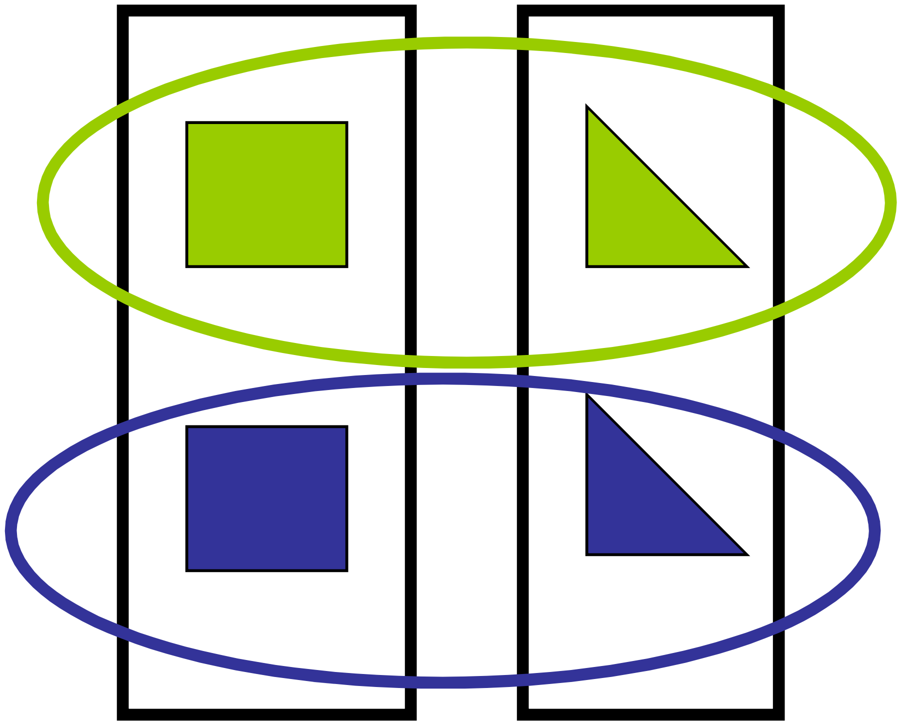
PAC-Bayesian Analysis of Co-clustering, Graph Clustering and Pairwise Clustering

Yevgeny Seldin

Motivation

- Clustering cannot be analyzed without specifying what it will be used for!

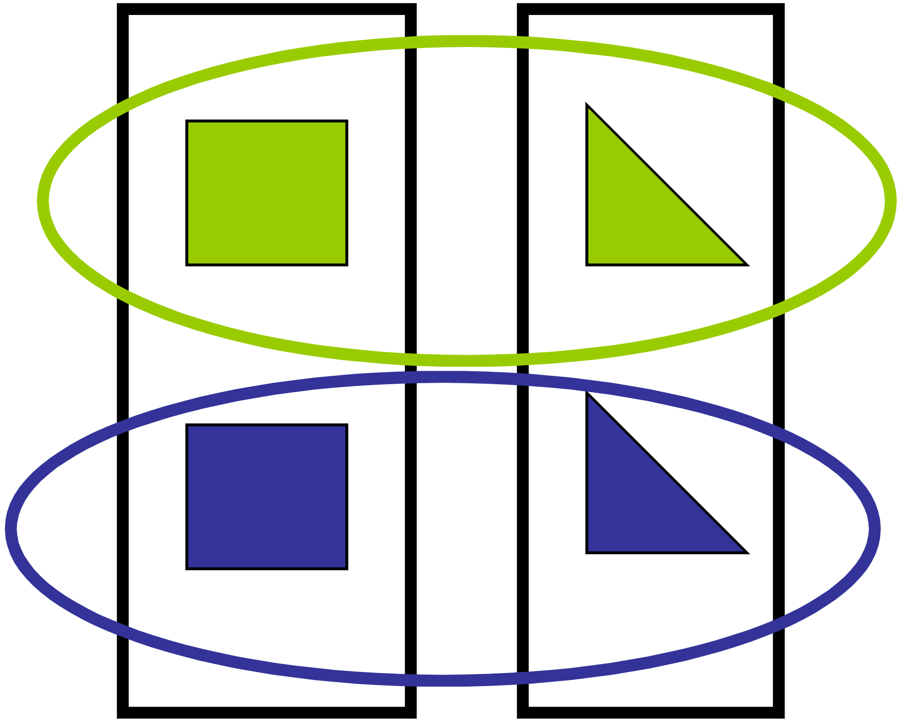
Example



Example

- Cluster then pack
- Clustering by shape is preferable

Evaluate the amount of time saved



How to define a clustering problem?

- Common pitfall: the goal is defined in terms of the solution
  - Graph cut
  - Spectral clustering
  - Information-theoretic approaches
- Which one to choose??? How to compare?

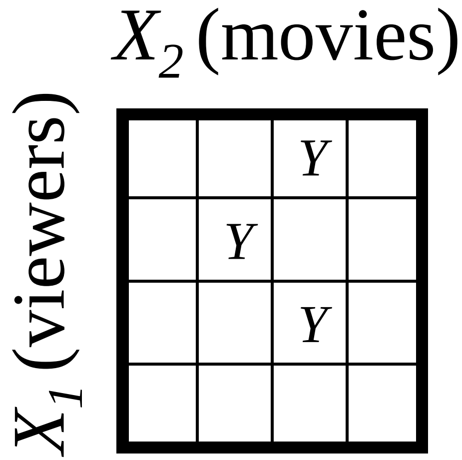
- Our goal: suggest problem formulation which is independent of the way of solution

Outline

- Two problems behind co-clustering
  - Discriminative prediction
  - Density estimation
- PAC-Bayesian analysis of discriminative prediction with co-clustering
- PAC-Bayesian analysis of graph clustering

Discriminative Prediction with Co-clustering

- Example: collaborative filtering
- Goal: find discriminative prediction rule $q(Y|X_1,X_2)$



Discriminative Prediction with Co-clustering

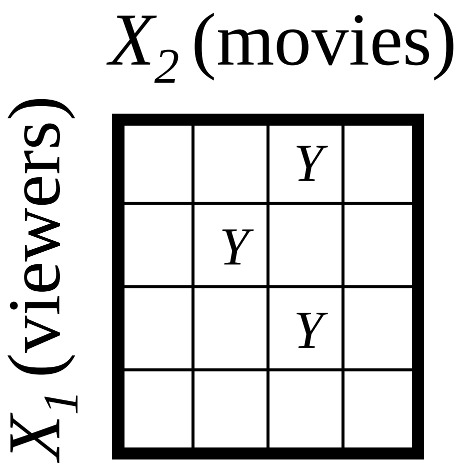
- Example: collaborative filtering
- Goal: find discriminative prediction rule $q(Y|X_1,X_2)$
- Evaluation:

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

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

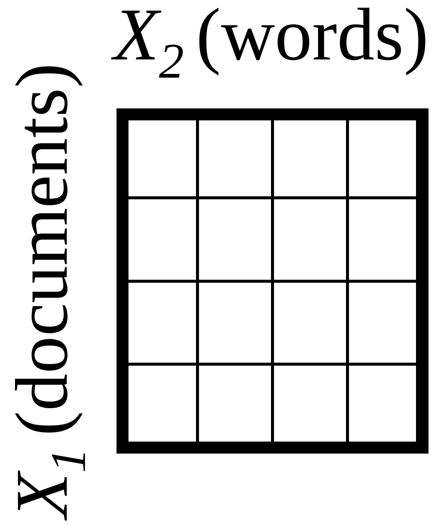
Expectation w.r.t. the classifier $q(Y|X_1,X_2)$

Given loss $l(Y,Y')$



Co-occurrence Data Analysis

- Example: words-documents co-occurrence data
- Goal: find an estimator $q(X_1,X_2)$ for the joint distribution $p(X_1,X_2)$

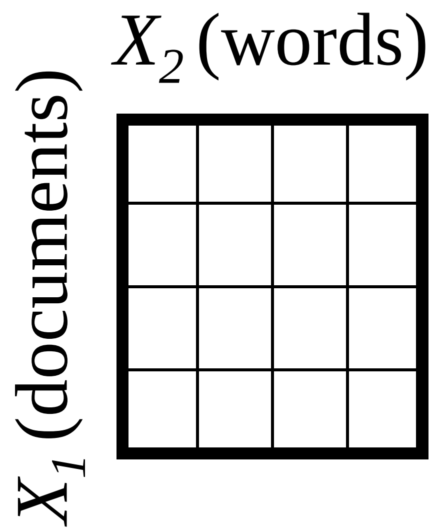


Co-occurrence Data Analysis

- Example: words-documents co-occurrence data
- Goal: find an estimator $q(X_1,X_2)$ for the joint distribution $p(X_1,X_2)$
- Evaluation:

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

The true distribution $p(X_1,X_2)$

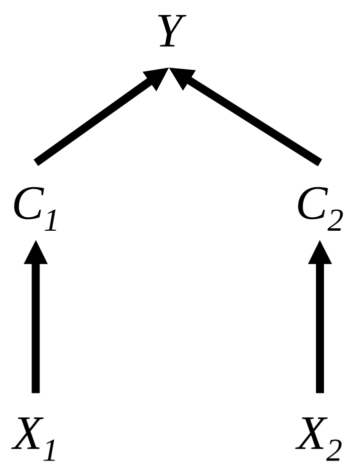


Outline

- PAC-Bayesian analysis of discriminative prediction with co-clustering

Discriminative prediction based on co-clustering

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



Denote:

$Q = \{q(C_1|X_1), q(C_2|X_2), q(Y|C_1,C_2)\}$

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

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

Generalization Bound

- With probability $\geq 1\text{-}\delta$:

$$kl(\hat{L}(Q) \,\|\, L(Q)) \leq \frac{\sum_i |X_i| I(X_i; C_i) + K}{N}$$

$$kl(\hat{L}(Q) \,\|\, L(Q)) = \hat{L}(Q) \ln \frac{\hat{L}(Q)}{L(Q)} + (1 - \hat{L}(Q)) \ln \frac{\hat{L}(Q)}{L(Q)}$$

- A looser, but simpler form of the bound:

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

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

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

Generalization Bound

- With probability $\geq 1\text{-}\delta$:

$$kl(\hat{L}(Q) \,\|\, L(Q)) \leq \frac{\sum_i |X_i| I(X_i; C_i) + K}{N}$$

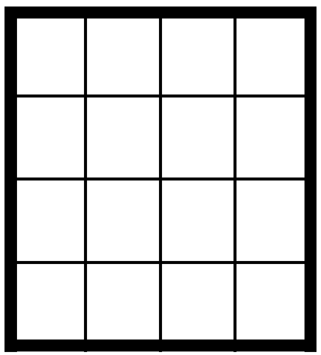
$$K = \underbrace{\sum_i |C_i| \ln |X_i|}_{\text{Logarithmic in } |X_i|} + \underbrace{\left(\prod_i |C_i|\right)}_{\text{Number of partition cells}} \ln |Y| + \underbrace{\ln(4N)/2 - \ln \delta}_{\text{PAC-Bayesian bound part}}$$

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

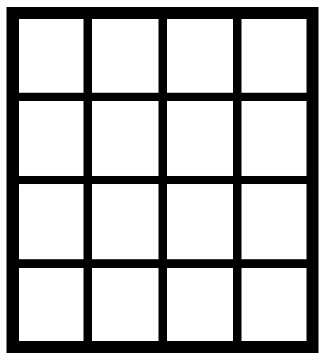
Generalization Bound

- With probability $\geq 1-\delta$:

$$kl(\hat{L}(Q) \| L(Q)) \leq \frac{\sum_i |X_i| I(X_i; C_i) + K}{N}$$



Low Complexity
$I(X_i;C_i) = 0$



High Complexity
$I(X_i;C_i) = \ln|X_i|$

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

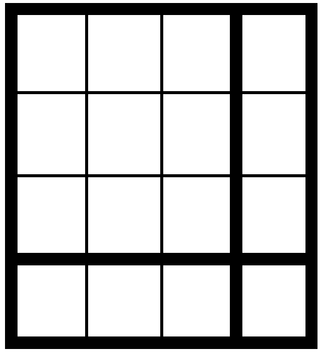
$Q = \{q(Y \mid C_1, C_2), q(C_1 \mid X_1), q(C_2 \mid X_2)\}$

Generalization Bound

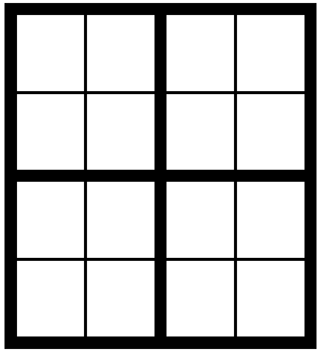
- With probability $\geq 1-\delta$:

$$kl(\hat{L}(Q) \,\|\, L(Q)) \leq \frac{\sum_i |X_i| I(X_i; C_i) + K}{N}$$

<u>Optimization tradeoff:</u>
Empirical loss vs.
“Effective” partition
complexity



Lower
Complexity



Higher
Complexity

Practice

- With probability $\geq$ 1-δ:

$$kl(\hat{L}(Q) \| L(Q)) \leq \frac{\sum_i |X_i| I(X_i;C_i) + K}{N}$$

- Replace with a trade-off:

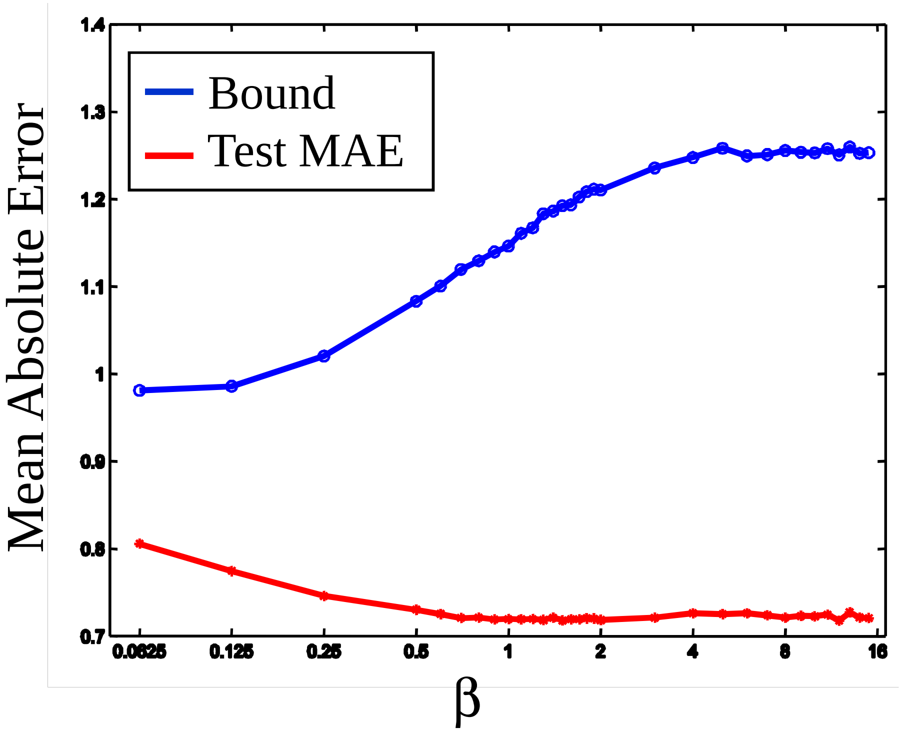
$$F(Q) = \beta N \hat{L}(Q) + \sum_i |X_i| I(X_i;C_i)$$

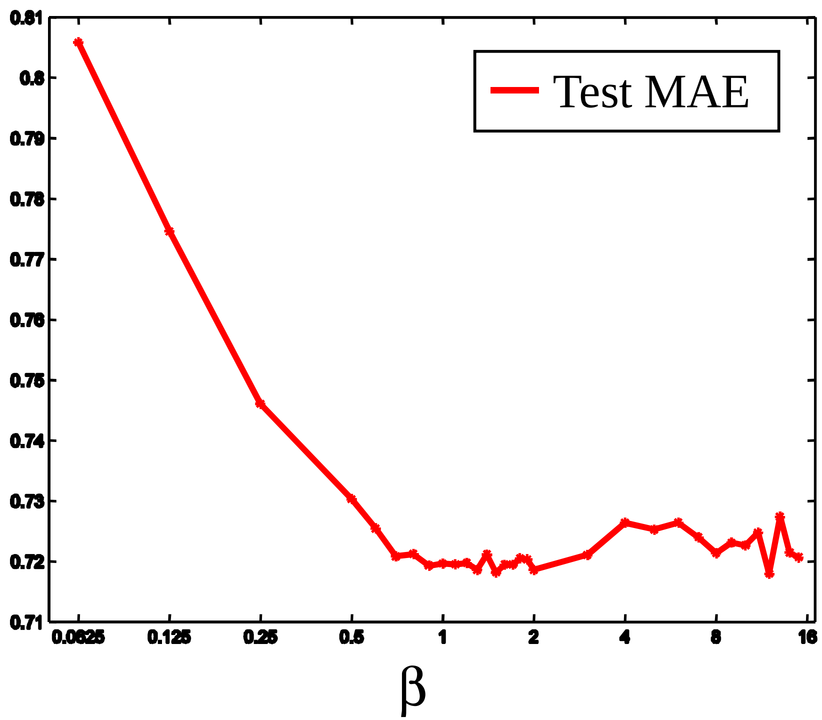
Application

- 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

13x6 Clusters

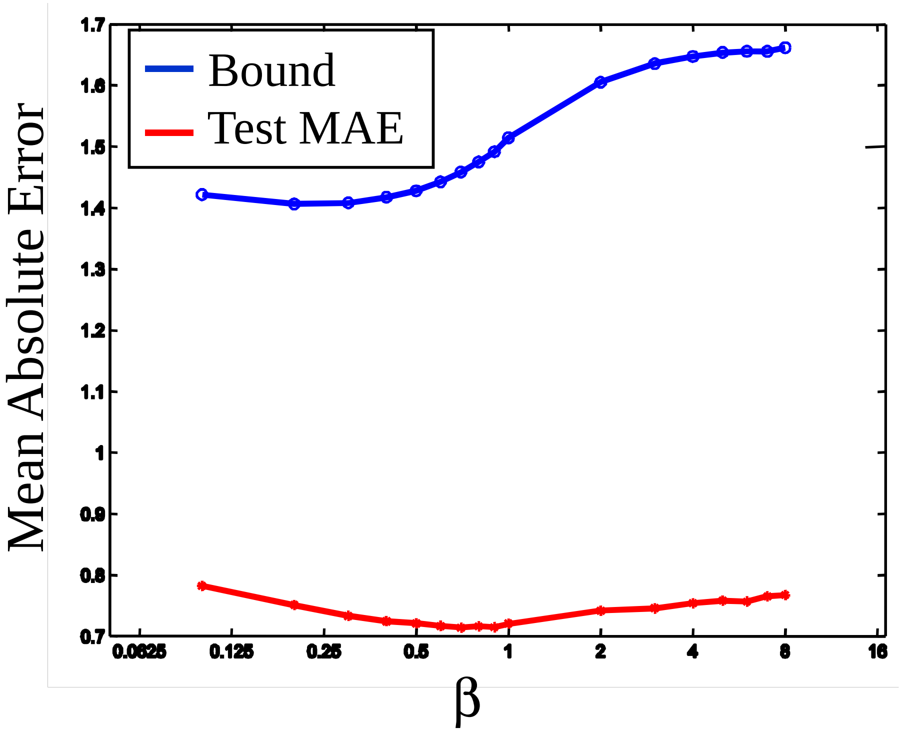
$$F(Q) = \beta N\hat{L}(Q) + \sum_i |X_i| I(X_i; C_i)$$

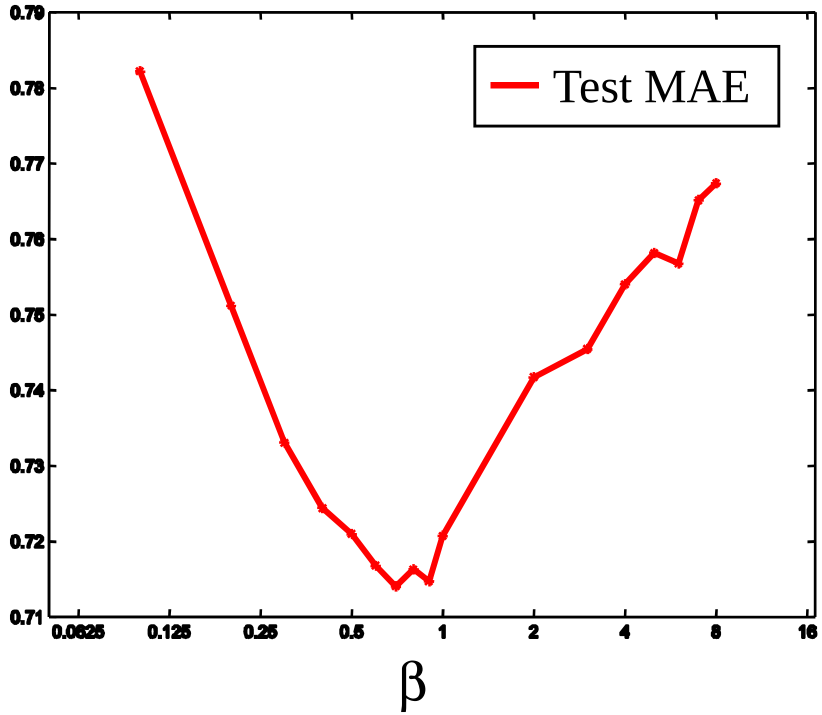




50x50 Clusters

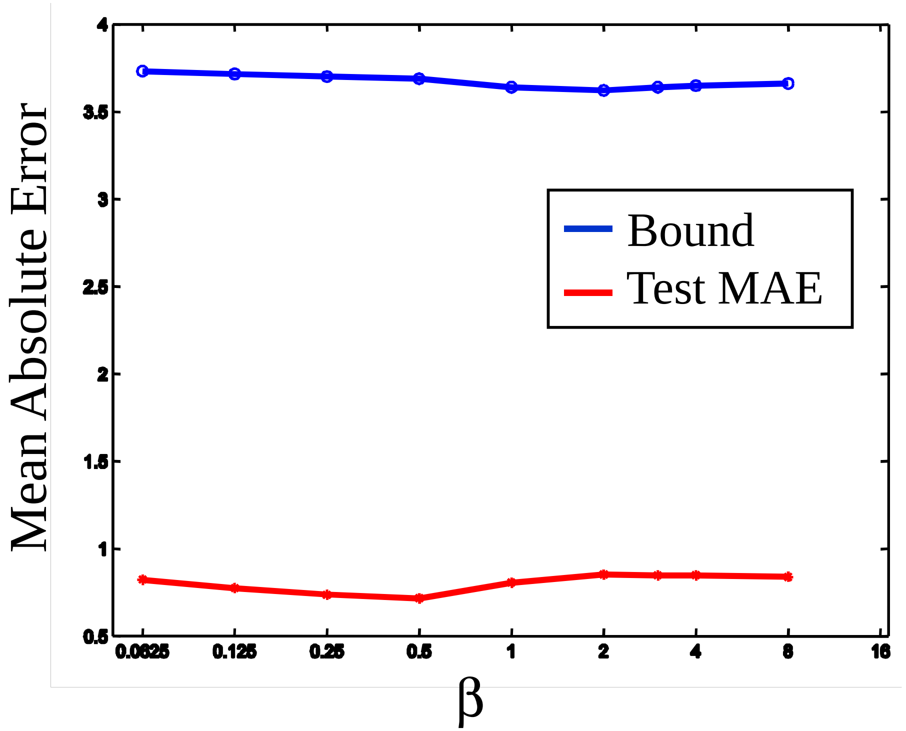
$$F(Q) = \beta N \hat{L}(Q) + \sum_i |X_i| I(X_i; C_i)$$

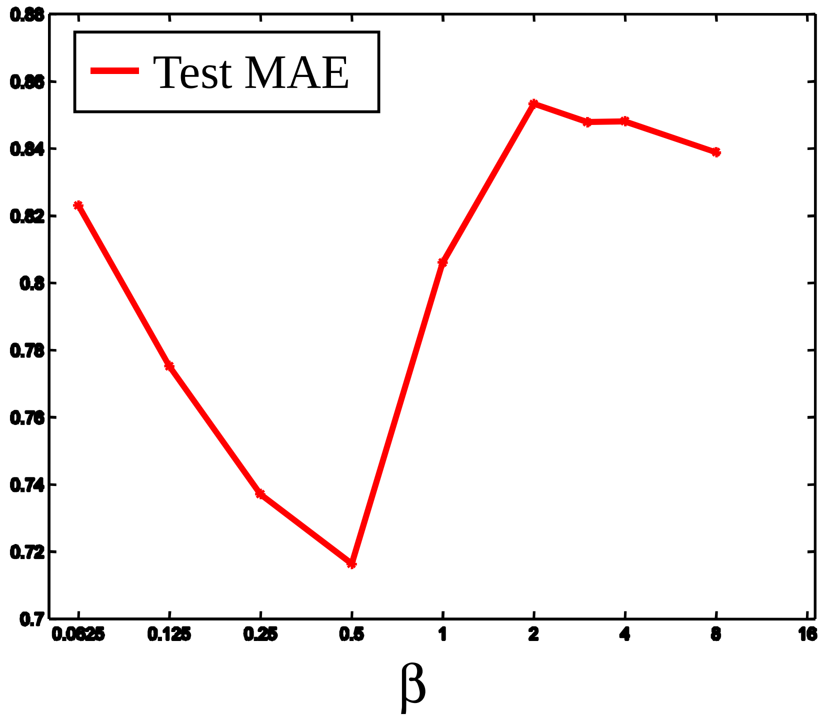




283x283 Clusters

$$F(Q) = \beta N\hat{L}(Q) + \sum_i |X_i| I(X_i; C_i)$$





Weighted Graph Clustering

- The weigts of the edges w_{ij} are generated by unknown distribution $p(w_{ij}|x_i,x_j)$

- Given a sample of size N of edge weights

- Build a model $q(w|x_1,x_2)$ such that
  $E_{p(x1,x2,w)}\ E_{q(w'|x1,x2)}\ l(w,w')$ is minimized

Other problems

- Pairwise clustering = clustering of a weighted graph
  - Edge weights = pairwise relations

- Clustering of unweigted graph
  - Present edges = weight 1
  - Absent edges = weight 0

Weighted Graph Clustering

- The weights of the links are generated according to:

  $q(w_{ij}|X_i,X_j) = \Sigma_{C_a,C_b}\, q(w_{ij}|C_a,C_b)\; q(C_a|X_i)\; q(C_b|X_j)$

- This is co-clustering with shared $q(C|X)$
  - Same bounds and (almost same) algorithms apply

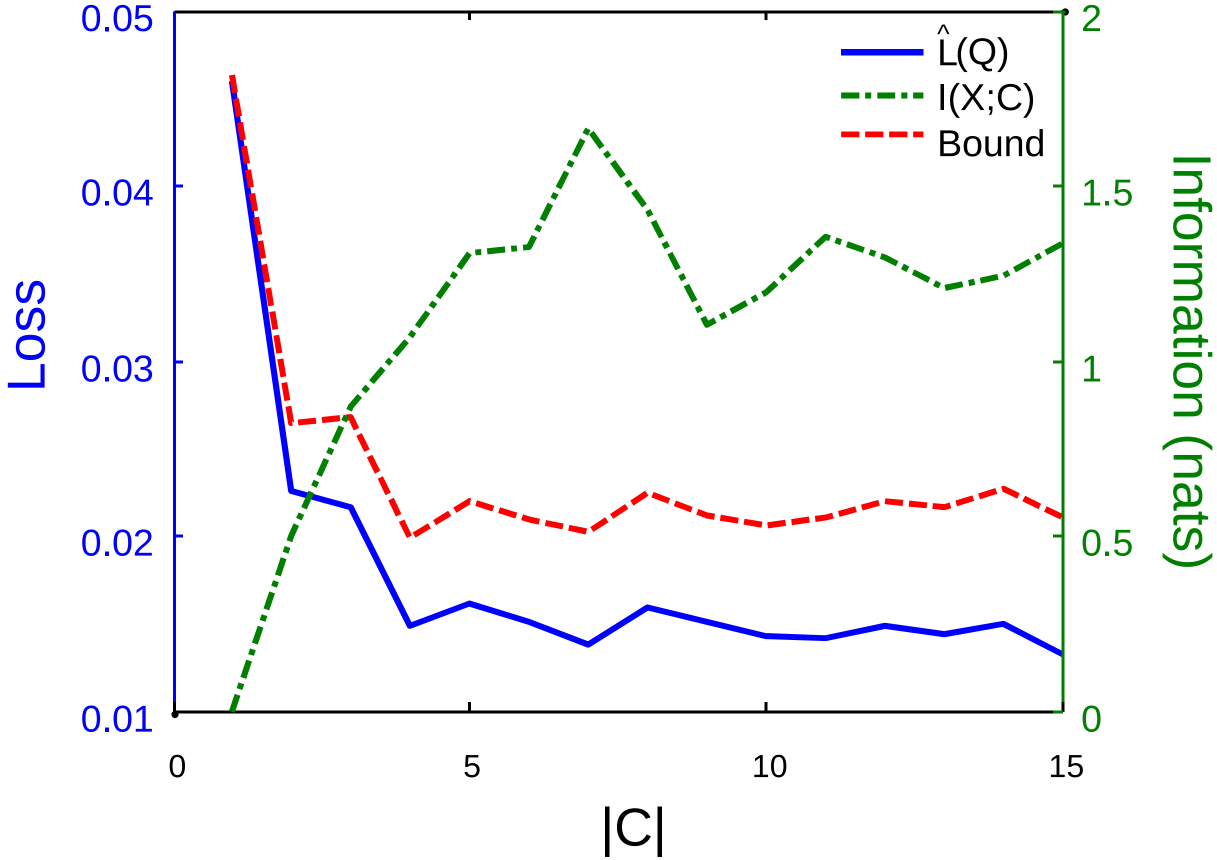
Application

- Optimize the trade-off

$$F(Q) = \beta N\hat{L}(Q) + |X| I(X;C)$$

- Kings dataset
  - Edge weights = exponentiated negative distance between DNS servers
  - |X| = 1740
  - Number of edges = 1,512,930

Graph Clustering Application



Relation with Matrix Factorization

- Co-clustering:
  - $g(X_1,X_2) = \Sigma_{C1,C2}\, q(C_1|X_1) g(C_1,C_2)\, q(C_2|X_2)$
  - $M \approx Q_1^T G Q_2$
- Graph clustering:
  - $g(X_1,X_2) = \Sigma_{C1,C2}\, q(C_1|X_1) g(C_1,C_2)\, q(C_2|X_2)$
  - $M \approx Q^T G Q$

Summary of main contributions

- Formulation of co-clustering and graph clustering (unsupervised learning) as prediction problems
- PAC-Bayesian analysis of co-clustering and graph clustering
  - Regularization terms
- Encouraging empirical results

Future Directions

- Practice:
  - More applications
- Theory:
  - Continuous domains
  - Multidimensional matrices

References

Co-clustering: Seldin & Tishby *JMLR* 2010 submitted, avail.online

Graph clustering: Seldin *Social Analytics* 2010