Dynamic non-parametric Mixture Models: the Recurrent Chinese Restaurant Process

Amr Ahmed and Eric Xing
School of Computer Science
Carnegie Mellon University
Clustering

• Important unsupervised learning problem
• Discovers latent structure in the data

Document collection

Topics/Themes

fMRI scan

Mental Processes

And Many Other applications
Clustering Dynamics

- The world is **dynamic**
- Data points arrive **sequentially**
- Number of clusters changes over time
- Cluster/theme representation **evolves** as well
Dynamic non-parametric Mixture Models

- Both the weights and the atoms are changing over time
- Different chains have different life span
- The DP of the next time-step depends on the samples observed in the previous time
- ...
The Big Picture

K-means

Fixed-dimensions Dynamic clustering

DPM

TDPM

Model Dimension

Time
K-Means: The Generative Process

- For data point $x_i$
  - Sample $c_i \sim \text{Multi}(\pi)$
  - Sample $x_i \sim f(\phi_{c_i})$

Cluster parameters DO NOT evolve over time
Number of clusters DOES NOT grow with the data
The Big Picture

K-means

\[ \pi \rightarrow C_i \rightarrow x_i \]

\[ \phi \rightarrow K \]

Fixed-dimensions Dynamic clustering

\[ \phi_{k,1} \rightarrow \phi_{k,2} \rightarrow \ldots \rightarrow \phi_{k,T} \]

\[ x_{1,n} \rightarrow c_{1,n} \rightarrow \pi_1 \]

\[ x_{2,n} \rightarrow c_{2,n} \rightarrow \pi_2 \]

\[ \ldots \]

\[ x_{T,n} \rightarrow c_{T,n} \rightarrow \pi_T \]

DPM

\[ \alpha \rightarrow \pi \rightarrow C_i \rightarrow x_i \]

\[ G_0 \rightarrow \phi \rightarrow \infty \]

TDPM

\[ G_{1,2} \rightarrow \phi_{1,2} \rightarrow \ldots \rightarrow \phi_{k,T} \]

\[ x_{1,n} \rightarrow c_{1,n} \rightarrow \pi_1 \]

\[ x_{2,n} \rightarrow c_{2,n} \rightarrow \pi_2 \]

\[ \ldots \]

\[ x_{T,n} \rightarrow c_{T,n} \rightarrow \pi_T \]

Time Model Dimension

Model Dimension Time
Modeling Clustering Dynamics

Generative Process

- For each time epoch $t$
  - Sample $\phi_{k,t} \sim P(\cdot | \phi_{k,t-1})$
  - For each data item at time $t$
    - Sample $c_{t,i} \sim \text{Multi}(\pi_t)$
    - Sample $x_{t,i} \sim f(\phi_c)$

- The parameters of each mixture change over time
- Number of mixture components DOES NOT change over time!
The Big Picture

K-means

DPM

Fixed-dimensions Dynamic clustering

TDPM
Dirichlet Process Mixture Model

• Three equivalent constructions:

Measure over Measures

Chinese Restaurant Process

The Stick-breaking construction
The Chinese Restaurant Process

- Customers correspond to data points
- Tables correspond to clusters/mixture components
- Dishes correspond to parameter of the mixtures
The Chinese Restaurant Process

For data point $x_i$
- Choose table $j \propto N_j$ and Sample $x_i \sim f(\phi_j)$
- Choose a new table $K+1 \propto \alpha$
  - Sample $\phi_{K+1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1})$

Generative Process

- The rich gets richer effect
CANNOT handle sequential data
The Big Picture

K-means

\[ \pi \]
\[ C_i \]
\[ x_i \]

Fixed-dimensions Dynamic clustering

\[ \phi_{k,1} \]
\[ \phi_{k,2} \]
\[ \phi_{k,T} \]

\[ x_{1,n} \]
\[ x_{2,n} \]
\[ x_{T,n} \]

\[ n_1 \]
\[ n_2 \]
\[ n_T \]

\[ c_{1,n} \]
\[ c_{2,n} \]
\[ c_{T,n} \]

TDPM [Xing 2005, Ahmed and Xing 2008]

\[ G_{10} \]

[Diagram showing dynamic clustering models and time models]
Temporal DPM [Xing 2005, Ahmed and Xing 2008]

- **Adapts the number of mixture components over time** --- a birth-death process
  - Mixture components can **die out**
  - New mixture components can **be born** at any time
  - Retained mixture components **parameters evolve** according to a **Markovian dynamics**
Temporal DPM

- Three equivalent constructions (see [Ahmed & Xing 2008])

Infinite limit of fixed-dimensional dynamic model.

Recurrent Chinese Restaurant

Time-dependent random measures
The Recurrent Chinese Restaurant Process

• The restaurant operates in epochs
• The restaurant is closed at the end of each epoch
• The state of the restaurant at time epoch $t$ depends on that at time epoch $t-1$
  – Can be extended to higher-order dependencies.
The Recurrent Chinese Restaurant Process

- Customers at time $T=1$ are seated as before:
  - Choose table $j \propto N_{j,1}$ and Sample $x_i \sim f(\phi_{j,1})$
  - Choose a new table $K+1 \propto \alpha$
    - Sample $\phi_{K+1,1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1,1})$

Dish eaten at table 3 at time epoch 1
OR the parameters of cluster 3 at time epoch 1
The Recurrent Chinese Restaurant Process

\[ \phi_{1,1}, \phi_{2,1}, \phi_{3,1}, \]  

\[ N_{1,1}=2, \quad N_{2,1}=3, \quad N_{3,1}=1. \]
\[ \phi_{1,1}, \phi_{2,1}, \phi_{3,1} \]

\[ \begin{align*}
T=1 & : N_{1,1}=2 & N_{2,1}=3 & N_{3,1}=1 \\
T=2 & : \frac{2}{6+\alpha} & \frac{3}{6+\alpha} & \frac{1}{6+\alpha} & \frac{\alpha}{6+\alpha} 
\end{align*} \]
$$\frac{2}{6 + \alpha}$$
Sample $\phi_{1,2} \sim P(. | \phi_{1,1})$
\[ \phi_{1,1} \]

\[ \phi_{2,1} \]

\[ \phi_{3,1} \]

\[ \text{T=1} \]

\[ N_{1,1}=2 \]

\[ N_{2,1}=3 \]

\[ N_{3,1}=1 \]

\[ \frac{1+2}{6+1+\alpha} \]

\[ \frac{3}{6+1+\alpha} \]

\[ \frac{1}{6+1+\alpha} \]

\[ \frac{\alpha}{6+1+\alpha} \]

\[ \text{T=2} \]

And so on ....
At the end of epoch 2

Newly born cluster

Died out cluster
Temporal DPM

- Can be extended to model higher-order dependencies
- Can decay dependencies over time
  - Pseudo-counts for table \( k \) at time \( t \) is

\[
\sum_{w=1}^{W} \left( e^{\frac{-w}{\lambda}} N_{k,t-w} \right)
\]
\[
N_{2,3} = \sum_{w=1}^{W} \left( e^{-\lambda w} \frac{e^{\lambda N_{k,t-w}}}{N_{k,t-w}} \right)
\]
TDPM Generative Power

DPM
\[ W = \infty \]
\[ \lambda = \infty \]

TDPM
\[ W = 4 \]
\[ \lambda = 0.4 \]

Independent DPMs
\[ W = 0 \]
\[ \lambda = ? \] (any)
Experiments

• **Simulated data**

• Chain dynamics is modeled as **random walk**

\[
\phi_{k,t} \mid \phi_{k,t-1} \sim N(\phi_{k,t-1}, \rho I)
\]

• **Gaussian emission:**

\[
x_{t,i} \mid c_{t,i} = k \sim N(\phi_k, \Sigma)
\]

• Simulated 30 epochs with 100 data points in each epoch

• Can TDPM recover the **ground truth clustering**?
  – Posterior inference ran using **Gibbs sampling** [Ahmed and Xing 2008]

• Compare with **fixed-dimension dynamic models**
Results

- Variation of information approximates the distance between two clusterings over the lattice of possible clusterings
  - Uses Entropy of and mutual information between the two clusterings
- GVI: applies VI globally by ignoring time
  - Measures global consistency
- LVI: applies VI on each epoch and average the results
  - Measures adaptability
TDPM Adaptability over Time

- Figure (c): Active Chains vs. Epoch
- Figure (e): Frequency vs. Chain Duration

Ground truth and estimated using TDPM are compared. Note the large gap in Figure (e) due to chains with a single data point, as mentioned in the figure caption.
Results: NIPS 12

• Building a **simple** dynamic **topic** model
• Chain dynamics is as before
• Emission model for document $x_{k,t}$ is:
  – Project $\phi_{k,t}$ over the **simplex**
  – Sample $x_{k,t} | c_{t,i} \sim \text{Multinomial}(., \text{Logistic}(\phi_{k,t}))$
• Unlike LDA here a **document** belongs to **one** **topic**
• Use this model to analyze **NIPS12** corpus
Results: NIPS 12

1994
- Training
- speech
- System
- Network
- Data
- Recognition
- Context
- word

1995
- System
- Context
- Network
- Recognition
- Training
- Word
- Model
- probabilities

1997
- Words
- Word
- recognition
- system
- Hmm
- Network
- Spectral
- user

1998
- Words
- Hmm
- Spectral
- User
- System
- Character
- Experts
- frequency

1996
- Data
- model
- learning
- Function
- Algorithm
- Set
- Neural
- Networks
- distribution

1997
- model
- data
- distribution
- algorithm
- learning
- gaussian
- function
- models

1998
- data
- model
- algorithm
- gaussian
- learning
- function
- parameters
- set

1999
- model
- data
- algorithm
- learning
- distribution
- gaussian
- parameters
- models
- mixture

T11: VLSI
- input
- time
- networks
- hopfield
- retrieval
- current
- weight
- voltage

T6: RL
- template
- time
- state
- policy
- input
- channel
- classification
- correlation
- set

T8: Graphical Models
- state
- template
- model
- time
- policy
- channel
- classification
- distribution

T10: Speech
- Chain Index
- Year

T12: Speech
- Chain Index
- Year
Summary

• **Framework** for non-parametric dynamic clustering
  – Number of clusters change over time
  – **Parameters** of each cluster change over time
  – Can be tailored for **different applications**
    • Tracking on 2-d space (Gaussian emission)
    • Topic evolution (Multinomial emission)

• Inference via Gibbs sampling
• Validated empirically over simulated and real data sets.
Questions?
Clustering as Mixture Modeling

• Cluster = Mixture
• Each data point is generated from one Mixture

How Many clusters

- Cross Validation
  - Data hungry!
- Information theoretic
  - AIC, MDL, etc.
- Non-Parametric
  - Manage model uncertainty
  - Integrate over different clustering configurations
  - Base for many interesting extensions
Dirichlet Process Mixture Model

• Allows the number of mixtures to grow with the data
• Also called non-parametric
  – Means the number of effective parameters grow with the data
  – Still have hyper-parameters that control the rate of growth
    • $\alpha$: how fast a new cluster/mixture is born?
    • $G_0$: Prior over mixture component parameters
The Recurrent Chinese Restaurant Process

What is wrong with the CRP?

- Customers SHOULD NOT sit indefinitely in the restaurant
- Equivalently, the same dish SHOULD NOT be eaten indefinitely
Results: NIPS 12

• Building a simple dynamic topic model
• Chain dynamics is as before
• Emission model for document $x_{k,t}$ is:
  – Project $\phi_{k,t}$ over the simplex
  – Sample $x_{k,t} | c_{t,i} \sim \text{Multinomial}(\cdot | \text{Logistic}(\phi_{k,t}))$
• Unlike LDA here a document belongs to one topic
• Use this model to analyze NIPS12 corpus