When is Clustering Hard?

Nati Srebro  
University of Toronto

Gregory Shakhnarovich  
Massachusetts Institute of Technology

Sam Roweis  
University of Toronto
Outline

• Clustering is Hard
• Clustering is Easy
• What we would like to do
• What we propose to do
• What we did
“Clustering”

• Clustering with respect to a specific model / structure / objective

• **Gaussian mixture model**
  – Each point comes from one of \( k \) “centers”
  – Gaussian cloud around each center
  – For now: unit-variance Gaussians, uniform prior over choice of center

• As an optimization problem:
  – Likelihood of centers:
    \[
    \sum_i \log( \sum_j \exp \left(-\frac{(x_i - \mu_j)^2}{2}\right) )
    \]
  – \( k \)-means objective—Likelihood of assignment:
    \[
    \sum_i \min_j (x_i - \mu_j)^2
    \]
Clustering is Hard

- Minimizing $k$-means objective is NP-hard
  - For some point configurations, it is hard to find the optimal solution.
  - But do these point configurations actually correspond to clusters of points?

- Likelihood-of-centers objective probably also NP-hard (I am not aware of a proof)

- Side note: for general metric spaces, hard to approximate $k$-mean to within factor $< 1.5$
“Clustering is Easy”, take 1: Approximation Algorithms

• $(1+\varepsilon)$-Approximation for k-means in time $O(2^{(k/\varepsilon)^{\text{const}}}nd)$ [Kumar Sabharwal Sen 2004]

$$\mu_1 = (5,0,0,0,\ldots,0) \quad 0.5 \ N(\mu_1,I) + 0.5 \ N(\mu_2,I)$$
$$\mu_2 = (-5,0,0,0,\ldots,0)$$

$$\text{cost}([\mu_1,\mu_2]) \approx \sum_i \min_j (x_i-\mu_j)^2 \approx d \cdot n$$
$$\text{cost}([0,0]) \approx \sum_i \min_j (x_i-0)^2 \approx (d+25) \cdot n$$
$$\Rightarrow [0,0] \text{ is a } (1+25/d)-\text{approximation}$$

• Need $\varepsilon < \text{sep}^2/d$, time becomes $O(2^{(kd)^{\text{const}}}n)$
“Clustering is Easy”, take 2: Data drawn from a Gaussian Mixture

\[ x_1, x_2, \ldots, x_n \sim \frac{1}{k} N(\mu_1, \sigma^2 I) + \frac{1}{k} N(\mu_2, \sigma^2 I) + \cdots + \frac{1}{k} N(\mu_k, \sigma^2 I) \]

\[ |\mu_i - \mu_j| > s \cdot \sigma \]

<table>
<thead>
<tr>
<th>Author</th>
<th>Condition on s</th>
<th>Condition on n</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dasgupta 1999</td>
<td>( s &gt; 0.5d^{\frac{1}{2}} )</td>
<td>( n = \Omega(k^{\log^2 \frac{1}{\delta}}) )</td>
<td>Random projection, then mode finding</td>
</tr>
<tr>
<td>Dagupta Schulamn 2000</td>
<td>( s = \Omega(d^{\frac{1}{4}}) )  (large d)</td>
<td>( n = \text{poly}(k) )</td>
<td>2 round EM with ( \Theta(k \cdot \log k) ) centers</td>
</tr>
<tr>
<td>Arora Kannan 2001</td>
<td>( s = \Omega(d^{\frac{1}{4}} \log d) )</td>
<td></td>
<td>Distance based</td>
</tr>
<tr>
<td>Vempala Wang 2004</td>
<td>( s = \Omega(k^{\frac{1}{4}} \log dk) )</td>
<td>( n = \Omega(d^3k^2 \log(dk/s\delta)) )</td>
<td>Spectral projection, then distances</td>
</tr>
</tbody>
</table>

General mixture of Gaussians:

- [Kannan Salmasian Vempala 2005] \( s = \Omega(k^{5/2} \log(kd)) \), \( n = \Omega(k^2d \cdot \log^5(d)) \)
- [Achlioptos McSherry 2005] \( s > 4k + o(k) \), \( n = \Omega(k^2d) \)
“Clustering isn’t hard—
it’s either easy,
or not interesting”
Effect of “Signal Strength”

Large separation, More samples

Lots of data—true solution creates distinct peak. Easy to find.

Computational limit

Just enough data—optimal solution is meaningful, but hard to find?

Informational limit

Not enough data—“optimal” solution is meaningless.

Small separation, Less samples
Effect of “Signal Strength”

Infinite data limit:
$E_x[\text{cost}(x;\text{model})] = \text{KL}(\text{true}||\text{model})$

Mode always at true model

Determined by
• number of clusters ($k$)
• dimensionality ($d$)
• separation ($s$)

Actual log-likelihood

Also depends on:
• sample size ($n$)

“local ML model” $\sim N(\text{true}; \frac{1}{n} J^{-1}_{\text{Fisher}})$

[Redner Walker 84]
Informational and Computational Limits

- What are the informational and computational limits?
- Is there a gap?
- Is there some minimum required separation for computational tractability?
- Is the learning the centers always easy given the true distribution?

Analytic, quantitative answers.
Independent of specific algorithm / estimator
Empirical Study

• Generate data from known mixture model
  – Uniform mixture of $k$ unit variance spherical Gaussians in $\mathbb{R}^d$
  – Distance $s$ between every pair of centers (centers at vertices of a simplex)

• Learn centers using EM
  – Spectral projection before EM
  – Start with $k \cdot \log k$ clusters and prune down to $k$

• Also run EM from true centers or true labeling
  (Cheating attempt to find ML solution)
EM with Different Bells and Whistles: Spectral Projection, Pruning Centers

Select $k \cdot \log k$ points at random as centers

PCA

EM

lift back

EM

prune

EM

prune

EM

EM

EM

EM

EM

EM

EM

EM

Select $k$ points at random as centers

PCA

EM

lift back

EM

EM

EM

candidate models
EM with Different Bells and Whistles: Spectral Projection, Pruning Centers

- **plain**
- **spectral projection**
- **prune down from \( k \log(k) \)**
- **spec proj: lift then prune**
- **spec proj: prune then lift**

**label error** = wrong edges in “same cluster” graphs
i.e., fraction of pairs of points that are in the same true cluster, but not the same recovered cluster, or visa versa
Behavior as a function of Sample Size

$k=16 \ d=1024 \ \text{sep}=6\sigma$

- △ “fair” EM
- × EM from true centers
- ○ Max likelihood (fair or not)
- • True centers

Difference between likelihood of “fair” EM runs and EM from true centers
  - • each run (random init)
  - △ run attaining max likelihood
Behavior as a function of Sample Size:
Lower dimension, less separation

$k=16$ $d=512$ $sep=4.0$

“fair” EM
EM from true centers
Max likelihood (fair or not)
True centers

Difference between likelihood of “fair” EM runs and EM from true centers
- each run (random init)
- run attaining max likelihood
Behavior as a function of Sample Size:
Lower dimension, less separation

$k=8$ $d=128$ $sep=3.0$

- △ “fair” EM
- × EM from true centers
- ○ Max likelihood (fair or not)
- • True centers

Difference between likelihood of “fair” EM runs and EM from true centers:
- • each run (random init)
- △ run attaining max likelihood
Behavior as a function of Sample Size:

Lower dimension, less separation

$k=8 \ d=128 \ sep=2.0$

- "fair" EM
- EM from true centers
- Max likelihood (fair or not)
- True centers

Difference between likelihood of "fair" EM runs and EM from true centers:
- each run (random init)
- run attaining max likelihood
Informational and Computational Limits as a function of $k$ and separation

\[ n \propto k^{1.5} - k^{1.6} \text{ for all } d, \text{ separation} \]
Informational and Computational Limits as a function of $d$ and separation

![Graph showing required sample size as a function of $d$ for different values of $s$. The graph includes lines for $s=4$, $s=6$, and $k=16$, with labels for computational and informational limits.]
Limitations of Empirical Study

• Specific optimization algorithm
  – Can only bound computational limit from above
• Do we actually find the optimum (max likelihood) solutions?
  – Can see regime in which EM fails even though there is a higher likelihood solution which does correspond to true model
  – But maybe there is an even higher likelihood solution the doesn’t?
• True centers always on a simplex
• Equal radius spherical Gaussians
Imperfect Learning

- So far, assumed data comes from specific model class (restricted Gaussian mixture)
- Even if data is not Gaussian, but clusters are sufficiently distinct and “blobby”, \( k \)-means / learning a Gaussian mixture model is easy.
- Can we give description of data for which this will be easy?

But for now, I’ll also be very happy with results on data coming from a Gaussian mixture…
Other Problems with Similar Behavior

• **Graph partitioning (correlation clustering)**
  – Hard in the worst case
  – Easy (using spectral methods) for large graphs with a “nice” statistically recoverable partition [McSherry 03]

• **Learning structure of dependency networks**
  – Hard to find optimal (max likelihood, or NML) structure in the worst case [S 04]
  – Polynomial-time algorithms for the large-sample limit [Narasimhan Bilmes 04]
Summary

• What are the informational and computational limits on Gaussian mixture clustering?
• Is there a gap?
• Is there some minimum required separation for computational tractability?
• Is the learning the centers always easy given the true distribution?
  • Analytic, quantitative answers
  • Hardness results independent of specific algorithm

• Limited empirical study:
  – There does seem to be a gap
  – Reconstruction via EM+spectral projection even from small separation (and a large number of samples)
  – Computational limit (very) roughly $\propto k^{1.5d}$