LAMBDA MEANS CLUSTERING
AUTOMATIC PARAMETER SEARCH AND DISTRIBUTED COMPUTING IMPLEMENTATION

MARCUS COMITER, MIRIAM CHA, HT KUNG, SURAT TEERAPITTAYANON
HARVARD UNIVERSITY
ICPR 2016
DECEMBER 6, 2016
TALK OUTLINE

• Motivation and Introduction
• Background
• Lambda Means
• Benefits of Lambda Means
• Results
• Extension to Distributed Framework
MACHINE LEARNING: VISION VS. REALITY
MACHINE LEARNING: VISION VS. REALITY

Vision
MACHINE LEARNING: VISION VS. REALITY

Vision

Reality
CLUSTERING

- Clustering is one of the most basic yet most powerful and fundamental of machine learning algorithms.

- But even in this simple setting, the choice of parameters are both difficult and greatly impact performance.
CLUSTERING

• Clustering is one of the most basic yet most powerful and fundamental of machine learning algorithms

• But even in this simple setting, the choice of parameters are both difficult and greatly impact performance
If machine learning is fundamentally a *data driven science*, shouldn't the use of machine learning itself follow a data driven methodology?
INTRODUCTION

• We present Lambda Means, a meta algorithm for the newly popular clustering algorithm DP-means

• Lambda Means automatically finds DP-means' main parameter ($\lambda$) automatically

• It finds $\lambda$ using the data itself on which the clustering is being performed
TALK OUTLINE

• Motivation and Introduction

• Background

• Lambda Means

• Benefits of Lambda Means

• Results

• Extension to Distributed Framework
DP-MEANS

• DP-means forms clusters of superior quality using a distance parameter $\lambda$ to ensure minimum separation between cluster centroids rather than specifying $k$ in advance

• B. Kulis and M. I. Jordan (the authors of DP-means) show that this new algorithm outperforms the traditional k-means algorithm!

• The algorithm forms a new cluster when a data point is found to be more than $\lambda$ distance away from all existing cluster centroids
Under an assumption that a sequence of data is drawn from a Dirichlet Process Mixture Model, B. Kulis and M. I. Jordan (the authors of DP-means) prove that there exists a lambda value such that when used by DP-means, the algorithm will discover the ground truth number of clusters $k$.

- $\mu_1, \ldots, \mu_k \sim G_0$
- $\pi \sim \text{Dir}(k, \pi_0)$
- $z_1, \ldots, z_n \sim \text{Discrete}(\pi)$
- $x_1, \ldots, x_n \sim \mathcal{N}(\mu_{z_i}, \sigma I)$

- $\mu$ corresponds to the mean of each of the clusters, drawn from some base distribution $G_0$, which is the prior distribution over the means
- $\pi=(\pi_1, \pi_2 \ldots)$ corresponds to the vector of probabilities of being in a cluster ($k \rightarrow \infty$)
- $z_i$ is an indicator of cluster assignment
- $x_i$ is a data point
In practice, without knowing the parameters of the distribution from which the data is drawn, it is unclear how to find the appropriate value of $\lambda$ for use with DP-means.

To solve this problem, a Farthest-first Heuristic requiring a user-provided approximation of $k$ can be used.

However, it is not easy to set $k$.

The choice of $k$ has a marked impact on the resulting value of $\lambda$. 
TALK OUTLINE

• Motivation and Introduction
• Background
• Lambda Means
• Benefits of Lambda Means
• Results
• Extension to Distributed Framework
LAMBDA MEANS

- As a solution for *automatically* finding the $\lambda$ parameter for use with DP-means, we present Lambda Means

- It finds $\lambda$ using the data itself on which the clustering is being performed

- Under an assumption that the data is generated by a Dirichlet Process Mixture Model, we formally prove that the $\lambda$ value found by Lambda Means is the same $\lambda$ used in generating the data (see Section III.D in our paper)
LAMBDAA MEANS

• The algorithm’s main mechanism is to decrease $\lambda$ at each iteration, automatically terminating at the proper $\lambda$ value.

• This has the effect of precipitating clusters at each iteration up to the point at which all clusters have been identified, but before the point at which true clusters are broken up into individual points.
ILLUSTRATION OF EFFECT OF DECREASING $\lambda$

**Iteration: T**  
**Lambda: Large**

A large value of lambda causes the two sets of points to be clustered together.

**Iteration: T + $\Delta T$**  
**Lambda: Small**

A small value of lambda causes the two sets of points to be clustered separately.
ILLUSTRATION OF EFFECT OF DECREASING $\lambda$

$D$, $d$: maximum and minimum distance between cluster centroids
$H$, $h$: maximum and minimum distance between data points
$N$: total # of data points

Clusters C1, C2A, C2B revealed
Clusters C1 and C2 revealed

Decreasing $\lambda$
LAMBDA MEANS

- Note that a naive implementation would generate the entire curve and then search for the elbow.

- Lambda Means replaces the need for this exhaustive search for the elbow of the curve.

- The algorithm uses the cumulative number of clusters formed as a signaling mechanism, continuing to iterate with smaller values of $\lambda$ until the stopping criteria is met.
TALK OUTLINE

• Motivation and Introduction

• Background

• Lambda Means

• **Benefits of Lambda Means**

• Results

• Extension to Distributed Framework
BENEFITS

• Lambda means is more robust then using a Farthest-first Heuristic, which requires a user-defined $k$

  • Reason 1: Setting this $k$ can be very difficult

  • Reason 2: If the initial approximation to $k$ is wrong, it negatively affects finding the correct $\lambda$
BENEFITS

- To show the effect of an incorrect $k$, we generate a dataset and then use the Farthest-first Heuristic with a number of different values of $k$ to derive $\lambda$.
- We find that $\lambda$ varies greatly based on the initial $k$ used.
BENEFITS

• The drawbacks of the farthest-first heuristic are clear:
  
  • The method is **brittle** to small changes in the approximation of $k$
  
  • The method has a **large impact** on the derived value of $\lambda$ as well as potentially on the **resulting cluster quality**
  
  • In contrast, Lambda Means automatically finds the $\lambda$ value without an initial approximation for $k$
TALK OUTLINE

• Motivation and Introduction
• Background
• Lambda Means
• Benefits of Lambda Means
• Results
• Extension to Distributed Framework
RESULTS

• We provide experimental evaluation of $\lambda$-means on both synthetic and real world data

• For synthetic data, we generate data with different values of inter-cluster variance $\rho$ and the intra-cluster variance $\sigma$

• For real-world data, we use the MNIST hand written digit dataset
RESULTS

• This figure shows that for synthetic data with a high value of $\rho/\sigma$, Lambda Means is able to automatically find the $\lambda$ value that maximizes AMI and NMI scores.

• NMI measures the amount of mutual information normalizing for number of clusters, and AMI measures the amount of mutual information accounting for chance.

• We can also judge Lambda Means by its ability to identify the correct number of clusters, which it does (as shown by the blue line).
RESULTS

• We now compare the AMI and NMI scores for Lambda Means and DP-means in Table I for additional values of $\rho/\sigma$, as well as for the MNIST dataset.

• Lambda Means outperforms DP-means where $\lambda$ is set via the Farthest-first heuristic.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Dataset</th>
<th>AMI</th>
<th>NMI</th>
<th>AMI</th>
<th>NMI</th>
<th>AMI</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Syn. $\frac{\rho}{\sigma} = 15$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda$-means</td>
<td></td>
<td>0.97</td>
<td>0.98</td>
<td>0.77</td>
<td>0.82</td>
<td>0.43</td>
<td>0.53</td>
</tr>
<tr>
<td>DP-means</td>
<td></td>
<td>0.87</td>
<td>0.92</td>
<td>0.52</td>
<td>0.78</td>
<td>0.32</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>Syn. $\frac{\rho}{\sigma} = 5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MNIST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TALK OUTLINE

• Motivation and Introduction
• Background
• Lambda Means
• Benefits of Lambda Means
• Results
• Extension to Distributed Framework
DISTRIBUTED RESULTS

- Lambda Means easily extends to the distributed framework under the optimistic concurrency control framework.

- We achieve within a factor of two away from a perfect speed-up in both the multicore and multi-processor distributed settings.

(a) Distributed cluster in the cloud
(b) Multicore
THANK YOU

MARCUS COMITER, MIRIAM CHA, HT KUNG, SURAT TEERAPITAYANON
HARVARD UNIVERSITY