Demystifying Information-Theoretic Clustering

Greg Ver Steeg, Aram Galstyan, Fei Sha and Simon DeDeo

(ICML 2014)

Discussion by: Chunyuan Li

August 7, 2015

Outline

- Information theory in clustering
- Proposed method: Consistency Violation Ratio
- Experimental results

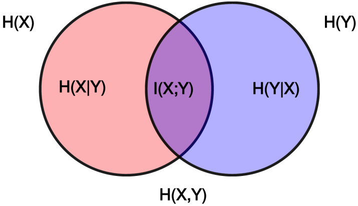
Information-theoretic clustering

- **Information-theoretic (IT) criteria**
  - Given samples drawn i.i.d. from a known distribution
  - Shannon entropy: minimum number of bits needed to encode the samples
- **Clustering**
  - Given samples of an unknown distribution,
  - We would like to label (encode), each sample to reflect some natural structure
- This paper: **Compression $\neq$ Clustering**

  Even if we knew the Shannon entropy of the distribution, a code that achieves this optimal compression does not necessarily reflect the natural structure of the underlying distribution.

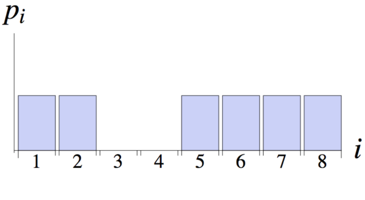
Information theory

- Reminder of basic IT concepts
  - Entropy: $H(\mathbf{X}) = \mathbb{E}[\log \frac{1}{p(\boldsymbol{x})}]$
  - Mutual information (MI):
    $I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y})$
- This paper: **Conditional Entropy, not Mutual information (MI)**
  MI determines how similar the joint distribution $p(\mathbf{X}, \mathbf{Y})$ is to the products of factored marginal distribution $p(\mathbf{X})p(\mathbf{Y})$

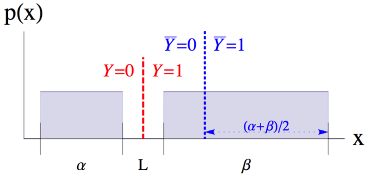


Pitfalls of IT clustering

- Discrete case
  Re-ordering of bins does not affect any IT quantity because they depend only on the values p_i.
- Continuous case
  $I(\mathbf{X};\mathbf{Y}) = H_0(\alpha/(\alpha+\beta))$, it reaches its maximum when splting the space into two equally sized masses of probability



(a)



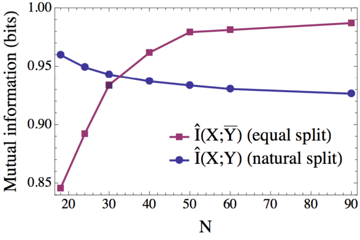
(b)

The knowledge to see Mystery

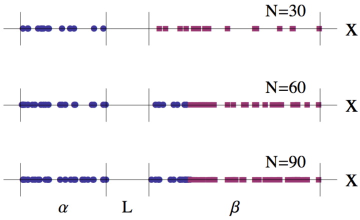
- What does maximizing MI tell us? $I(\mathbf{X};\mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X})$
  - $H(\mathbf{Y})$ is maximized for equally sized clusters.
  - $H(\mathbf{Y}|\mathbf{X})$ should be 0 for any exact partitioning of the input space
  - These two terms compete
- In pratice, entropy is estimated according to density. The non-parametric estimator is based on k-nn.
- **Uncertainty comes from the case near the boundary between the two clusters**
- **The percentage of points near the boundary will decrease as N increases**

Mystery about the empirical success

- When N is small
  - Natural clustering is preferred
- When N is large
  - Equal-sized clusters will be preferred
- More data leads to a less desirable result
- Tests with previous information-theoretic clustering objectives focused on small, nearly balanced datasets, so that these shortcomings went unnoticed.



(a)



(b)

Coarse-Graining

- Clustering is a coarse-graining
- Shannon's Axiom of Consistency under Coarse-Graining: Uncertainty should grow with k if there are k equally likely events, *e.g.*
  - Discrete case
    $h(p_1, p_2, p_3) = h(p_1, p_2 + p_3) + (p_2 + p_3)h(p_2/(p_2 + p_3), p_3/(p_2 + p_3))$
  - Continuous case
    $H(p(\boldsymbol{x})) = H(p(y)) + p(y = 0)H(p(\boldsymbol{x}|y = 0)) + p(y = 1)H(p(\boldsymbol{x}|y = 1))$
- For $y = f(\boldsymbol{x})$, it implies

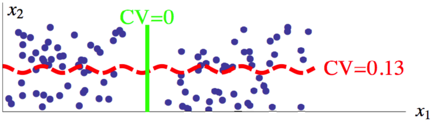
$$H(X) = H(X|Y) + H(Y)$$

- Start with an unbiased estimator and search for coarse-grainings that lead to consistent entropy estimates. These coarse-grainings are called as *natural*.

Consistency Violation (CV)

$$CV = \hat{H}(Y) + \hat{H}(X|Y) - \hat{H}(X) \quad (1)$$

- CV as a measure of how well we can estimate the global entropy given the entropy of clusters of data points
- RHS also recovers $\hat{H}(Y|X) = \hat{H}(Y) + \hat{H}(X|Y) - \hat{H}(X)$
- Advantages
  - The estimated uncertainty about cluster labels will be as low as possible, even for small amounts of data
  - The coarse-graining will be natural in the sense that we do not violate information-theoretic axioms



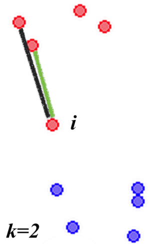
Entropy estimation

- Nonparametric estimator
  - $H(\mathbf{X}) = \mathbb{E}[\log \frac{1}{p(\boldsymbol{x})}] \approx \frac{1}{N}\sum_{i=1}^{N}[\log \frac{1}{p(\boldsymbol{x}^{(i)})}] \approx \frac{1}{N}\sum_{i=1}^{N}[\log \frac{\epsilon_{i,k}^{d}}{k/N}]$
  - $\hat{H}(\mathbf{X}) \approx \log \frac{N}{k} + \frac{d}{N}\sum_{i=1}^{N}[\log \epsilon_{i,k}] + c_{k,N}$
  - $c_{k,N}$: a constant factor
- Nonparametric estimation of conditional entropy

$$\hat{H}(Y|X) = \frac{d}{N}\sum_{i=1}^{N} \log \frac{\bar{\epsilon}_{i,k}}{\epsilon_{i,k}} \tag{2}$$

- Example:

green line: $\epsilon_{i,k}$
dark line: $\bar{\epsilon}_{i,k}$



Conditional entropy for clustering

- CV is expected to be small under arbitrary resampling with limited data
- By considering all possible resamplings with α,

$$\hat{H}_{\alpha,k}(Y|X) = \mathbb{E}_\alpha[\hat{H}_k(Y|X)] \tag{3}$$

  where α is an independent probability for each point to be removed.
- Total consistency violation

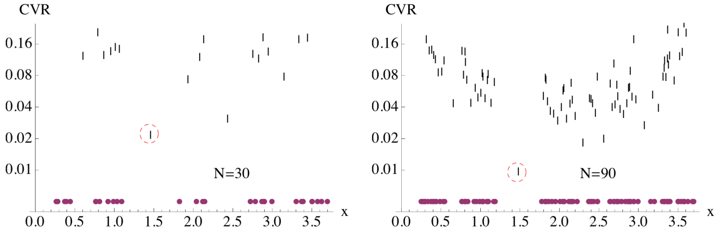
$$\hat{H}_{\alpha,k}(Y|X) = \int_0^1 d\alpha \hat{H}_{\alpha,k}(Y|X) \tag{4}$$

- Define $\hat{H}_T(Y|X) \equiv \hat{H}_{\alpha,k=1}(Y|X)$ We can search for partitions that minimize the Consistency Violation Ratio (CVR)

$$\hat{H}_T(Y|X)/\hat{H}(Y) \tag{5}$$

Experiment 1: Test CVR with N

- Setup
  - 2 unform distributions in 1D (slide 5, (b))
  - (a) $N = 30$, (b) $N = 90$.
- Results
  - Adding more data does not make the clusters harder to distinguish.



(a) (b)

Experiment 2: Compare CVR with others

- Setup
  - 2 unform distributions in 2D
  - Other ojective functions, including MI and Nonparametric Information Clustering (NIC) (Faivishevsky et al. 2010)
  - $r^* = \arg opt_r$ Objective$(\mathbf{Y}_r, \mathbf{X})$
- Results
  - CVR is the only objective to prefer the correct partition over a wide range of parameter values