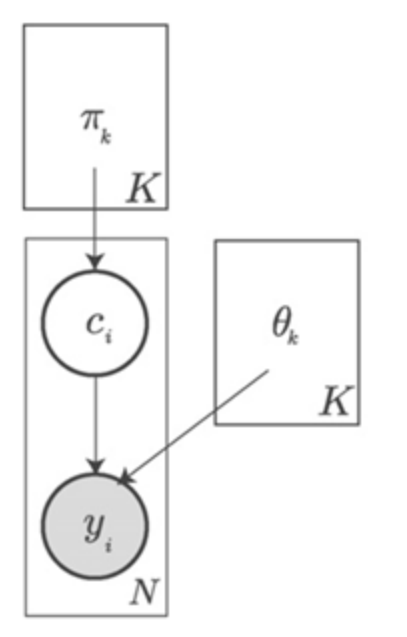
$$\Pr(X_0 = A) = \pi_A$$
$$\Pr(X_0 = B) = \pi_B$$

$$\Pr(X_{n+1} = A | X_n = A) = p_{AA}$$
$$\Pr(X_{n+1} = B | X_n = A) = p_{AB}$$
$$\Pr(X_{n+1} = A | X_n = B) = p_{BA}$$
$$\Pr(X_{n+1} = B | X_n = B) = p_{BB}$$

$$\Pr(Y_i = k | X_i = A) = \frac{1}{6}$$
$$\Pr(Y_i = k | X_i = B) = \frac{1}{4}$$

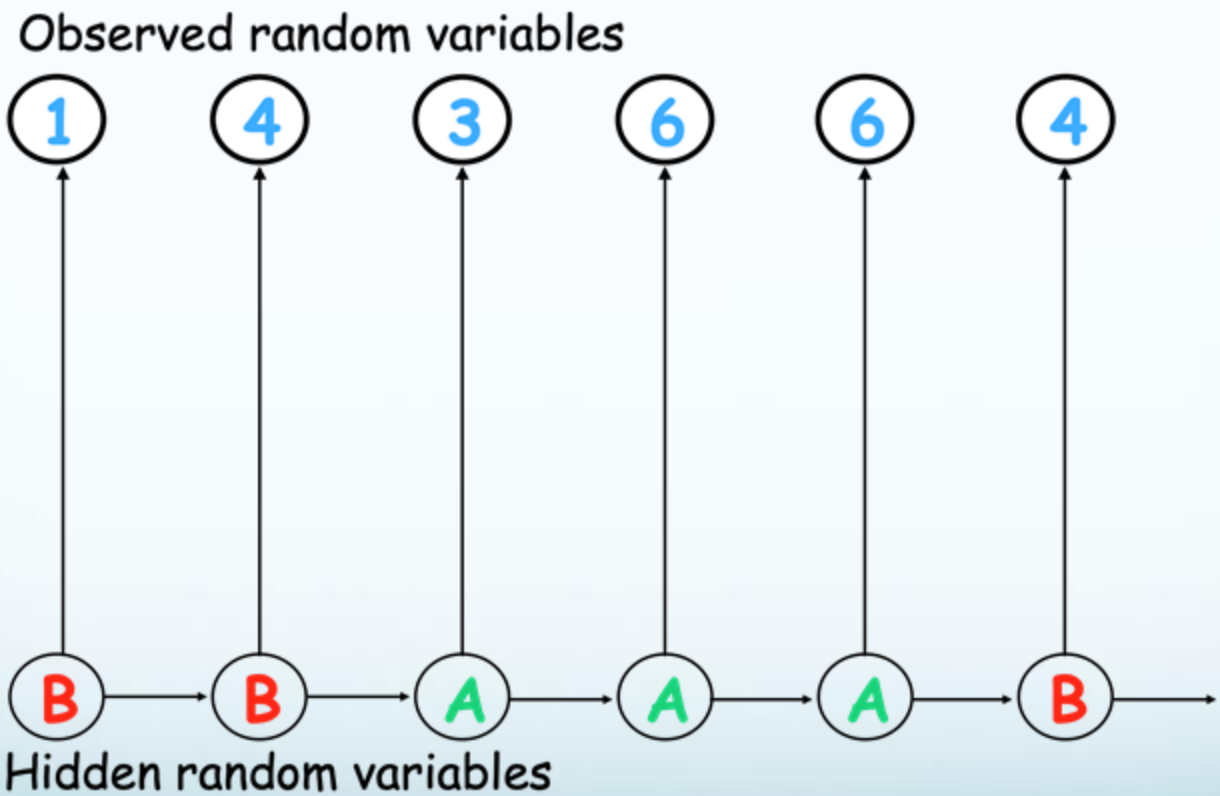
Graphical Models

- Graphical models describe models with latent variables.
- Shaded circles are observed random variables; unshaded circles are latent random variables.
- Parameters are shown in boxes.
- Numbers in the bottom right of each box indicate the number of copies (these are called plates).
- The edges encode conditional independence.



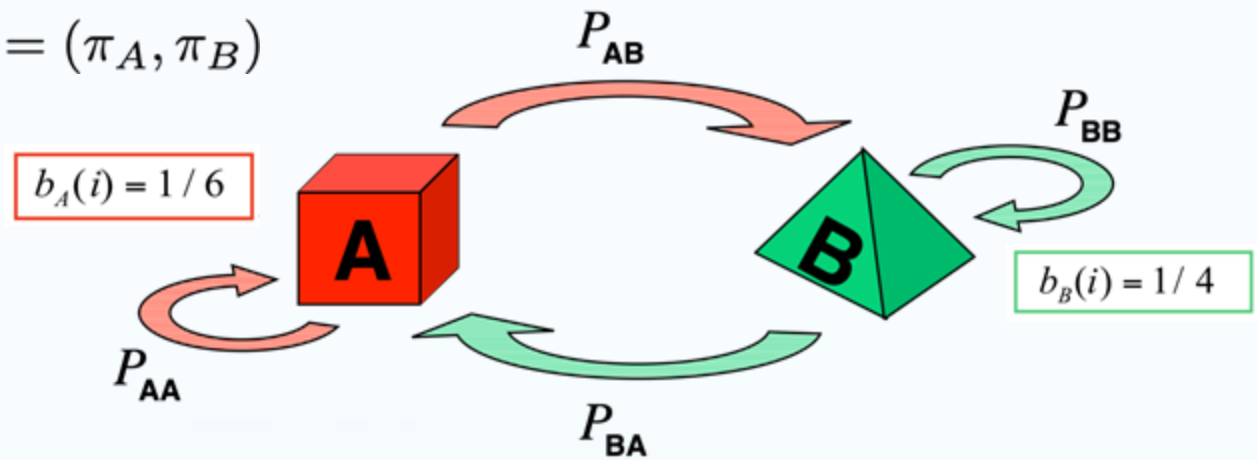
Wood and Black, 2008

Hidden Markov models (HMMs) as graphical models



A hidden Markov model

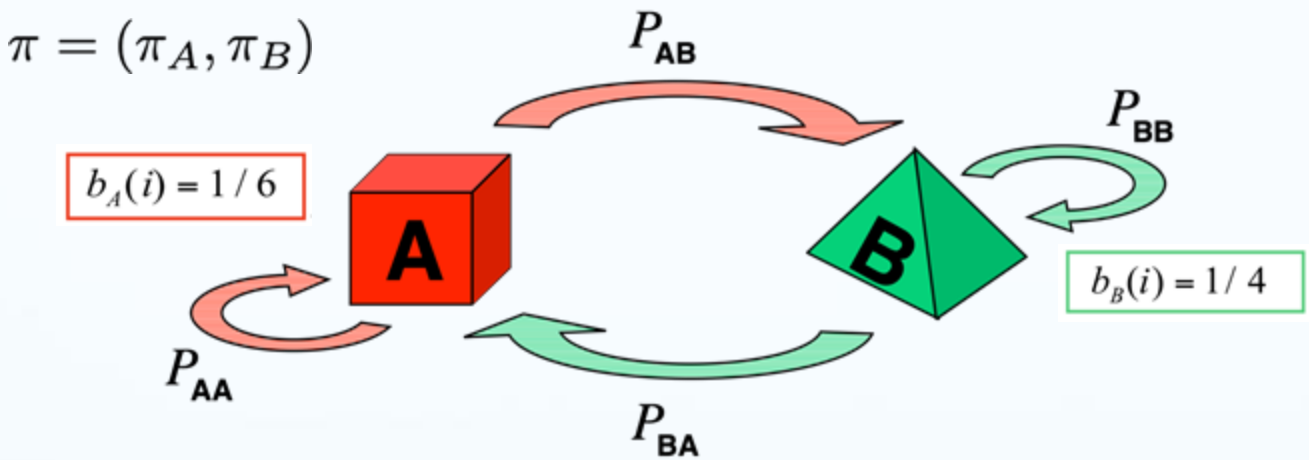
$\pi = (\pi_A, \pi_B)$



$$\Pr(Y_0 = 1, Y_1 = 4, Y_2 = 3, Y_3 = 6, Y_4 = 6, Y_5 = 4) = ?$$

Solved with the forward algorithm.

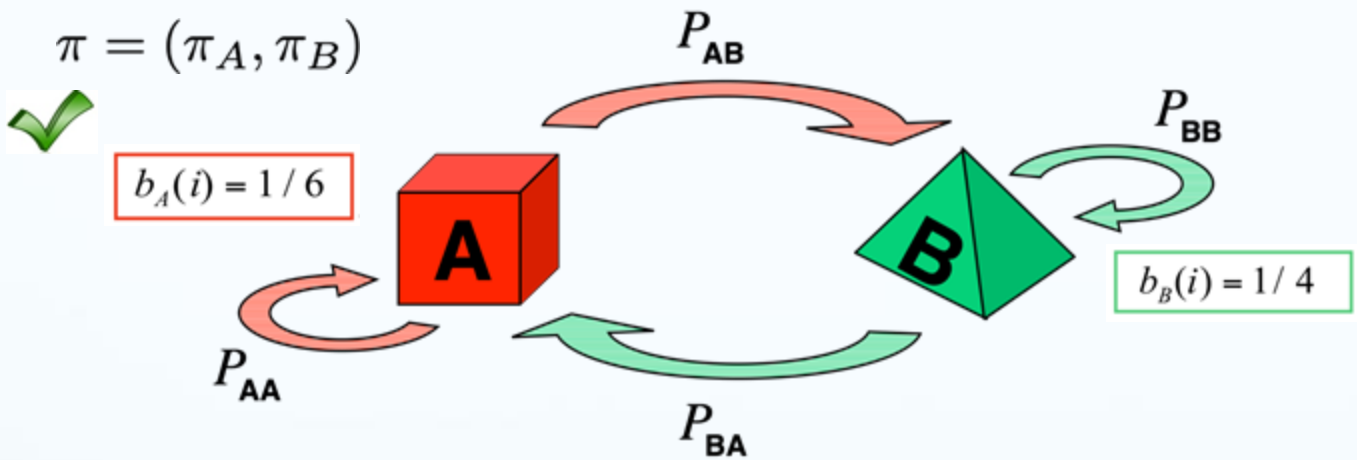
A hidden Markov model



What is the most likely sequence of states of the Markov chain to have resulted in 1,4,3,6,6,4?

Solved with the Viterbi algorithm.

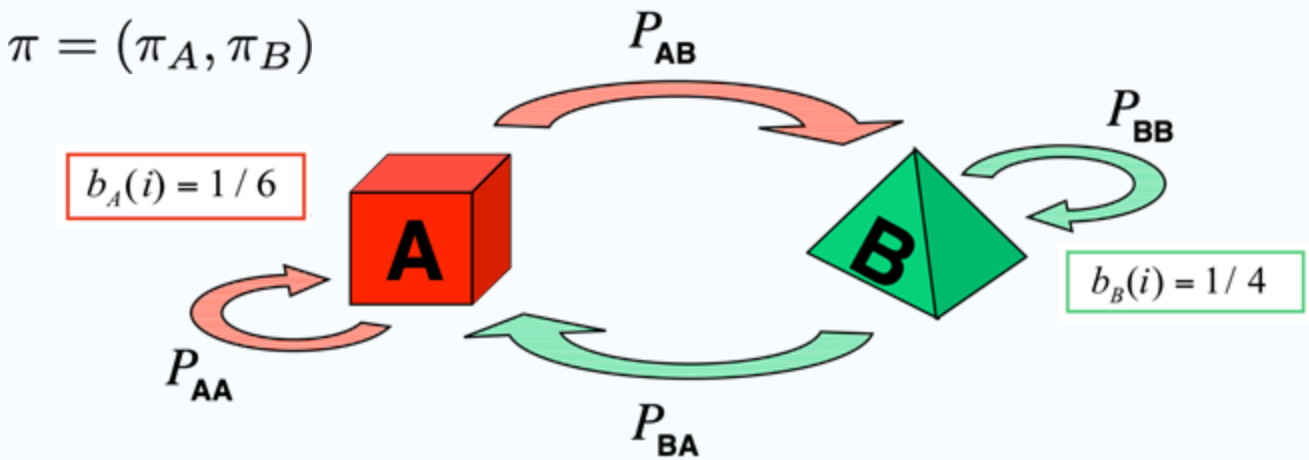
A hidden Markov model



What is the probability that $X_3 = B$ if $Y_0=1, Y_1=4, Y_2=3, Y_3=6, Y_4=6, Y_5=4$?

Solved with the forward-backward algorithm.

A hidden Markov model



*Given multiple sequences of numbers (observations of **Y**), estimate parameters for the model*

This is expectation-maximization algorithm

Today's Learning Objectives

Students will be able to:

- ✓ Review: Clustering
- ✓ Hidden Markov Models (HMM)
- Inference with HMM: Viterbi Algorithm
- Expectation-Maximization (EM) Algorithm
- EM and Clustering: Gaussian Mixture Models

Hidden Markov Models

- *Components:*
  - *Observed variables*
    - *Emitted symbols*
  - *Hidden variables*
  - *Relationships between them*
    - *Represented by a graph with transition probabilities*
- ***Goal:** Find the most likely explanation for the observed variables*

The occasionally dishonest casino

A casino uses a fair die most of the time, but occasionally switches to a loaded one

Fair die: Prob(1) = Prob(2) = . . . = Prob(6) = 1/6

Loaded die: Prob(1) = Prob(2) = . . . = Prob(5) = 1/10,

Prob(6) = ½

These are the ***emission*** probabilities

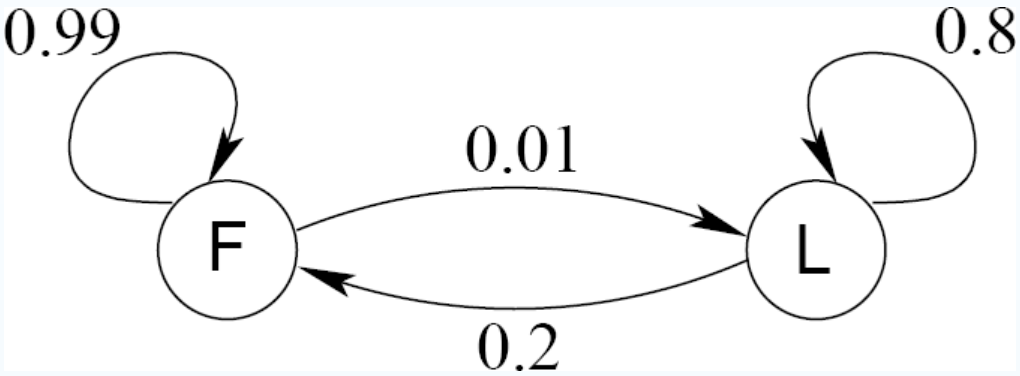
Transition probabilities

Prob(Fair –> Loaded) = 0.01

Prob(Loaded –>Fair) = 0.2

Transitions between states obey a Markov process

An HMM for the occasionally dishonest casino



The occasionally dishonest casino

- *Known:*
  - *The structure of the model*
  - *The transition probabilities*
- *Hidden: What the casino did*
  - *FFFFFLLLLLLFFFF...*
- *Observable: The series of die tosses*
  - *3415256664666153...*
- *What we must infer:*
  - *When was a fair die used?*
  - *When was a loaded one used?*
    - *The answer is a sequence FFFFFFFLLLLLFFF...*

Making the inference

Model assigns a probability to each explanation of the observation:

$P(326|FFL)$

$= P(3|F) \cdot P(F \rightarrow F) \cdot P(2|F) \cdot P(F \rightarrow L) \cdot P(6|L)$

$= 1/6 \cdot 0.99 \cdot 1/6 \cdot 0.01 \cdot ½$

Maximum Likelihood: Determine which explanation is most likely

–Find the path ***most likely*** to have produced the observedsequence

Total probability: Determine probability that observed sequence was produced by the HMM

- Consider ***all*** paths that could have produced the observed sequence

Notation

- x is the sequence of symbols emitted by model
  - x_i is the symbol emitted at time i
- A **path**, π, is a sequence of states
  - The i-th state in π is π_i
- a_{kr} is the probability of making a transition from state k to state r:

$$a_{kr} = Pr(\pi_i = r \mid \pi_{i-1} = k)$$

- $e_k(b)$ is the probability that symbol b is emitted when in state k

$$e_k(b) = Pr(x_i = b \mid \pi_i = k)$$

The occasionally dishonest casino

$$x = \langle x_1, x_2, x_3 \rangle = \langle 6,2,6 \rangle$$

$$\pi^{(1)} = FFF$$

$$\pi^{(2)} = LLL$$

$$\pi^{(3)} = LFL$$

The most probable path

The most likely path π^* satisfies

$$\pi^* = \arg\max_{\pi} Pr(x, \pi)$$

To find π^*, consider all possible ways the last symbol of x could have been emitted

Let

$$v_k(i) = \text{Prob. of path } \langle \pi_1, \cdots, \pi_i \rangle \text{ most likely to emit } \langle x_1, \cdots, x_i \rangle \text{ such that } \pi_i = k$$

Then

$$v_k(i) = e_k(x_i) \max_r (v_r(i-1) a_{rk})$$

The Viterbi Algorithm

- *Initialization* $(i = 0)$

$$v_0(0) = 1, \quad v_k(0) = 0 \text{ for } k > 0$$

- *Recursion* $(i = 1, \ldots, L)$: *For each state* k

$$v_k(i) = e_k(x_i) \max_r \left(v_r(i-1) a_{rk} \right)$$

- *Termination:*

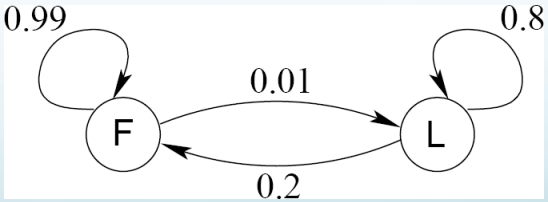
$$Pr(x, \pi^*) = \max_k \left(v_k(L) a_{k0} \right)$$

To find π^**, use trace-back, as in dynamic programming*

Viterbi: Example

		x		
	ε	6	2	6
π B	1	0	0	0
π F	0			
π L	0			

$$v_k(i) = e_k(x_i)\max_r(v_r(i-1)a_{rk})$$



Total probabilty

Many different paths can result in observation x.

The probability that our model will emit x is

$$\Pr(x) = \sum_{\pi} \Pr(x, \pi)$$

Total Probability

If HMM models a family of objects, we want total probability to peak at members of the family. (Training)

Total probability

$\Pr(x)$ can be computed in the same way as probability of most likely path.

Let

$$f_k(i) = \text{Prob. of observing } \langle x_1, \ldots, x_i \rangle$$
$$\text{assuming that } \pi_i = k$$

Then

$$f_k(i) = e_k(x_i) \sum_r f_r(i-1) a_{rk}$$

and

$$\Pr(x) = \sum_k f_k(L) a_{k0}$$

The Forward Algorithm

- *For next time, consider how to calculate probability of sequence and being in state at time i….*
- *How does this related to the Viterbi algorithm?*

Backward Algorithm and Posterior Decoding

- *For next time, consider how to calculate probability of of the rest of the sequence and being in state at time i….*

- *Also consider how we can calculate posterior probabilities*

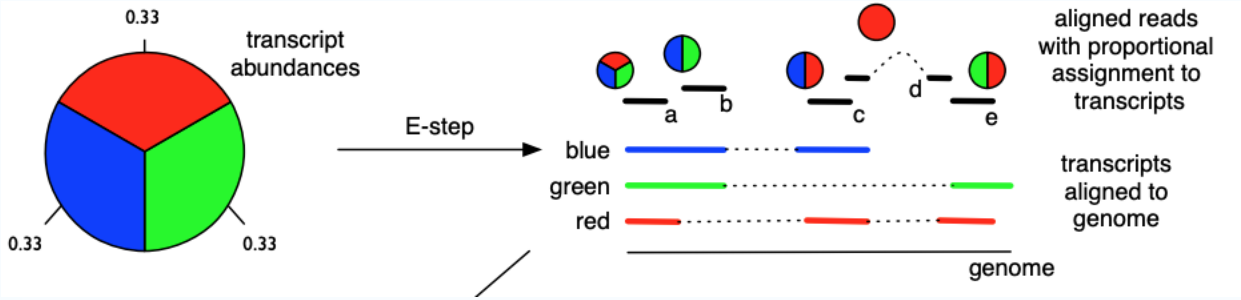
$$P(\pi_i = k \mid x) = \frac{f_k(i) \cdot b_k(i)}{P(x)}$$

Today's Learning Objectives

Students will be able to:

- ✓ Review: Clustering
- ✓ Hidden Markov Models (HMM)
- ✓ Inference with HMM: Viterbi Algorithm
- Expectation-Maximization (EM) Algorithm
- EM and Clustering: Gaussian Mixture Models

Expectation-Maximization

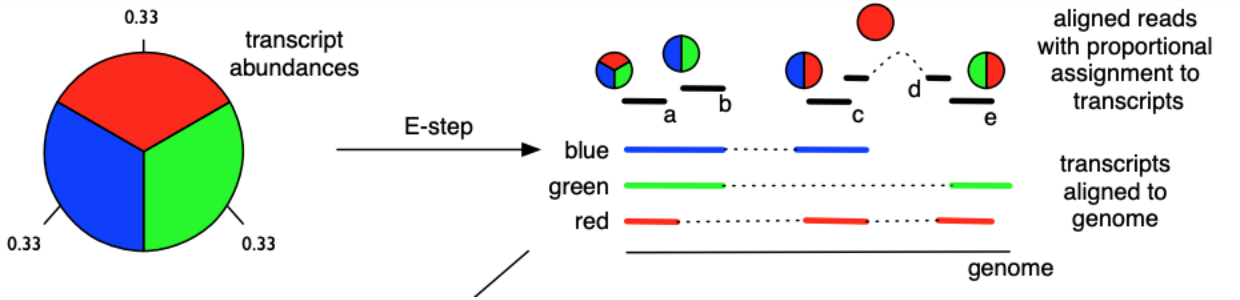


*Given multiple sequences of numbers (observations of **Y**), estimate parameters for the model*

<u>Goal: Find parameters the maximize likelihood of data observed</u>

This is expectation-maximization algorithm

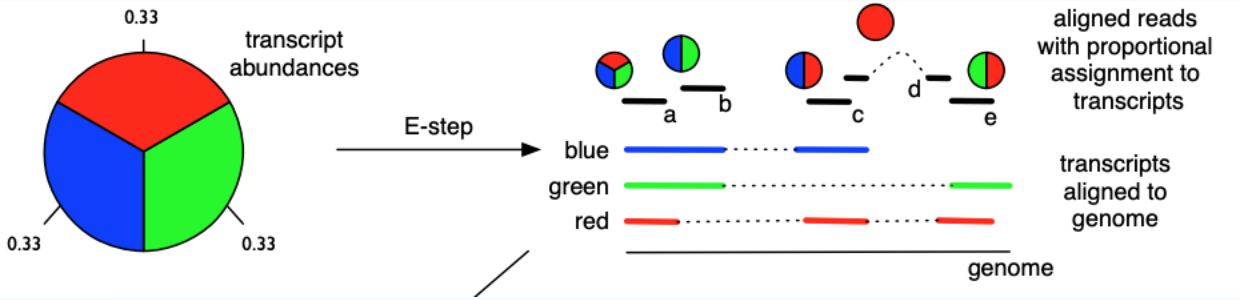
Expectation-Maximization



Given 5 sequences and we want to see how they align to 3 different genes: red, blue and green.

Assumption: reads are generated based on how much of gene there is, transcript abundance

Expectation-Maximization



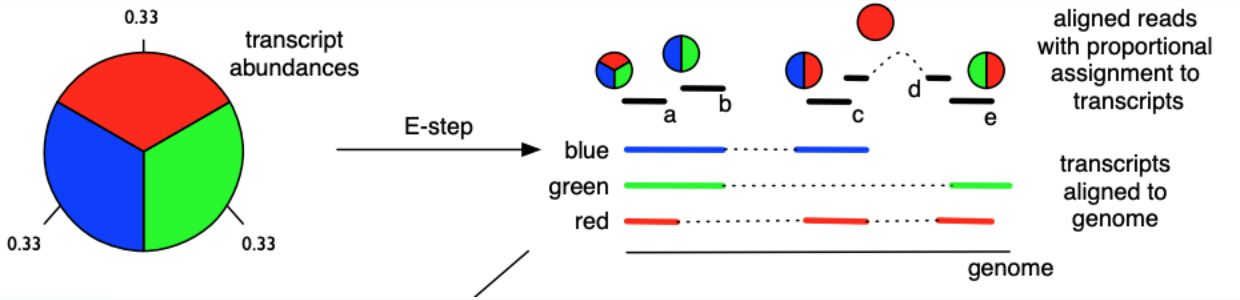
Given 5 sequences and we want to see how they align to 3 different genes: red, blue and green.

What is hidden?

What is observed?

What are parameters?

Expectation-Maximization



Observations of reads aligned

	a	b	c	d	e
red	1	0	1	1	1
green	1	1	0	0	1
blue	1	1	1	0	0

Latent variables and likelihood of data

- Use latent variables to express likelihood of data given parameters:

$$Pr(Y|\alpha) = \sum_{i=1}^{k} Pr(Z = i|\alpha) Pr(Y|Z = i, \alpha)$$

Maximizing the likelihood function

- The expectation of the log likelihood function is iteratively maximized:

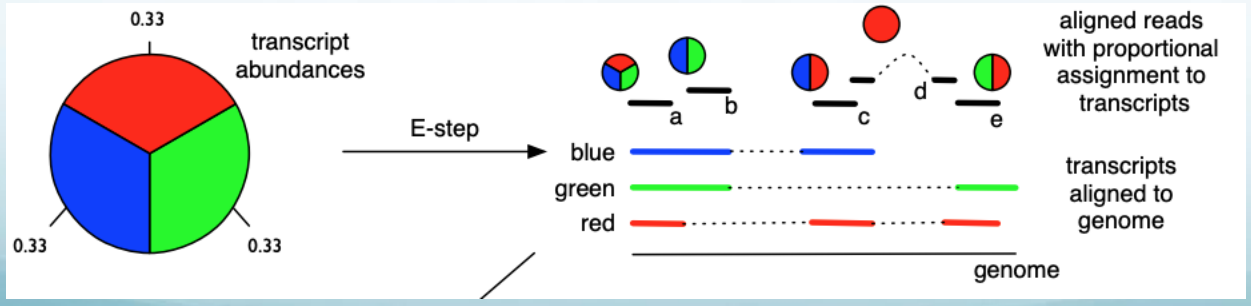
$$Q(\alpha|\alpha^T) = E_{Z|Y,\alpha^T}[\log Pr(Y, Z|\alpha)]$$

1. Computing the expected value of *z*.
2. Maximizing the likelihood conditioned on z.

Maximizing the likelihood function

- What is this for our transcript example?

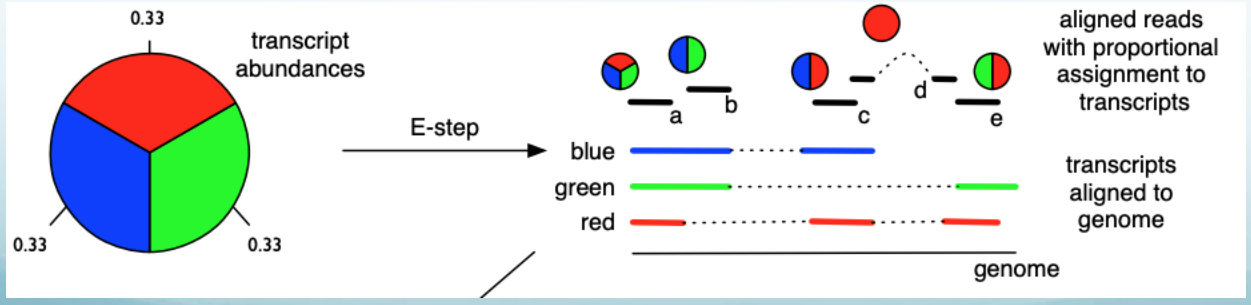
$$Q(\alpha|\alpha^T) = \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{y_{k,n}\alpha_k^{(t)}}{\sum_{l=1}^{K} y_{ln}\alpha_l^{(t)}}\log(y_{k,n}\alpha_k)$$



Expectation step

- For each read and transcript, calculate likelihood of gene given data and current alpha

$$p(Z_n = k | Y_n; \alpha^{(t)}) = \frac{y_{k,n}\alpha_k^{(t)}}{\sum_{l=1}^{K} y_{l,n}\alpha_l^{(t)}}$$



Expectation step

- What does this look like if we assume to start all transcripts equally likely?

$$p(Z_n = k | Y_n; \alpha^{(t)}) = \frac{y_{k,n}\alpha_k^{(t)}}{\sum_{l=1}^{K} y_{l,n}\alpha_l^{(t)}}$$

	a	b	c	d	e
red	1	0	1	1	1
green	1	1	0	0	1
blue	1	1	1	0	0

Expectation step

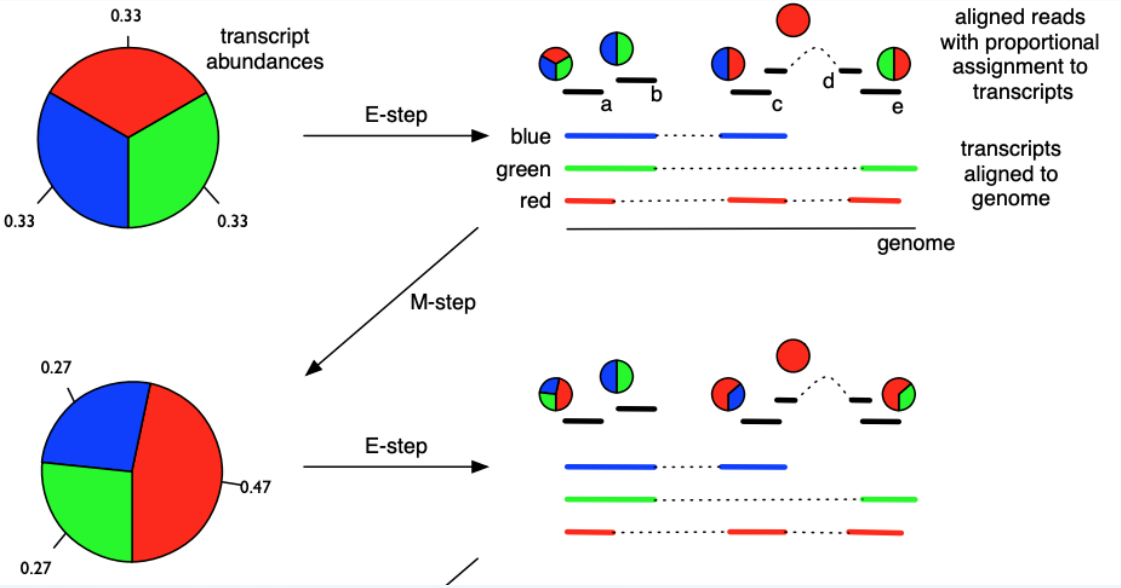
- What does this look like if we assume to start all transcripts equally likely?

$$p(Z_n = k | Y_n; \alpha^{(t)}) = \frac{y_{k,n}\alpha_k^{(t)}}{\sum_{l=1}^{K} y_{l,n}\alpha_l^{(t)}}$$

	a	b	c	d	e
red	0.333333	0.0	0.5	1.0	0.5
green	0.333333	0.5	0.0	0.0	0.5
blue	0.333333	0.5	0.5	0.0	0.0

Maximization Step

- Choose the new parameters that maximize the log likelihood



$$\alpha^{t+1} = arg\_max_{\alpha} Q(\alpha|\alpha^{T})$$

Maximization Step

- Choose the new parameters that maximize the log likelihood

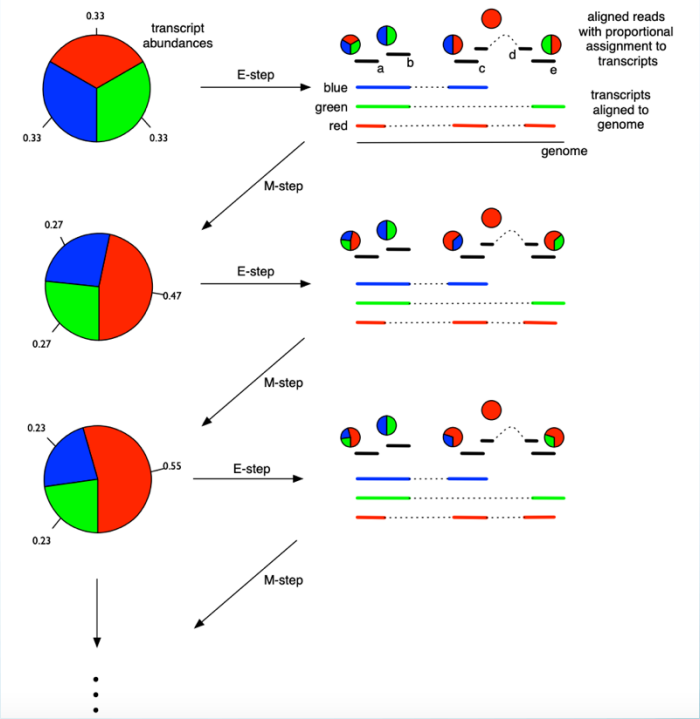
$$\alpha_k^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} \frac{y_{k,n} \alpha_k^{(t)}}{\sum_{l=1}^{K} y_{l,n} \alpha_l^{(t)}} \quad .$$

	a	b	c	d	e
red	0.333333	0.0	0.5	1.0	0.5
green	0.333333	0.5	0.0	0.0	0.5
blue	0.333333	0.5	0.5	0.0	0.0

```
[0.46666667 0.26666667 0.26666667]
```

When does the EM algorithm stop?

- The algorithm has the property that the log likelihood is non-decreasing (subject to the assumption that the hidden variables do not have 0 probability).
- A stopping criteria is usually based on measuring the incremental improvement in the log likelihood or a number of iterations.
- The algorithm guarantees convergence a local maxima of the likelihood function but not necessarily a global maximum.



Today's Learning Objectives

Students will be able to:

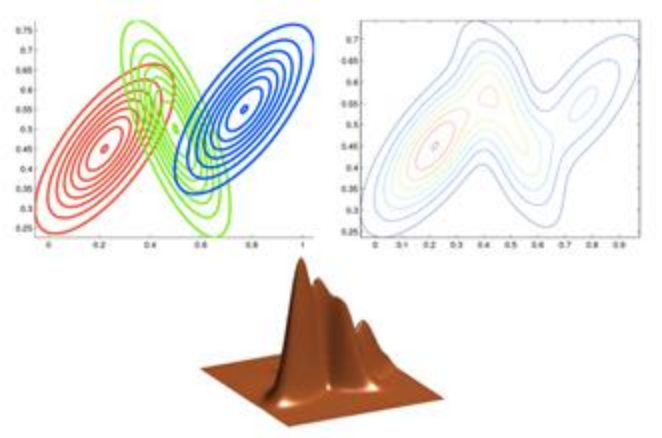
- ✓ Review: Clustering
- ✓ Hidden Markov Models (HMM)
- ✓ Inference with HMM: Viterbi Algorithm
- ✓ Expectation-Maximization (EM) Algorithm
- EM and Clustering: Gaussian Mixture Models

Gaussian Mixture Models

- A Gaussian mixture model is a family of distributions of the form

$$p(\mathbf{x}) = \sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i).$$

- The π_i are *mixing coefficients* that sum to 1.
- Estimate the mean and covariance matrices from data as parameters for GMM from data
- Each Gaussian distribution represents a cluster.
- This provides a generative model for clustering.



Zemel et al., 2016

Gaussian Mixture Model

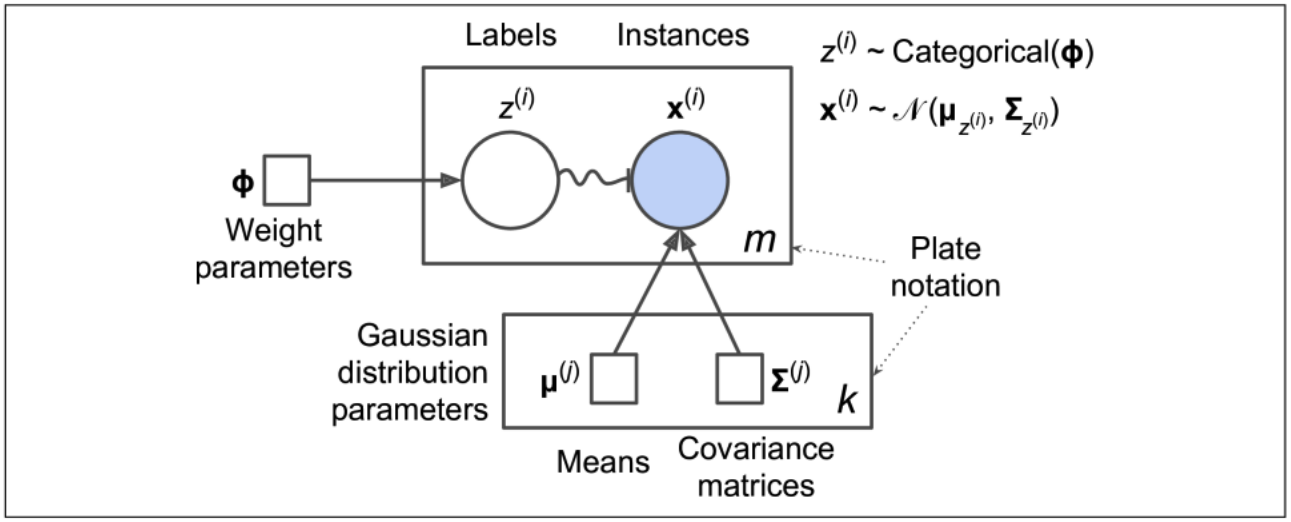


Figure 9-16. Gaussian mixture model

Aurelien Geron. 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd. ed.). O'Reilly Media, Inc

Maximum likelihood estimation

- The log likelihood function that must be maximized is:

$$\log Pr(x|\pi, \mu, \sigma) = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{k} \pi N(x|\mu_i, \Sigma_i) \right)$$

- Expectation Step: Calculate clusters
- Maximization Step: Update parameters

Shortcoming of EM algorithm for clustering

- How to choose the number of clusters (k)?
- The EM algorithm can get stuck in local maxima; initialization is corncer
- EM algorithm may be slow.
- If data not from a Gaussian mixture process, EM algorithm may converge to a poor solution.

Your turn:
GMM on GPU data

Please get the Jupyter notebook for GPU data:

Go to:

The data file on BruinLearn Week 7 Module:

- `sgemm_product.csv`

Notebook:
https://colab.research.google.com/drive/1iqVnrE7LKQyW_UXExzV3Q06EP10Q2Bvk?usp=sharing

Save a copy to your Google Drive and keep notes there...

Today's Learning Objectives

Students will be able to:

- ✓ Review: Clustering
- ✓ Hidden Markov Models (HMM)
- ✓ Inference with HMM: Viterbi Algorithm
- ✓ Expectation-Maximization (EM) Algorithm
- EM and Clustering: Gaussian Mixture Models

Citations:

Shah. C. (2020) A hands-on introduction to data science. Cambridge University Press.

Yu, B., & Barter, R. L. (2024). Veridical data science: The practice of responsible data analysis and decision making. The MIT Press.

Shah. C. (2020) A hands-on introduction to data science. Cambridge University Press.

Thank You