A PAC-Bayesian Approach to Formulation of Clustering Objectives

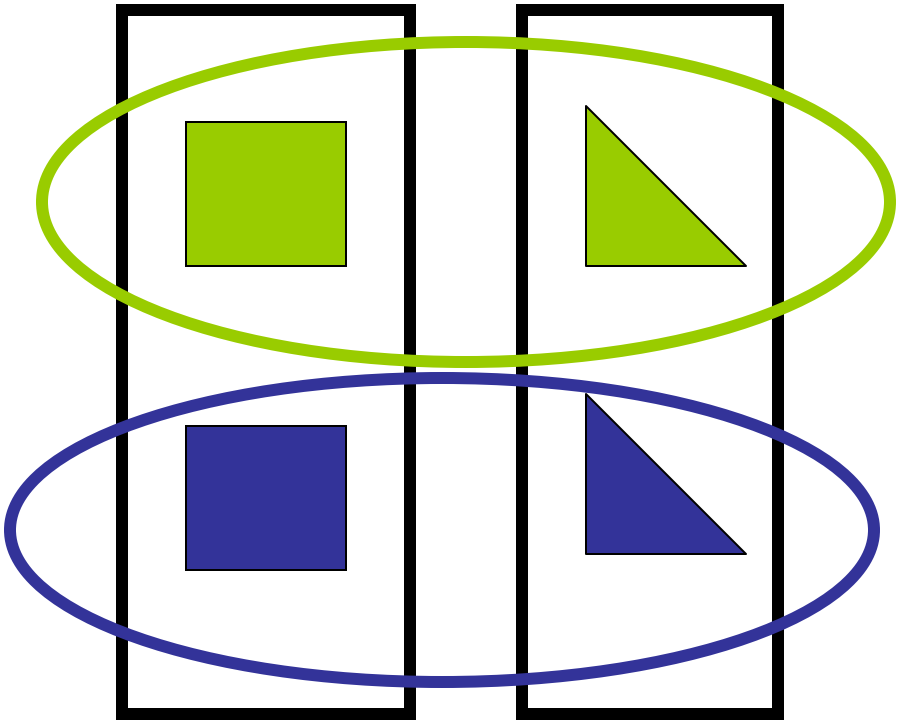
Yevgeny Seldin

Joint work with Naftali Tishby

Motivation

- Clustering tasks are often ambiguous
- It is hard to compare between solutions, especially if based on different objectives

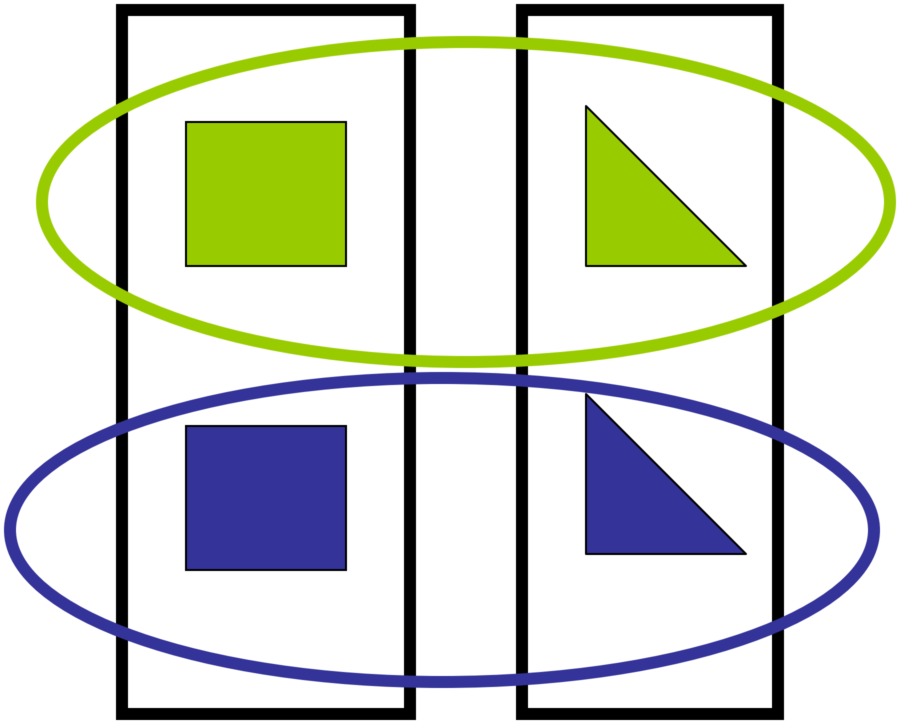
Example



Motivation

- Many structures co-exist simultaneously
- The problem of comparison of solutions cannot be resolved by testing any property of the clustering itself

Example



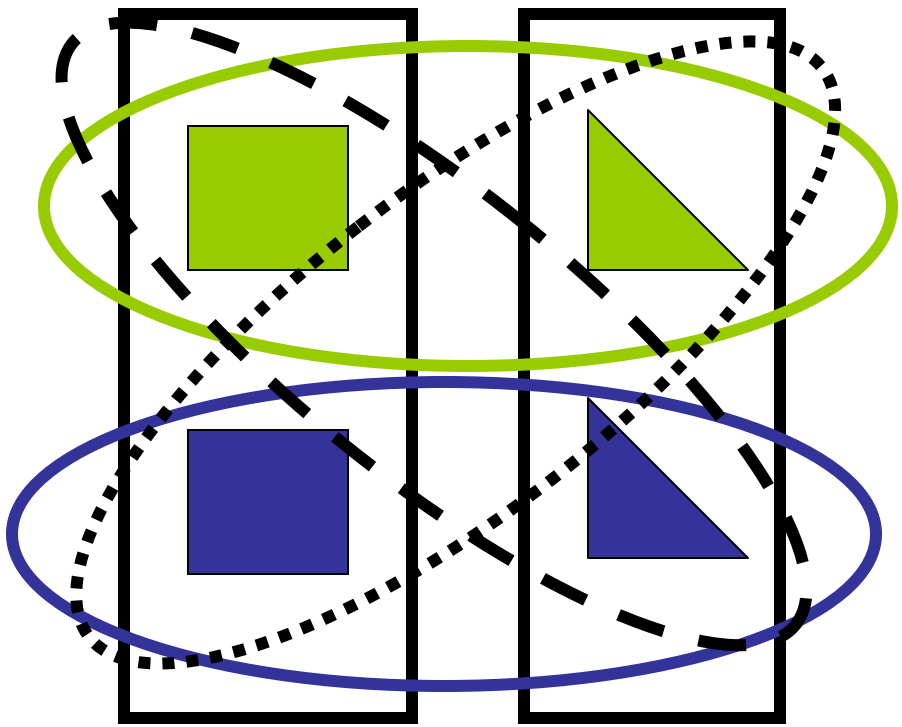
Motivation

- Clustering depends on our needs

Example

Recyclable

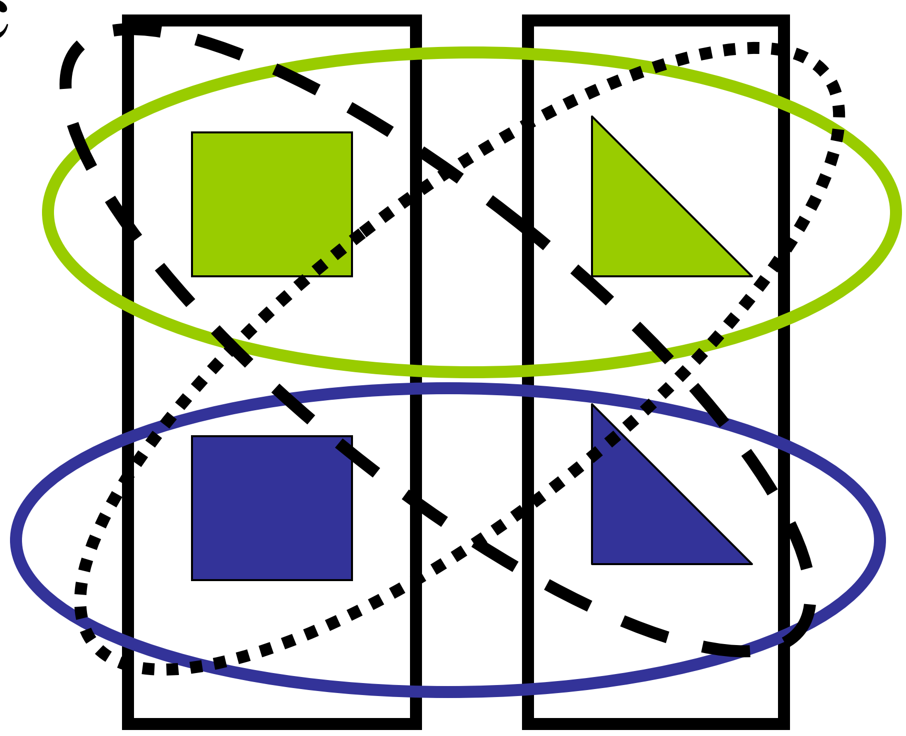
Non-Recyclable



Motivation

- Inability to compare solutions is problematic for advancement and improvement

Example



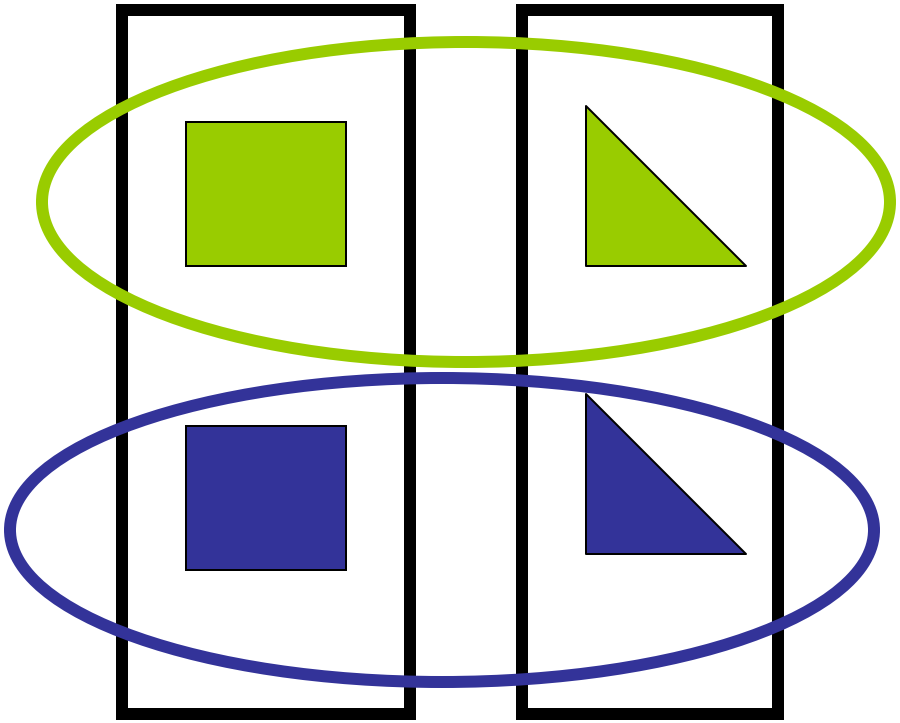
Thesis

- We do not cluster the data just for the sake of clustering, but rather to facilitate a solution of some higher level task
- The quality of clustering should be evaluated by its contribution to the solution of that task
- We should put more effort into identification and formulation of the tasks which are solved via clustering

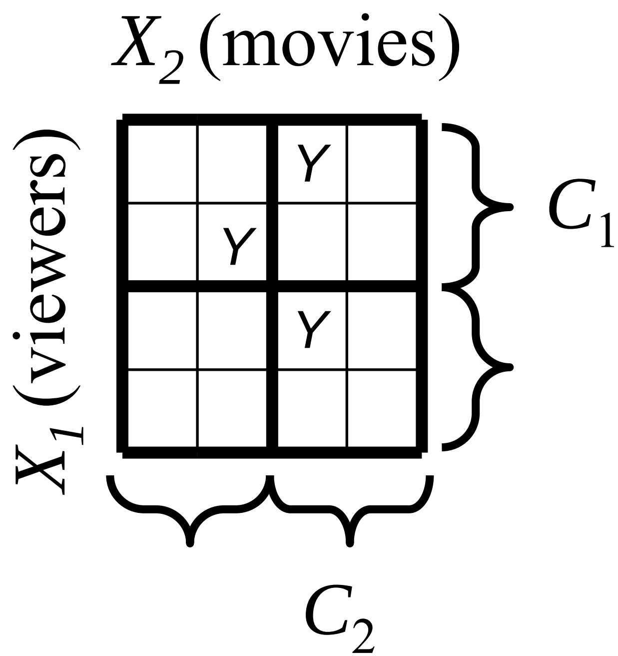
Example

- Cluster then pack
- Clustering by shape is preferable

Evaluate the amount of time saved

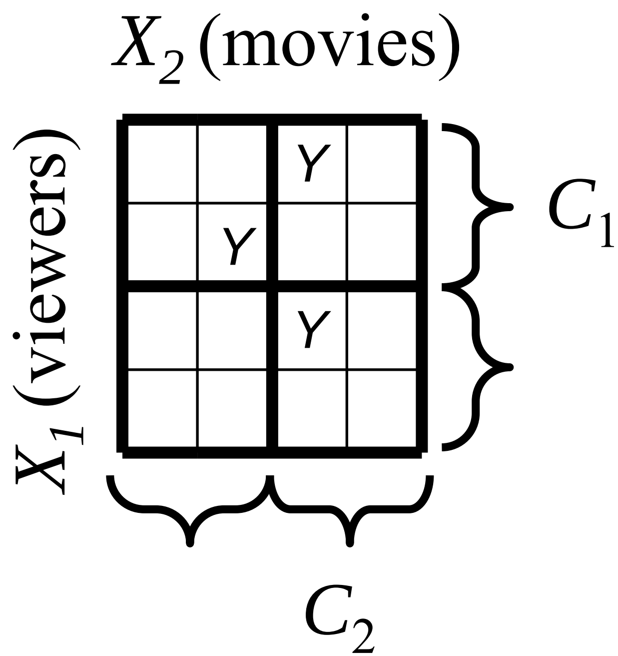


Proof of Concept
Collaborative Filtering via Co-clustering



Model: $q(Y \mid X_1, X_2) = \sum_{C_1, C_2} q(Y \mid C_1, C_2) q(C_1 \mid X_1) q(C_2 \mid X_2)$

Evaluation



- How well are we going to predict the new ratings

Analysis

$$L(q) = E_{p(X_1,X_2,Y)} E_{q(Y'|X_1,X_2)} l(Y,Y')$$

Expectation w.r.t. the true distribution $p(X_1,X_2,Y)$ (unrestricted)

Expectation w.r.t. the classifier $q(Y|X_1,X_2)$

Given loss $l(Y,Y')$

- Model-independent comparison
  - Does not depend on the form of q
- We can compare any two co-clusterings
- We can compare clustering-based solution to any other solution (e.g. Matrix Factorization)

PAC-Bayesian Analysis of Co-clustering

$$q(Y \mid X_1, X_2) = \sum_{C_1, C_2} q(Y \mid C_1, C_2) q(C_1 \mid X_1) q(C_2 \mid X_2)$$

$$L(q) = E_{p(X_1, X_2, Y)} E_{q(Y' | X_1, X_2)} l(Y, Y')$$

PAC-Bayesian Analysis of Co-clustering

$$q(Y \mid X_1, X_2) = \sum_{C_1, C_2} q(Y \mid C_1, C_2) q(C_1 \mid X_1) q(C_2 \mid X_2)$$

$$L(q) = E_{p(X_1, X_2, Y)} E_{q(Y' \mid X_1, X_2)} l(Y, Y')$$

$$\hat{L}(q) = E_{\hat{p}(X_1, X_2, Y)} E_{q(Y' \mid X_1, X_2)} l(Y, Y')$$

PAC-Bayesian Analysis of Co-clustering

$$q(Y \mid X_1, X_2) = \sum_{C_1, C_2} q(Y \mid C_1, C_2) q(C_1 \mid X_1) q(C_2 \mid X_2)$$

$$L(q) = E_{p(X_1, X_2, Y)} E_{q(Y' \mid X_1, X_2)} l(Y, Y') \qquad \hat{L}(q) = E_{\hat{p}(X_1, X_2, Y)} E_{q(Y' \mid X_1, X_2)} l(Y, Y')$$

$$L(q) \leq \hat{L}(q) + \sqrt{\frac{\sum_i |X_i| I(X_i; C_i) + K}{2N}}$$

PAC-Bayesian Analysis of Co-clustering

$$q(Y \mid X_1, X_2) = \sum_{C_1, C_2} q(Y \mid C_1, C_2) q(C_1 \mid X_1) q(C_2 \mid X_2)$$

$$L(q) = E_{p(X_1, X_2, Y)} E_{q(Y' \mid X_1, X_2)} l(Y, Y') \qquad \hat{L}(q) = E_{\hat{p}(X_1, X_2, Y)} E_{q(Y' \mid X_1, X_2)} l(Y, Y')$$

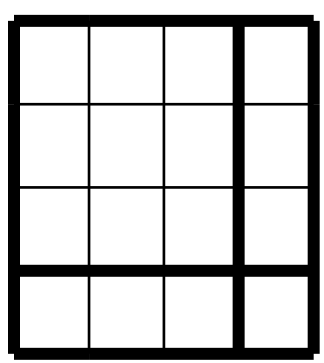
$$L(q) \leq \hat{L}(q) + \sqrt{\frac{\sum_i |X_i| I(X_i; C_i) + K}{2N}}$$

- We can compare any two co-clusterings
- We can find a locally optimal co-clustering
- We can compare clustering-based solution to any other solution (e.g. Matrix Factorization)

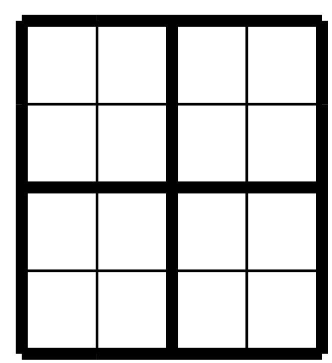
Bound Meaning

$$L(q) \le \hat{L}(q) + \sqrt{\frac{\sum_i |X_i| I(X_i; C_i) + K}{2N}}$$

- Trade-off between empirical performance and effective complexity



4 unbalanced partitions



$\binom{4}{2} = 6$ balanced partitions

Application

- Replace with a trade-off

$$L(q) \leq N\hat{L}(q) + \beta \sum_i |X_i| I(X_i; C_i)$$

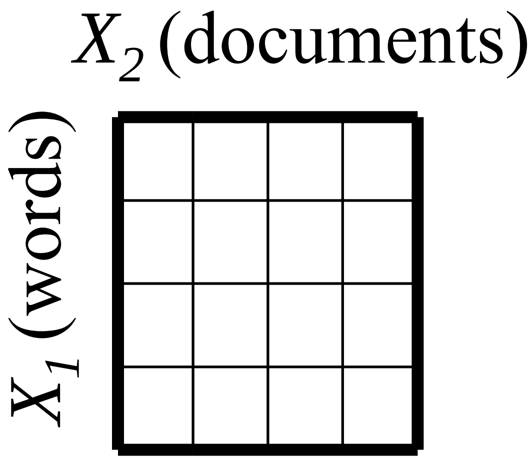
Application

- Replace with a trade-off

$$L(q) \leq N\hat{L}(q) + \beta \sum_i |X_i| I(X_i; C_i)$$

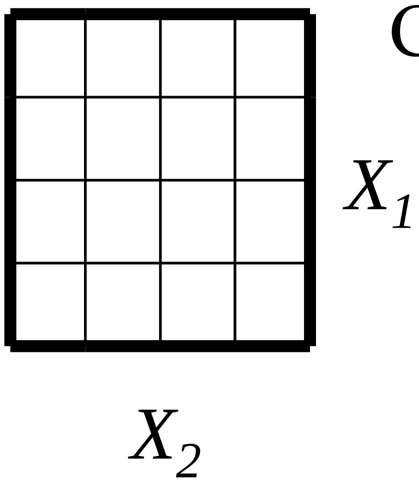
- MovieLens dataset
  - 100,000 ratings on 5-star scale
  - 80,000 train ratings, 20,000 test ratings
  - 943 viewers x 1682 movies
  - State-of-the-art Mean Absolute Error (0.72)
  - The optimal performance is achieved even with 300x300 cluster space

Co-occurrence Data Analysis



- Approached by
  - Co-clustering [Slonim&Tishby’01,Dhillon et.al.’03,...]
  - Probabilistic Latent Semantic Analysis [Hofmann’99,…]
  - …
- No theoretical comparison of the approaches
- No model order selection criteria

Suggested Approach

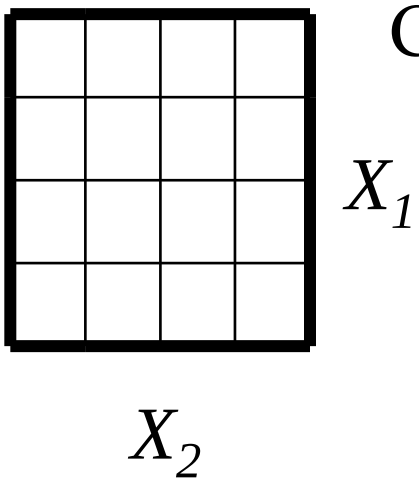


Co-occurrence events are generated by $p(X_1,X_2)$

$q(X_1,X_2)$ – a density estimator

Evaluate the ability of q to predict new co-occurrences (out-of-sample performance of q)

Suggested Approach



Co-occurrence events are generated by $p(X_1,X_2)$

$q(X_1,X_2)$ – a density estimator

Evaluate the ability of q to predict new co-occurrences (out-of-sample performance of q)

$$L(q) = -E_{p(X_1,X_2)} \ln q(X_1, X_2)$$

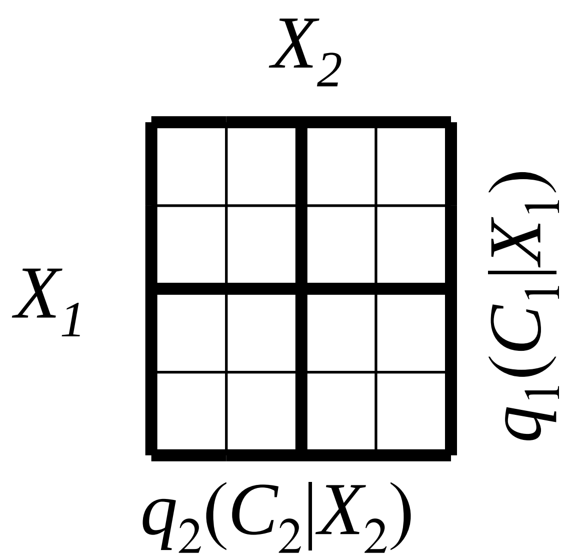
The true distribution $p(X_1,X_2)$ (unrestricted)

- Possibility of comparison of approaches
- Model order selection

Density Estimation with Co-clustering

- Model: $q(X_1, X_2) = \sum_{C_1, C_2} q(C_1, C_2) q(X_1 | C_1) q(X_2 | C_2)$
- With probability ≥ 1-δ:

$$-E_{p(X_1,X_2)} \ln q(X_1, X_2) \le -I(C_1;C_2) + \ln(|C_1||C_2|)\sqrt{\frac{\sum_i |X_i| I(X_i;C_i) + K_1}{2N}} + K_2$$



Density Estimation with Co-clustering

- Model: $q(X_1,X_2)=\sum_{C_1,C_2} q(C_1,C_2)q(X_1 \mid C_1)q(X_2 \mid C_2)$
- With probability ≥ 1-δ:

$$-E_{p(X_1,X_2)} \ln q(X_1,X_2) \le -I(C_1;C_2) + \ln(|C_1||C_2|)\sqrt{\frac{\sum_i |X_i| I(X_i;C_i) + K_1}{2N}} + K_2$$

- Related work
  - Information-Theoretic Co-clustering [Dhillon et. al. '03]: maximize $I(C_1;C_2)$ alone
  - PAC-Bayesian approach provides regularization and model order selection

Future work

- Formal analysis of clustering
  - Points are generated by $p(X)$, $X \in \mathrm{R}^{d}$
  - $q(X)$ is an estimator of $p(X)$
    - E.g. Mixture of Gaussians: $q(X) = \sum_i \lambda_i \mathrm{N}(\mu_i,\sigma_i)$
  - Evaluate $-\mathrm{E}_{p(X)} \ln q(X)$
  - Model order selection
  - Comparison of different approaches

Relation to Other Approaches to Regularization and Model Order Selection in Clustering

- Information Bottleneck (IB)
  - [Tishby, Pereira & Bialek ’99, Slonim, Friedman & Tishby ’06, …]
- Minimum Description Length (MDL) principle
  - [Barron,Rissanen&Yu’98, Grünwald ’07, …]
- Stability
  - [Lange, Roth, Braun & Buhmann ’04, Shamir & Tishby ’08, Ben-David & Luxburg ’08, …]

Relation with IB

= The “relevance variable” *Y* was a prototype of a “high level task”

≠ IB does not analyze generalization directly

- Although there is a post-factum analysis [Shamir,Sabato&Tishby ’08]
- There is a slight difference in the resulting tradeoff

≠ IB returns the complete curve of the trade-off between compression level and quality of prediction (no model order selection)

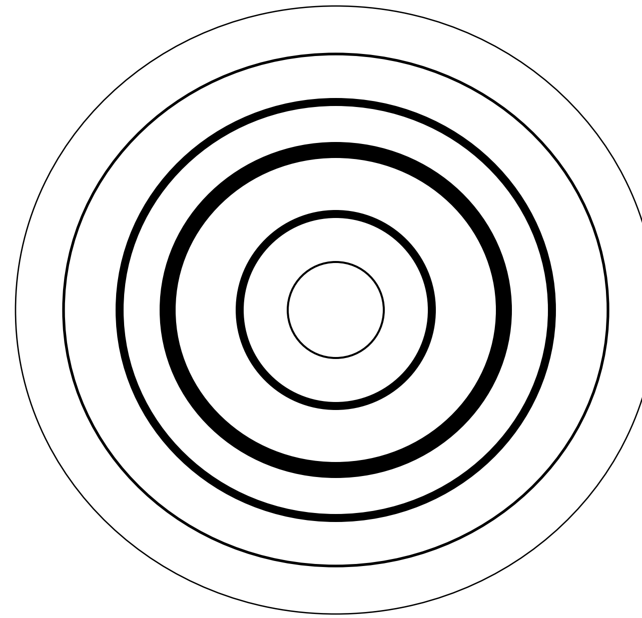
≠ PAC-Bayesian approach suggests a point which provides optimal prediction at a given sample size

Generalization ≠ MDL

- MDL is not concerned with generalization
- MDL solutions can overfit the data
  - [Kearns,Mansour,Ng,Ron ’97], [Seldin ’09]

Generalization ≠ Stability

- Example: "Gaussian ring"
  - Mixture of Gaussians estimation is not stable
  - If we increase the size of the sample and the number of Gaussians to infinity it will converge to the true distribution
- "Meaning" of the clusters is different



Some high level remarks

(For future work)

Clustering and Humans

- Clustering represents a structure of the world
- By clustering objects we ignore irrelevant properties of the objects and concentrate on the relevant ones (relevance is application dependent)
- We communicate by using a structured description of the world
- Clustering is tightly related to object naming
- There must be advantages to such a representation

What Kind of Tasks Clustering is Required for?

- Classification - ???
- Memory efficiency
- Computational efficiency
- Communication efficiency
- Multi-task learning and Transfer learning
- Transfer learning and Control
- Robustness
- Your ideas…

Summary

- In order to deliver better clustering algorithms and understand their outcome we have to identify and formalize their potential applications
- Clustering algorithms should be evaluated by their contribution in the context of their potential application