Information Theoretic Clustering using Kernel Density Estimation

Shashank Singh [1] Bryan Hooi[1]

10-715 Advanced Introduction to Machine Learning

October 21, 2014

[1]Machine Learning Dept. and Dept. of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA

Background

- Between 2010-2012, several papers proposed an approach to nonparametric clustering based on maximizing the estimated mutual information between the data points and their labels (MIMax)
- Steeg et al., 2014, showed that MIMax was asymptotically biased towards clusters of equal sample size, and thus sometimes performed *worse* with more data

Background

- Intead, Steeg et al. used the axiomatic foundations of information theory to justify an approach based on minimizing the estimated conditional entropy $\hat{H}(Y|X)$ of the labels (Y) given the data (X)
- They proposed an algorithm using a k-nearest neighbor estimate $\hat{H}(Y|X)$

Main Contributions

Our work. . .

- provides further motivation for Conditional Entropy Minimization in terms of Minimum Description Length (MDL)
- suggests a principled approach to determining the number of clusters using MDL
- provides a theoretical link between clustering CHMin and the K-means algorithm
- provides a novel approach to Conditional Entropy clustering via Kernel Density Estimation (CHMin)
- empirically compares the performance of CHMin on synthetic and real datasets with K-means and Hierachical Clustering

Theoretical Results

Theoretical Results

Minimum Description Length (MDL)

- Principle of parsimony
- Select the hypothesis that compresses the data the most.

Two-stage MDL

$$\underset{H}{\text{minimize}} \quad L(H) + L(D|H)$$

Conditional Entropy Minimization and MDL

Theorem

Under the conditions:

- *Fixed number of clusters K*
- *Estimate $\hat{p}$ as a mixture of a parametric distribution (e.g. mixture of Gaussians)*

Minimizing description length is equivalent to minimizing estimated CE $\hat{H}(Y) + \hat{H}(X|Y)$.

Implications

- Justifies minimizing CE
- Can use MDL to select the number of clusters K

Selecting number of clusters using MDL

Theorem

To select the number of clusters K using MDL, we minimize

$$\hat{H}(Y) + \hat{H}(X|Y) + \log^*(K) + Kd(\log(2B) + \frac{1}{2}\log(n)) + \log(K!)$$

- Can be seen as $\hat{H}(Y) + \hat{H}(X|Y) +$ penalty on K
- Penalty grows as $O(($no. of parameters$) \times \log n)$
    - Same as BIC

Conditional Entropy and the K-Means Algorithm

Theorem

- *Using a Gaussian kernel function* $K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$*, the estimated conditional entropy* $\hat{H}(X|Y)$ *satisfies:*

$$\hat{H}(X|Y) \leq \log(h) + \frac{1}{2}\log(2\pi) + \frac{1}{2h^2 n} \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \mu_k)^2$$

- *Minimizing the K-means objective* $\sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \mu_k)^2$ *is equivalent to minimizing an upper bound for* $\hat{H}(X|Y)$*.*

- Use K-means to initialize gradient descent for conditional entropy (CE) minimization

Empirical Results

Empirical Results

Intuition

Why do we want to minimize

$$\frac{\hat{H}(Y|X)}{\hat{H}(Y)}?$$

- Points with similar x-values and different y values increase $\hat{H}(Y|X)$
- Having a small range of y values decreases $\hat{H}(Y)$

$\Rightarrow$ minimizing the objective causes nearby x-values to have similar y-values

CHMin: A Simple Optimization Procedure

Want to solve:

$$\min_{y_1,\cdots,y_n \in \{0,1\}} \frac{\hat{H}(Y|X)}{\hat{H}(Y)}.$$

We use gradient descent + rescaling into $[0,1]$; i.e., repeatedly:

1. $$y \leftarrow y - \alpha \nabla_y \frac{\hat{H}(Y|X)}{\hat{H}(Y)}$$

2. $$y \leftarrow \frac{y - \min_i y_i}{\max_i y_i - \min_i y_i}$$

For $K > 2$ clusters, use soft clustering: rescale onto convex hull of $(0,0,\cdots,0,1),(0,0,\cdots,1,0),\cdots,(1,0,\cdots,0,0)$.

CHMin: Parameter Selection

KDE Bandwidth: Literature suggests undersmoothing (relative to optimal density derivative estimate). In practice, Silverman's Rule of Thumb seems to work better than AMISE.

KDE Kernel: We use a Gaussian kernel, but, for well-separated clusters, bounded kernels (e.g., Epanechnikov, Uniform) work very well (converge quickly).

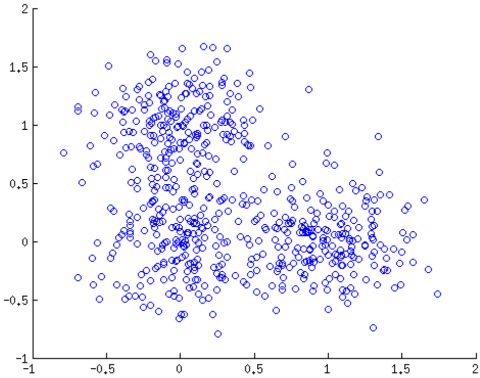
Gradient Step Size: Anything approaching 0 slowly appears to work ($1/\log i, 1/\sqrt{i}$, etc.); affects convergence, but not final result

Initialization: K-means + random restarts (1-2 seems sufficient)

Three Gaussians

3 spherical Gaussians in $\mathbb{R}^3$

- Very easy data set



Three Gaussians

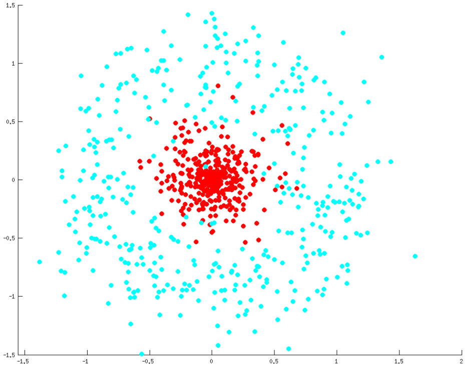
Three clusters (chance $= 0.33$)

CHMin	K-means++	HC (complete)	HC (average)
0.991	**0.998**	0.984	0.994

Concentric Circles: Results

Two concentric circles in $\mathbb{R}^2$

- Not linearly-separable
- 2/3 of data points in inner cluster - MIMax doesn't work well.



Concentric Circles: Results

Two clusters (chance $= 0.5$)

CHMin	K-means++	HC (complete)	HC (average)
0.894	0.671	0.677	0.605

Iris

Cluster 3 iris species using 4 flower measurements (150 samples)

- One fairly distinct, linearly separable cluster.
- Two overlapping clusters.
- Chance $= 0.33$.

CHMin	K-means++	HC (complete)	HC (average)
0.929	0.893	0.840	0.906

Wine

Cluster 3 wine source using 13 chemical properties (178 samples)

- One cluster is fairly distinct and linearly separable. Remaining two overlap.
- Chance = 0.33

CHMin	K-means++	HC (complete)	HC (average)
0.675	**0.702**	0.674	0.612

- Difficulty in high-dimensional nonparametric density estimate
- Improved performance on (arbitrary) 5 feature subset:

CHMin	K-means++	HC (complete)	HC (average)
0.700	0.494	0.500	0.500

Empirical Conclusions

- CHMin works well on a number of (relatively small) datasets
- Scales poorly with dimension
  - Only depends on pairwise distances, so could combine with dimension reduction

Future Work

- Empirically, how does CHMin fare against other nonparametric clustering approaches (e.g., MIMax, mean shift)
- Empirically, how well does MDL identify number of clusters?
- Can other optimization procedures speed up convergence?
- Can we adapt error bounds from kernel density estimation?