PAC-Bayesian Analysis in Unsupervised Learning

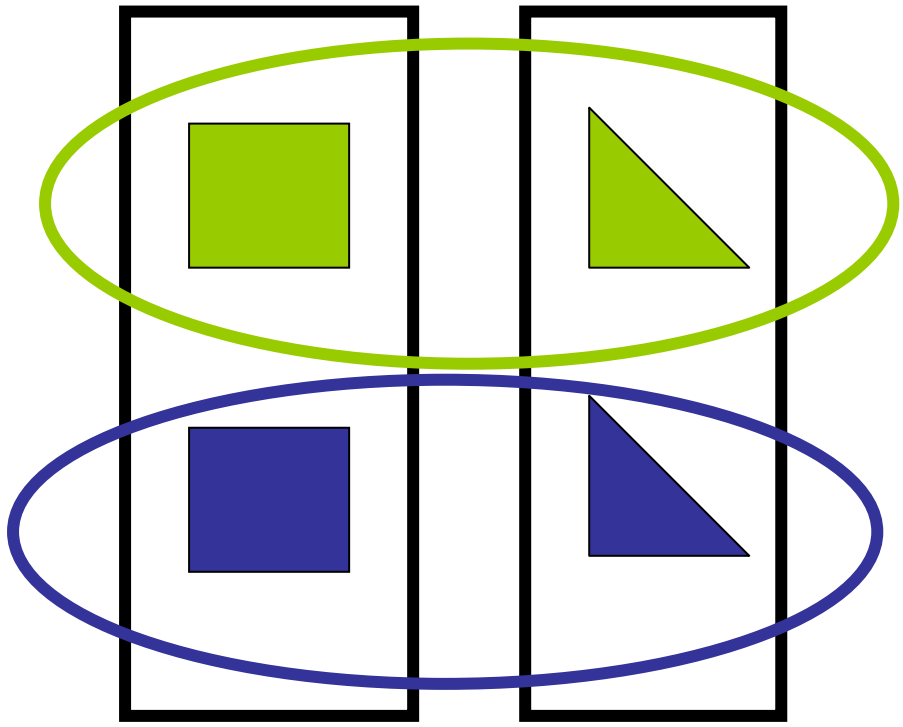
Yevgeny Seldin

Joint work with Naftali Tishby

Motivation

- Clustering and more general structure learning tasks are often ambiguous
- Many structures co-exist simultaneously
- The problem of comparison of solutions cannot be resolved by testing any property of the clustering itself

Example



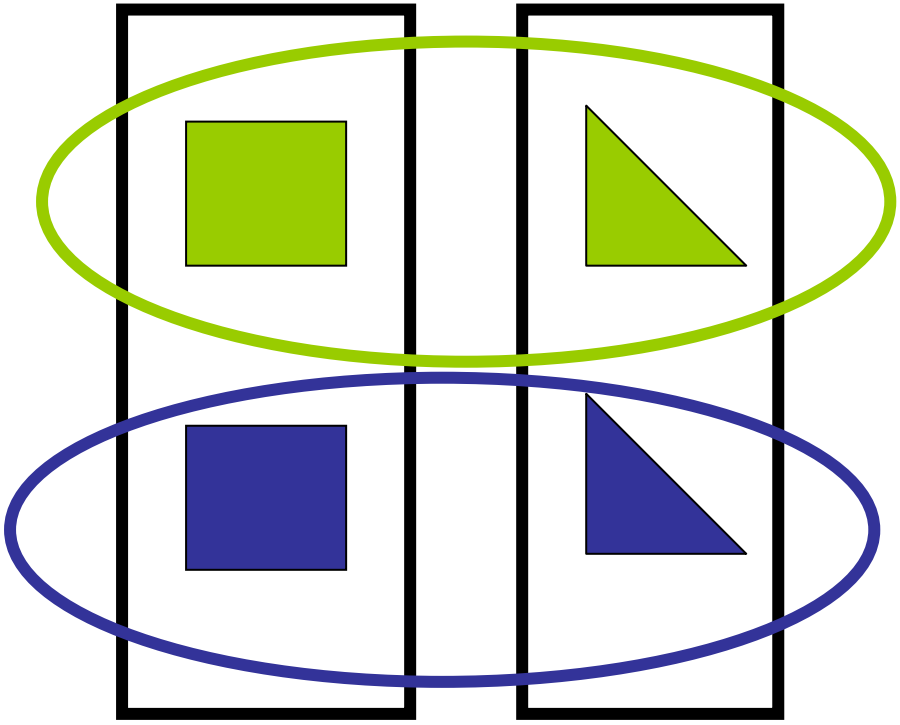
Motivation

- We do not cluster the data just for the sake of it, 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
- The main obstacle in the development of unsupervised learning is the absence of good formulations of its application contexts

Example

- Cluster then pack
- Clustering by shape is preferable

- Evaluate the amount of time saved

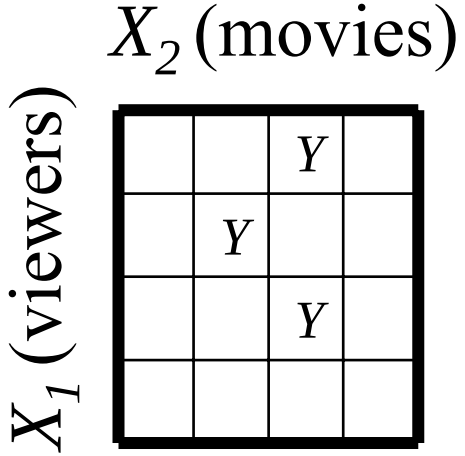


Outline

- Two problems behind co-clustering
  - Discriminative prediction
  - Density estimation
- PAC-Bayesian analysis of discriminative prediction with co-clustering
  - Combinatorial priors
- PAC-Bayesian bound for discrete density estimation
  - Density estimation with co-clustering
- Extensions

Discriminative Prediction with Co-clustering

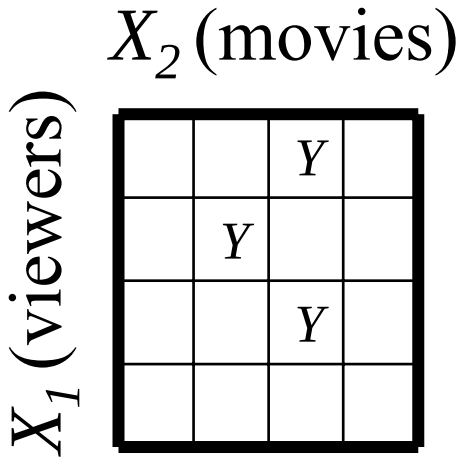
- Example: collaborative filtering
- Goal: find discriminative prediction rule $q(Y|X_1,X_2)$



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 $q(Y|X_1,X_2)$

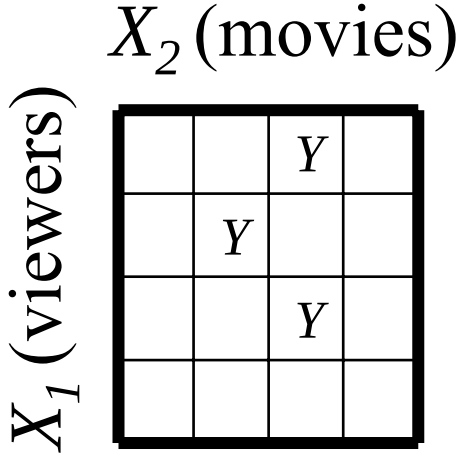
Given loss $l(Y,Y')$

Discriminative Prediction with Co-clustering

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

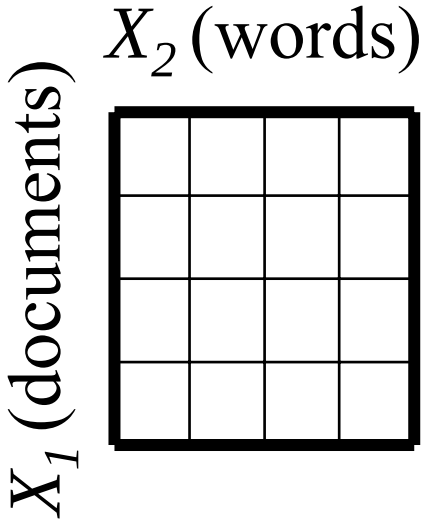
$$L(q) = E_{p(X_1,X_2,Y)} E_{q(Y'|X_1,X_2)} l(Y,Y')$$

- Model-independent comparison



Co-occurrence Data Analysis

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

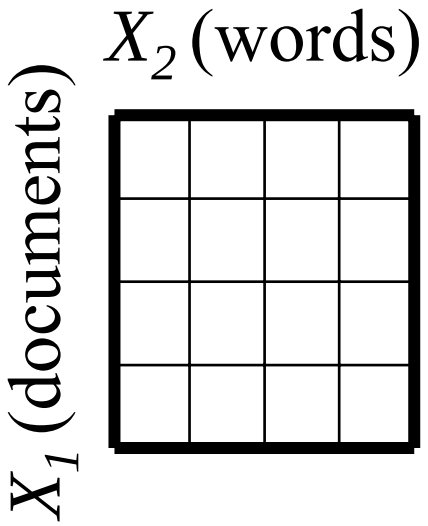


Co-occurrence Data Analysis

- 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 $p(X_1,X_2)$

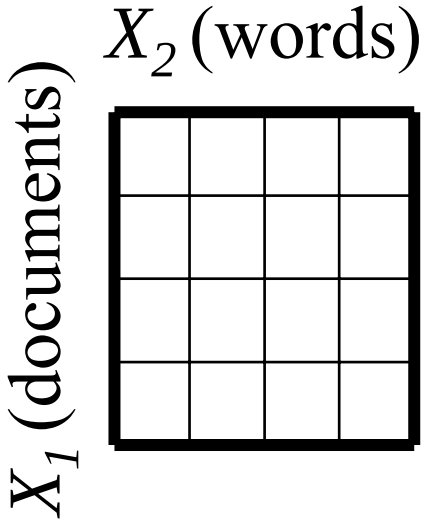


Density Estimation with Co-clustering

- 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:

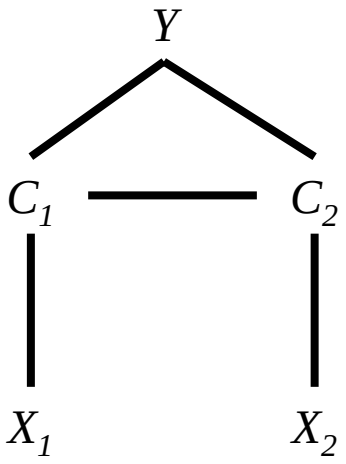
$$L(q) = -E_{p(X_1,X_2)} \ln q(X_1, X_2)$$

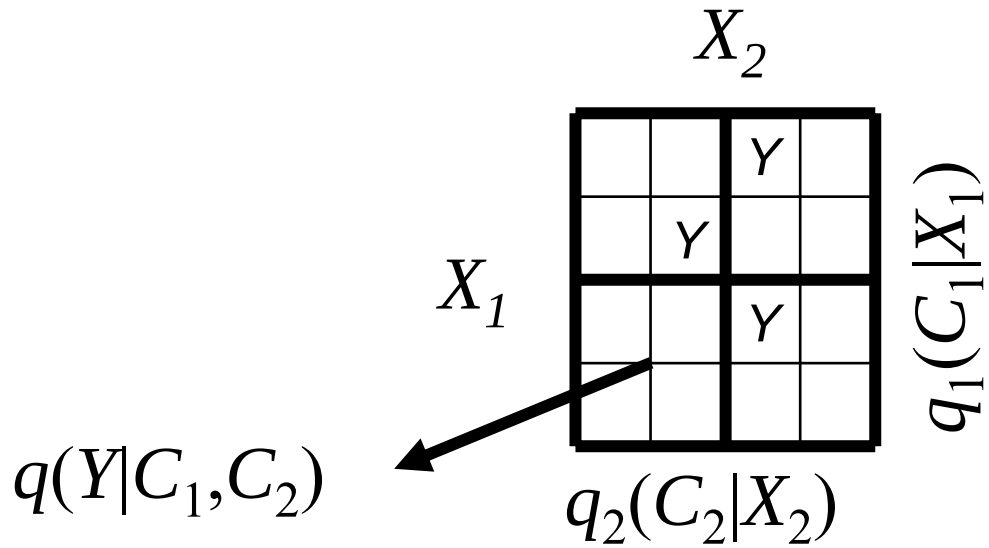
- Model-independent comparison



Discriminative prediction based on co-clustering

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

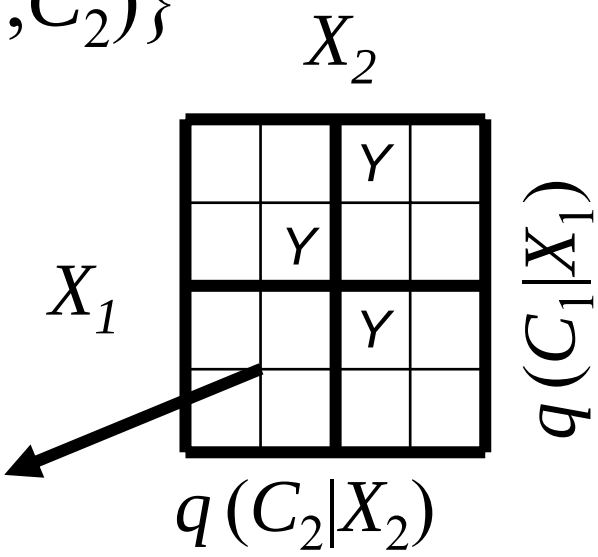




PAC-Bayesian Analysis

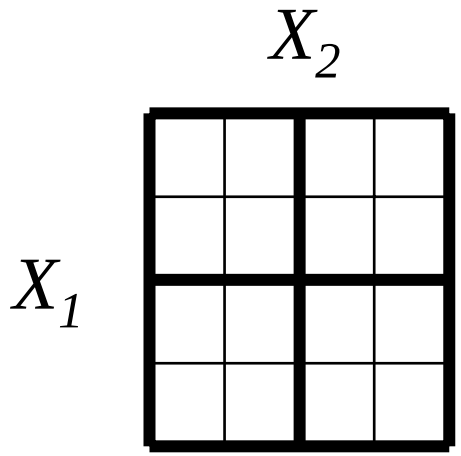
$$q(Y \mid X_1, X_2) = \sum_{C_1, C_2} q(Y \mid C_1, C_2) q(C_1 \mid X_1) q(C_2 \mid X_2)$$

- H – all hard partitions + labels for partition cells
- P – combinatorial (next slide)
- Q – $\{q(C_1|X_1), q(C_2|X_2), q(Y|C_1,C_2)\}$
- $q(x_i,c_i) = 1/|X_i|\ q(c_i|x_i)$



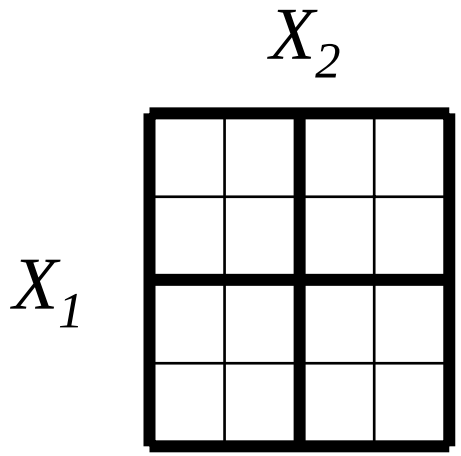
Prior Construction

- $|X_i|$ possibilities to choose $|C_i|$



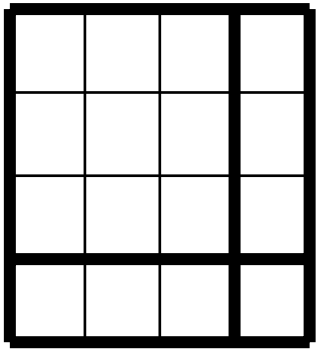
Prior Construction

- $|X_i|$ possibilities to choose $|C_i|$
- $\leq |X_i|^{|C_i|-1}$ possibilities to choose a cardinality profile along dimension i

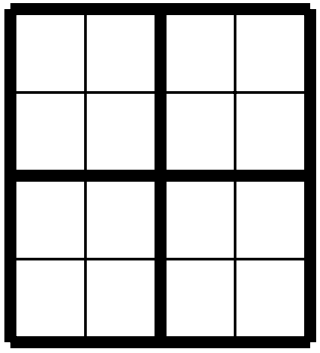


Prior Construction

- $|X_i|$ possibilities to choose $|C_i|$
- $\leq |X_i|^{|C_i|-1}$ possibilities to choose a cardinality profile along dimension i
- $\binom{|X_i|}{n_i^1,..,n_i^{|C_i|}} \leq e^{|X_i|H(C_i)}$ possibilities to assign X_i-s to C_i-s



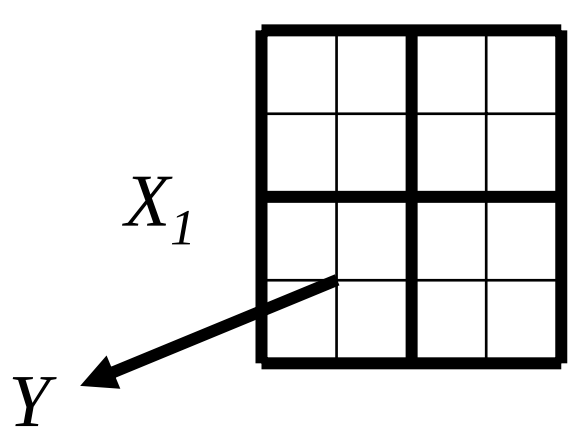
4 unbalanced partitions



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

Prior Construction

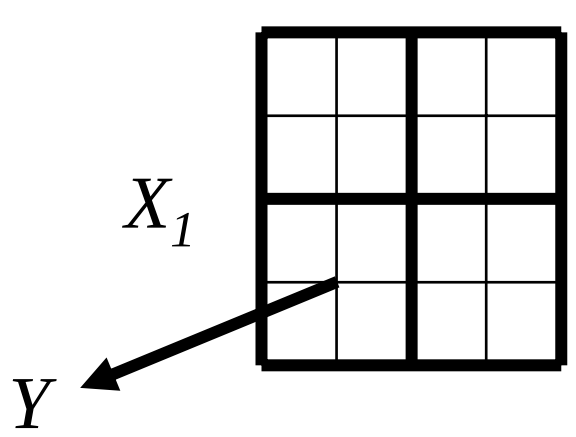
- $|X_i|$ possibilities to choose $|C_i|$
- $\leq |X_i|^{|C_i|-1}$ possibilities to choose a cardinality profile along dimension i
- $\binom{|X_i|}{n_i^1,..,n_i^{|C_i|}} \leq e^{|X_i|H(C_i)}$ possibilities to assign X_i-s to C_i-s
- $|Y|^{|C_1||C_2|}$ possibilities to assign labels to partition cells



Prior Construction

- $|X_i|$ possibilities to choose $|C_i|$
- $\leq |X_i|^{|C_i|-1}$ possibilities to choose a cardinality profile along dimension i
- $\binom{|X_i|}{n_i^1,..,n_i^{|C_i|}} \leq e^{|X_i|H(C_i)}$ possibilities to assign X_i-s to C_i-s
- $|Y|^{|C_1||C_2|}$ possibilities to assign labels to partition cells

$$P(h) \geq e^{\sum_i [-|X_i|H(C_i)-|C_i|\ln|X_i|]-|C_1||C_2|\ln|Y|}$$



Calculation of $D(Q||P)$

$$P(h) \geq e^{\sum_i [-|X_i| H(C_i) - |C_i| \ln |X_i|] - |C_1||C_2| \ln |Y|}$$

- $q(X_i, C_i) = 1/|X_i|\ q(C_i|X_i) \Rightarrow q(C_i) = 1/|X_i| \sum_{x_i} q(C_i|x_i)$

Calculation of $D(Q||P)$

$$P(h) \geq e^{\sum_i [-|X_i| H(C_i) - |C_i| \ln |X_i|] - |C_1||C_2| \ln |Y|}$$

- $q(X_i, C_i) = 1/|X_i|\ q(C_i|X_i) \Rightarrow q(C_i) = 1/|X_i| \sum_{xi} q(C_i|x_i)$
- $D(Q||P) = \mathrm{E}_{Q(h)} - \ln P(h) - H(Q)$

Calculation of $D(Q||P)$

$$P(h) \geq e^{\sum_i [-|X_i| H(C_i) - |C_i| \ln |X_i|] - |C_1||C_2| \ln |Y|}$$

- $q(X_i, C_i) = 1/|X_i|\ q(C_i|X_i) \Rightarrow q(C_i) = 1/|X_i| \sum_{x_i} q(C_i|x_i)$
- $D(Q||P) = \mathrm{E}_{Q(h)} - \ln P(h) - H(Q)$
- $\mathrm{E}_{Q(h)} - \ln P(h) \leq \sum_i [|X_i| H(C_i) + |C_i| \ln |X_i|] + |C_1||C_2| \ln |Y|$

Calculation of $D(Q||P)$

$$P(h) \geq e^{\sum_i [-|X_i| H(C_i) - |C_i| \ln |X_i|] - |C_1||C_2| \ln |Y|}$$

- $q(X_i, C_i) = 1/|X_i|\ q(C_i|X_i) \Rightarrow q(C_i) = 1/|X_i| \sum_{x_i} q(C_i|x_i)$
- $D(Q||P) = \mathrm{E}_{Q(h)} - \ln P(h) - H(Q)$
- $\mathrm{E}_{Q(h)} - \ln P(h) \leq \sum_i [|X_i| H(C_i) + |C_i| \ln |X_i|] + |C_1||C_2| \ln |Y|$
- $-H(Q) \leq -|X_i| H(C_i|X_i)$

Calculation of $D(Q||P)$

$$P(h) \geq e^{\sum_i [-|X_i|H(C_i)-|C_i|\ln|X_i|]-|C_1||C_2|\ln|Y|}$$

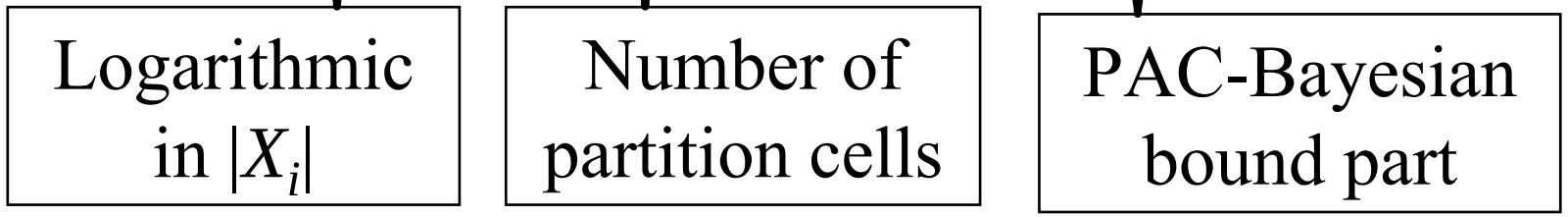
- $q(X_i,C_i) = 1/|X_i|\ q(C_i|X_i) \Rightarrow q(C_i)=1/|X_i|\ \Sigma_{xi} q(C_i|x_i)$
- $D(Q||P) = \mathrm{E}_{Q(h)}\text{-}\ln P(h) - H(Q)$
- $\mathrm{E}_{Q(h)}\text{-}\ln P(h) \leq \Sigma_i[|X_i|H(C_i)+|C_i|\ln|X_i|]+|C_1||C_2|\ln|Y|$
- $-H(Q) \leq -|X_i|H(C_i|X_i)$
- $D(Q||P) \leq \Sigma_i[|X_i|I(X_i;C_i)+|C_i|\ln|X_i|]+|C_1||C_2|\ln|Y|$

Generalization Bound

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

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

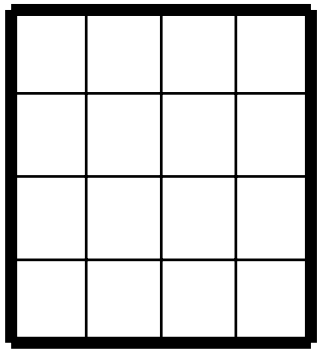
$$K = \sum_i |C_i| \ln |X_i| + \left(\prod_i |C_i| \right) \ln |Y| + \ln(4N)/2 - \ln \delta$$



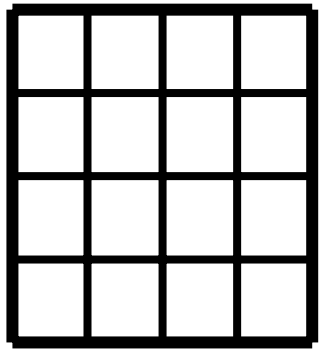
Generalization Bound

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

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



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



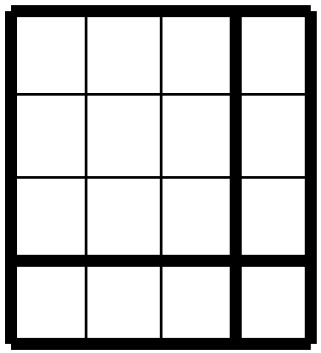
High Complexity
$I(X_i;C_i) = \ln|X_i|$

Generalization Bound

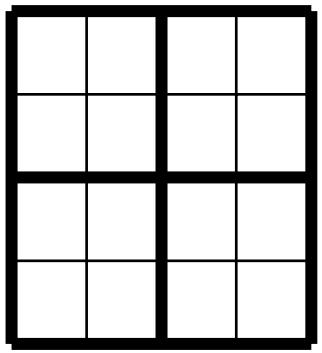
- With probability $\geq 1-\delta$:

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

Optimization tradeoff:
Empirical loss vs.
"Effective" partition
complexity



Lower
Complexity



Higher
Complexity

Application

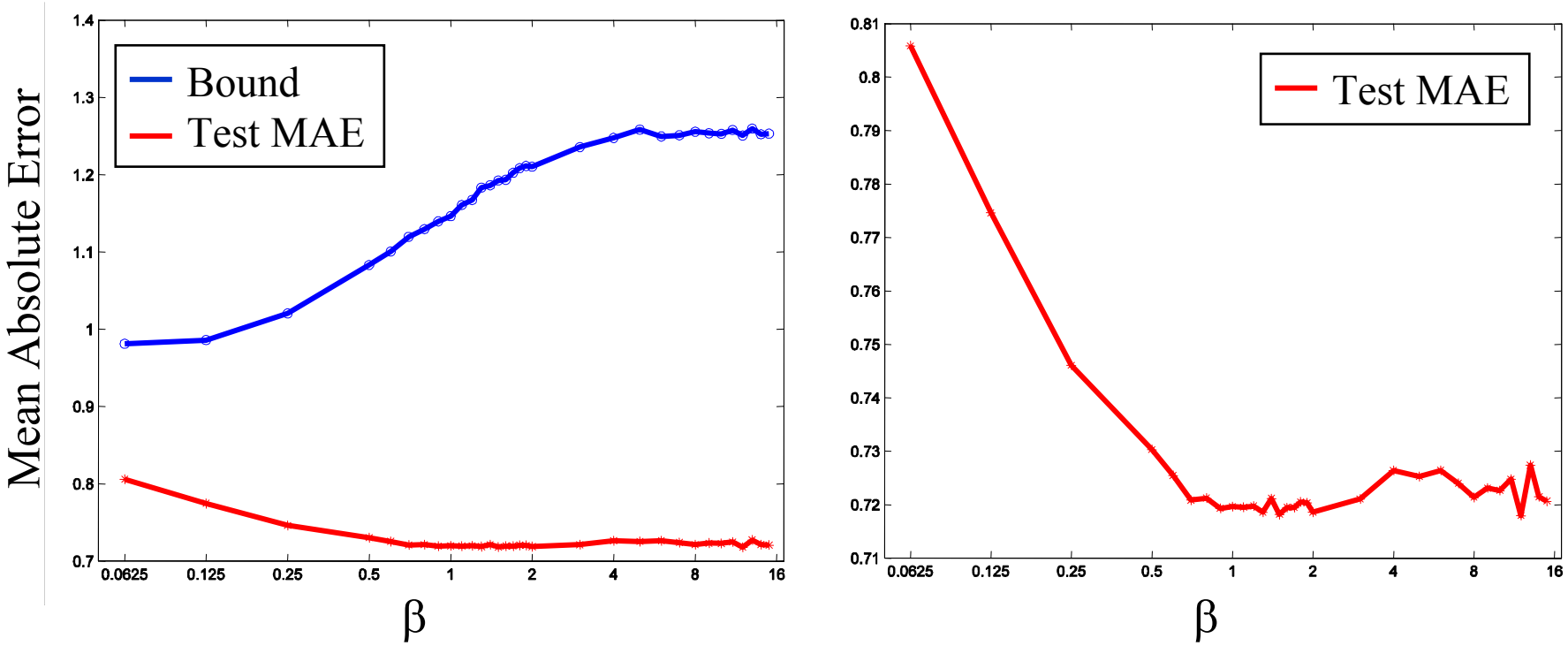
- Replace with a trade-off

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

- 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

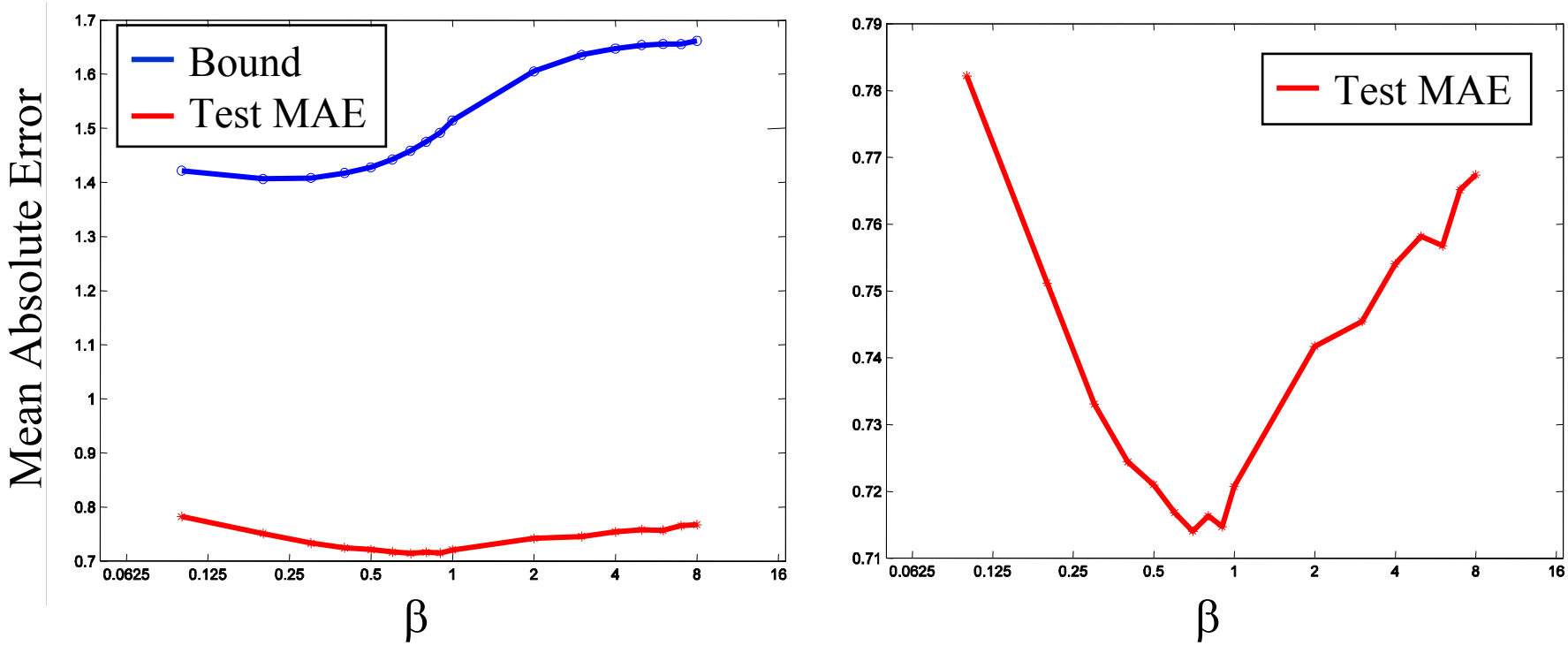
13x6 Clusters

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



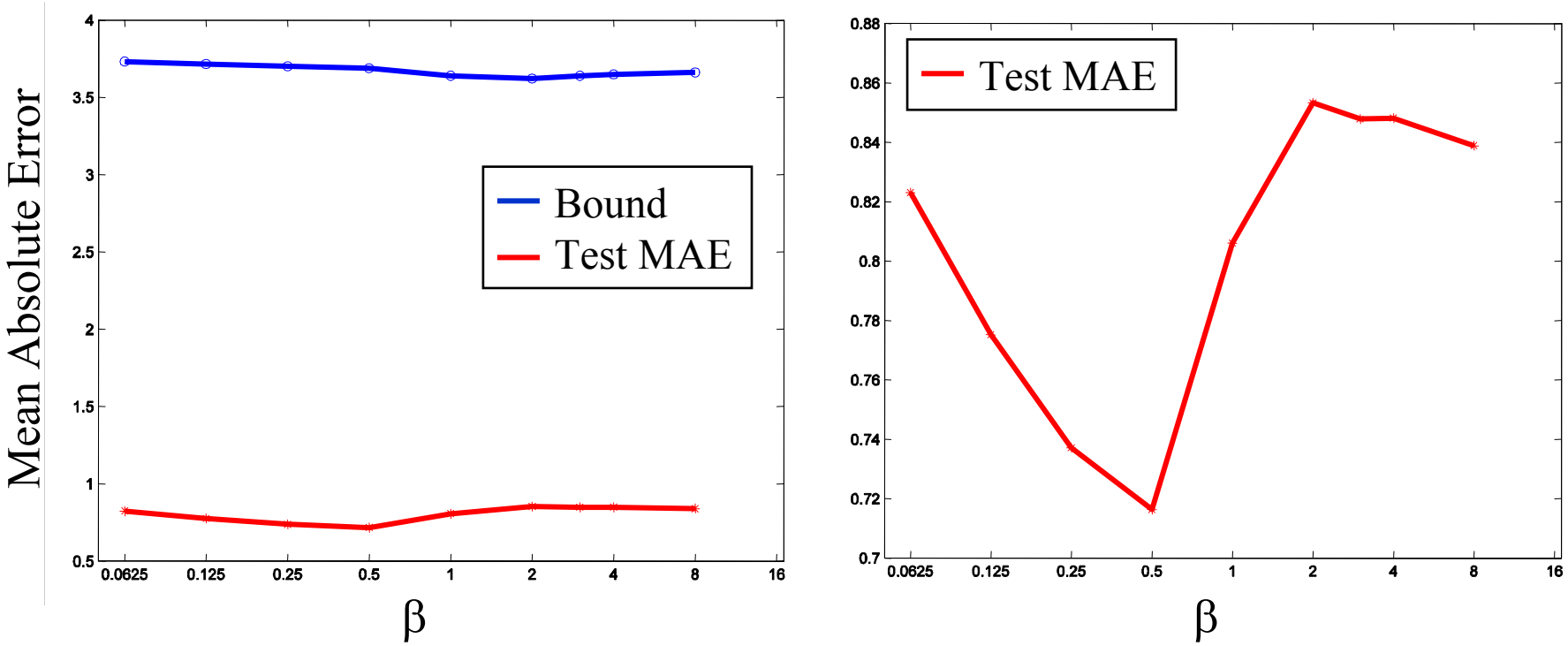
50x50 Clusters

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



283x283 Clusters

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



PAC-Bayesian Bound for Discrete Density Estimation

- X – sample space, H – hypothesis space
- Each $h \in H$ is a function $h: X \to Z$; Z - finite
- $p_h(Z) = \mathrm{P}_{X \sim p(X)}\{h(X)=Z\}$

PAC-Bayesian Bound for Discrete Density Estimation

- X – sample space, H – hypothesis space
- Each $h \in H$ is a function $h: X \to Z$; Z - finite
- $p_h(Z) = \mathrm{P}_{X \sim p(X)}\{h(X)=Z\}$
- P – prior over H, Q – posterior over H
- $p_Q(Z) = \mathrm{E}_{Q(h)} p_h(Z)$

PAC-Bayesian Bound for Discrete Density Estimation

- X – sample space, H – hypothesis space
- Each $h \in H$ is a function $h: X \to Z$; Z - finite
- $p_h(Z) = \mathrm{P}_{X \sim p(X)}\{h(X)=Z\}$
- P – prior over H, Q – posterior over H
- $p_Q(Z) = \mathrm{E}_{Q(h)} p_h(Z)$

- PAC-Bayesian Bound for Density Estimation:
  - With probability $\geq 1-\delta$ for all Q simultaneously:

$$D(\hat{p}_Q(Z) \,\|\, p_Q(Z)) \leq \frac{D(Q \,\|\, P) + (|Z| - 1)\ln(N+1) - \ln\delta}{N}$$

Special case: bound for classification

- Density estimation: with probability ≥ 1-δ:

$$D(\hat{p}_Q(Z) \,\|\, p_Q(Z)) \le \frac{D(Q \,\|\, P) + (|Z| - 1)\ln(N+1) - \ln\delta}{N}$$

- Classification = density estimation of the error variable
  - If Z is the error variable, then $|Z|=2$ and $p_Q(Z)=L(Q)$:

$$D(\hat{L}(Q) \,\|\, L(Q)) \le \frac{D(Q \,\|\, P) + \ln(N+1) - \ln\delta}{N}$$

Proof Idea

$$D(\hat{p}_Q(Z) \,\|\, p_Q(Z)) \leq \frac{D(Q \,\|\, P) + (|Z| - 1)\ln(N+1) - \ln \delta}{N}$$

- Change of measure inequality:

$$\mathrm{E}_{Q(h)}\phi(h) \leq D(Q\|P) + \ln \mathrm{E}_{P(h)}\mathrm{e}^{\phi(h)}$$

Proof Idea

$$D(\hat{p}_Q(Z) \,\|\, p_Q(Z)) \le \frac{D(Q \,\|\, P) + (|Z| - 1)\ln(N+1) - \ln \delta}{N}$$

- Change of measure inequality:

  $\mathrm{E}_{Q(h)}\phi(h) \le D(Q\|P) + \ln \mathrm{E}_{P(h)} \mathrm{e}^{\phi(h)}$

- Choose:

  $\phi(h) = ND(\hat{p}_h(Z) \,\|\, p_h(Z))$

- Show that $\mathrm{E}_S e^{\phi(h)} \le (N+1)^{|Z|-1}$ $\qquad S = \{X_1,..,X_N\}$

Proof Idea

$$D(\hat{p}_Q(Z) \,\|\, p_Q(Z)) \le \frac{D(Q \,\|\, P) + (|Z| - 1)\ln(N+1) - \ln \delta}{N}$$

- Change of measure inequality:

  $\mathrm{E}_{Q(h)}\phi(h) \le D(Q\|P) + \ln \mathrm{E}_{P(h)} \mathrm{e}^{\phi(h)}$

- Choose:

  $\phi(h) = ND(\hat{p}_h(Z) \,\|\, p_h(Z))$

- Show that $\mathrm{E}_S e^{\phi(h)} \le (N+1)^{|Z|-1}$ $\qquad S = \{X_1,..,X_N\}$

- $\mathrm{E}_S \mathrm{E}_{P(h)}\, e^{\phi(h)} = \mathrm{E}_{P(h)} \mathrm{E}_S e^{\phi(h)} \le (N+1)^{|Z|-1}$

A Bound on $\mathrm{E}_S e^{ND(\hat{p}_h(Z)\|p_h(Z))}$

$$E_S e^{ND(\hat{p}_h(Z)\|p_h(Z))} = \sum_{n_1,..,n_{|Z|}:\sum_i n_i = N} \binom{N}{n_1,..,n_{|Z|}} \prod_{i=1}^{|Z|} p_i^{N\hat{p}_i} e^{ND(\hat{p}\|p)}$$

$$\hat{p}_h(Z) = \left\{ \frac{n_1}{N},..., \frac{n_{|Z|}}{N} \right\}$$

A Bound on $\mathrm{E}_S e^{ND(\hat{p}_h(Z)\|p_h(Z))}$

$$E_S e^{ND(\hat{p}_h(Z)\|p_h(Z))} = \sum_{n_1,..,n_{|Z|}:\sum_i n_i = N} \binom{N}{n_1,..,n_{|Z|}} \prod_{i=1}^{|Z|} p_i^{N\hat{p}_i} e^{ND(\hat{p}\|p)}$$

$$\leq \sum_{n_1,..,n_{|Z|}:\sum_i n_i = N} e^{NH(\hat{p})} e^{N\sum_i \hat{p}_i \ln p_i} e^{ND(\hat{p}\|p)}$$

A Bound on $\mathrm{E}_S e^{ND(\hat{p}_h(Z)\|p_h(Z))}$

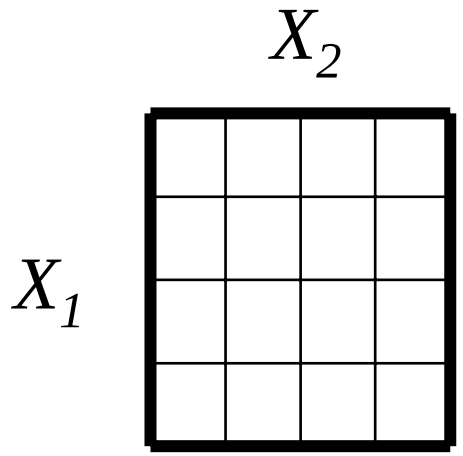
$$E_S e^{ND(\hat{p}_h(Z)\|p_h(Z))} = \sum_{n_1,..,n_{|Z|}:\sum_i n_i = N} \binom{N}{n_1,..,n_{|Z|}} \prod_{i=1}^{|Z|} p_i^{N\hat{p}_i} e^{ND(\hat{p}\|p)}$$

$$\leq \sum_{n_1,..,n_{|Z|}:\sum_i n_i = N} e^{NH(\hat{p})} e^{N\sum_i \hat{p}_i \ln p_i} e^{ND(\hat{p}\|p)}$$

$$= \sum_{n_1,..,n_{|Z|}:\sum_i n_i = N} 1 = \binom{N+|Z|-1}{|Z|} \leq (N+1)^{|Z|-1}$$

Density Estimation with Co-clustering

- Given: $\hat{p}(X_1, X_2)$
- Estimate: $p(X_1,X_2)$
- Minimize: $-\mathrm{E}_{p(X_1,X_2)} \ln q(X_1,X_2)$

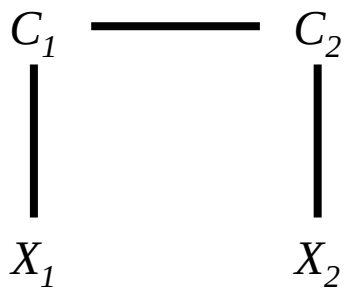


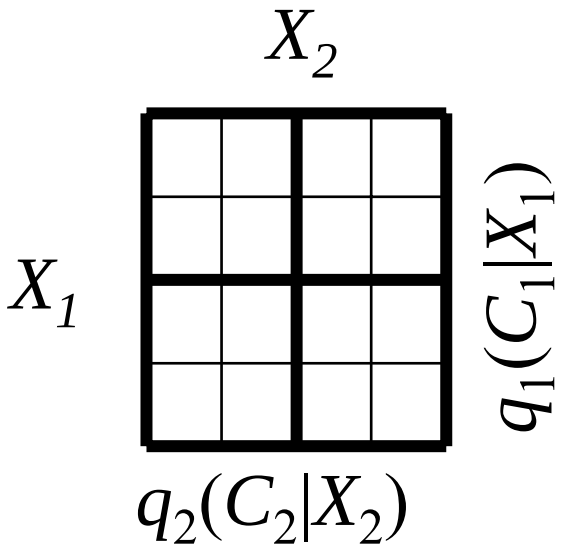
Density Estimation with Co-clustering

- Model: $Q = \{q(C_1|X_1), q(C_2|X_2)\}$

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

$$= \sum_{C_1, C_2} q(C_1, C_2) \prod_{i=1}^{2} \frac{q(X_i)}{q(C_i)} q(C_i \mid X_i)$$



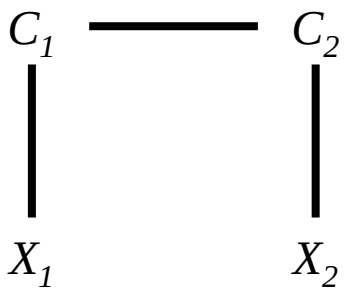


Density Estimation with Co-clustering

- Model: $Q = \{q(C_1|X_1), q(C_2|X_2)\}$

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

$$= \sum_{C_1, C_2} q(C_1, C_2) \prod_{i=1}^{2} \frac{q(X_i)}{q(C_i)} q(C_i \mid X_i)$$

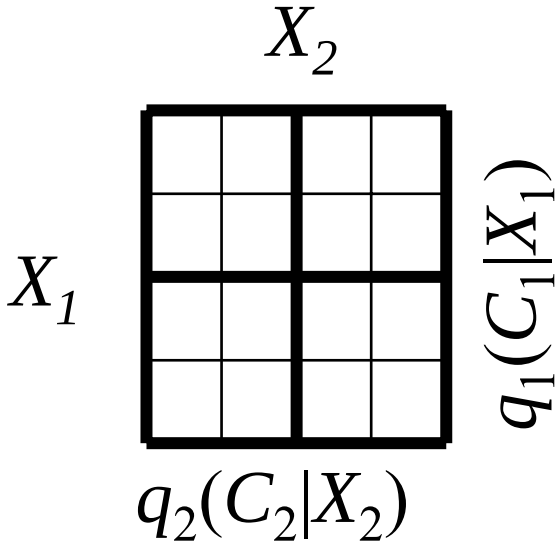


- Estimator:

$$\tilde{p}_Q(C_1, C_2) \propto \hat{p}_Q(C_1, C_2) + \gamma_1$$

$$\tilde{p}_Q(C_i) \propto \hat{p}_Q(Ci) + \gamma_2$$

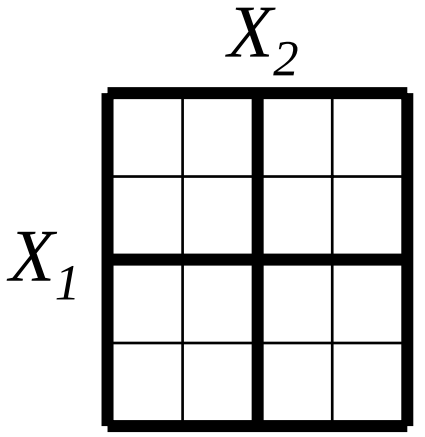
$$\tilde{p}(X_i) \propto \hat{p}(X_i) + \gamma_3$$



Density Estimation with Co-clustering

- Model: $Q = \{q(C_1|X_1), q(C_2|X_2)\}$

$$q(X_1, X_2) = \sum_{C_1, C_2} \tilde{p}_Q(C_1, C_2) \prod_{i=1}^{2} \frac{\tilde{p}(X_i)}{\tilde{p}_Q(C_i)} q(C_i \mid X_i)$$



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

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

Density Estimation with Co-clustering

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

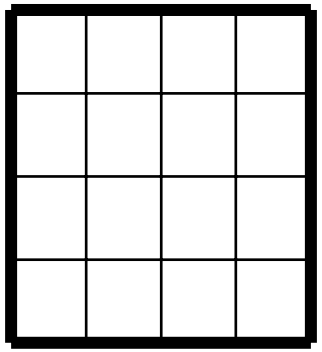
$$-E_{p(X_1,X_2)} \ln q(X_1, X_2) \leq -I(C_1;C_2) + \ln(|C_1||C_2|)\sqrt{\frac{\sum_i |X_i| I(X_i;C_i) + K_1}{2N}} + K_2$$

- 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

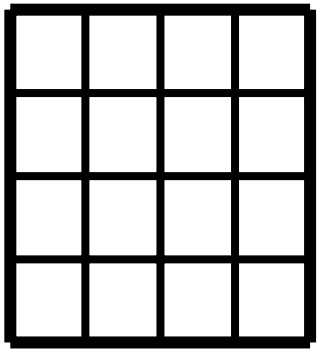
Density Estimation with Co-clustering

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

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



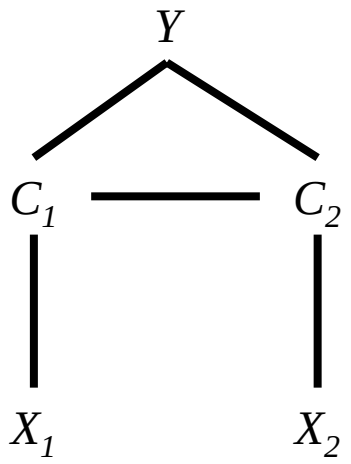
$I(C_1,C_2) = 0$
$I(X_i;C_i) = 0$



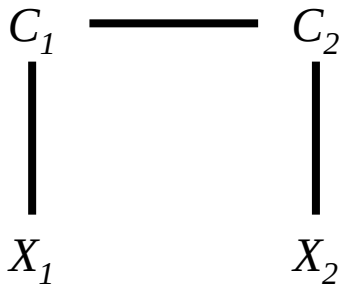
$I(C_1,C_2) = I(X_1,X_2)$
$I(X_i;C_i) = \ln|X_i|$

$$K_2 = \sum_i \hat{H}(X_i) + \ldots$$

Graphical Models



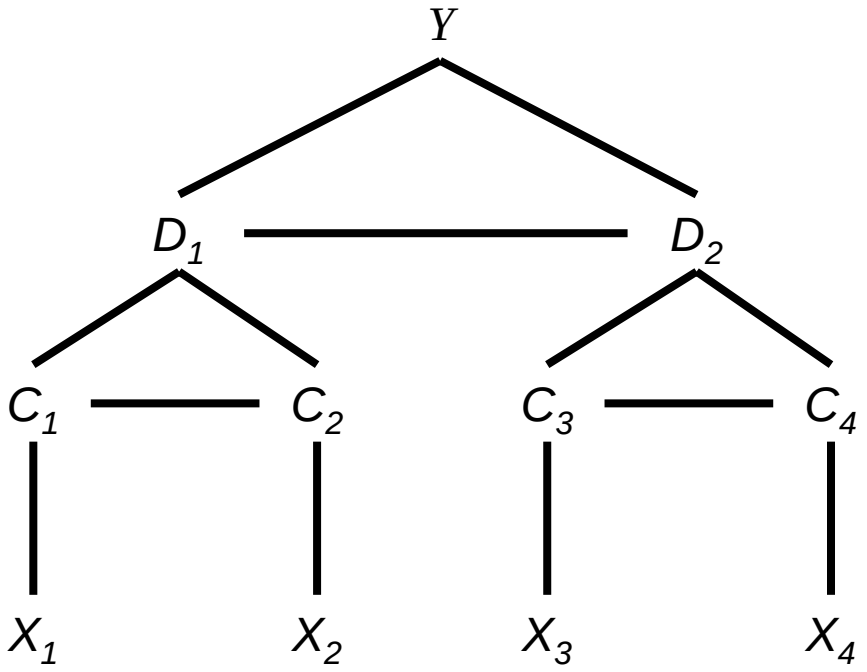
Discriminative
Prediction



Density
Estimation

Tree-Shaped Graphical Models

- Generalization bound:
  - Trade-off between empirical performance and the mutual information propagated up the tree
- Future work:
  - Optimization algorithms
  - Applications
  - More general graphical models



Graph Clustering, Pairwise 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 the co-clustering model with shared $q(C|X)$
  - Same bounds and algorithms apply

Summary of main contributions

- PAC-Bayesian analysis of unsupervised learning
  - Co-clustering
  - Tree-shaped graphical models
  - Graph clustering, pairwise clustering
- PAC-Bayesian bound for discrete density estimation
- Combinatorial priors $\Rightarrow$ mutual information regularization terms

Future Directions

- PAC-Bayesian analysis of unsupervised learning
  - Clustering
  - Continuous density estimation
  - General graphical models

- PAC-Bayesian analysis of reinforcement learning

References

Seldin & Tishby ICML 2008, AISTATS 2009, JMLR 2010 submitted