Discriminative Clustering with Cardinality Constraints

The 2018 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, Canada

Anh T. Pham (PhD student), **Raviv Raich**, and **Xiaoli Z. Fern**

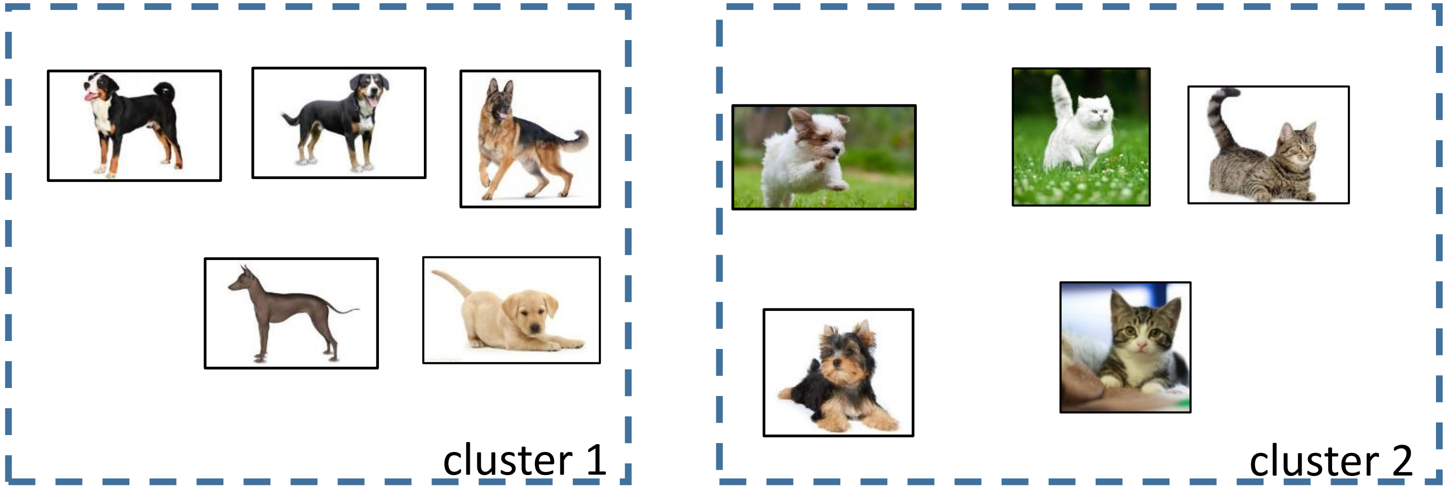
School of EECS, Oregon State University, Corvallis, OR 97331-5501, USA

{phaman,raich,xfern}@eecs.oregonstate.edu

Clustering

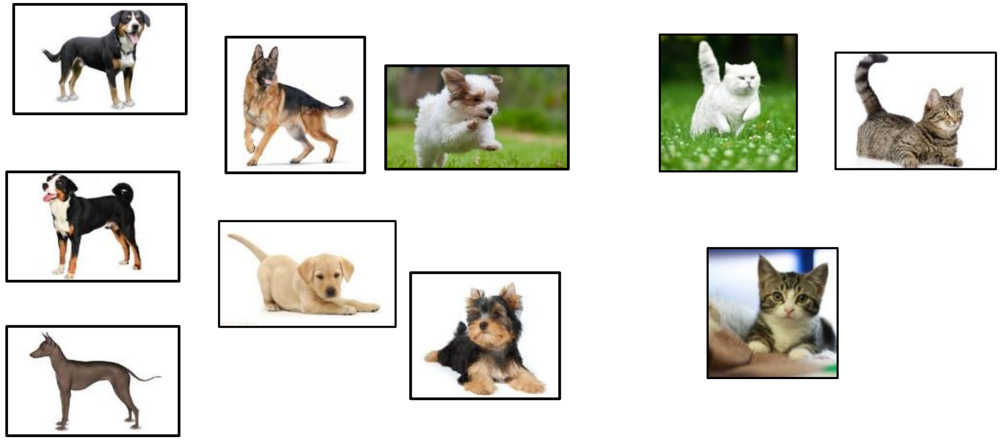
- Clustering is one of the most important tasks in machine learning [Jain’PRL10]: e.g., displaying news and search engines.

- **Goal:** grouping similar objects in the same cluster



Clustering results

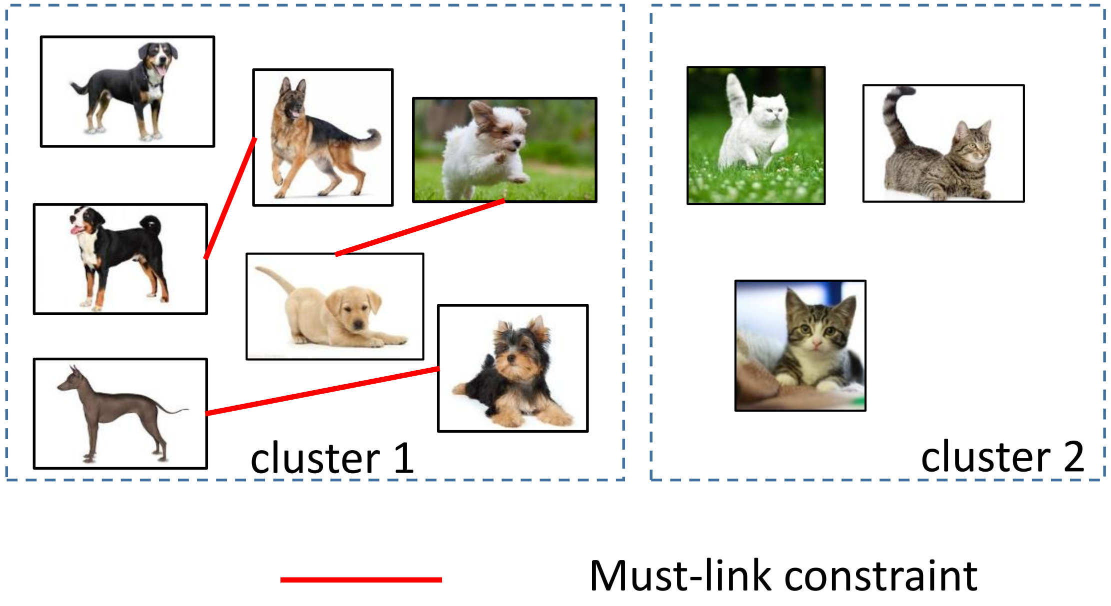
Constrained Clustering



Constrained Clustering

Instance-level constraints

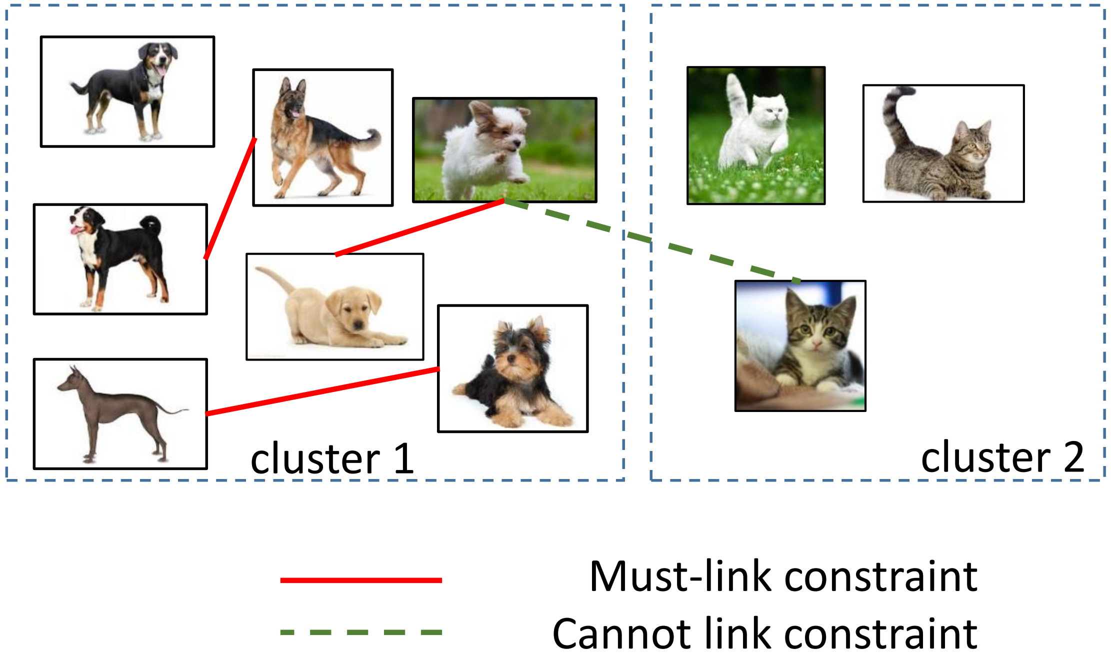
Clustering with pair-wise constraints



Constrained Clustering

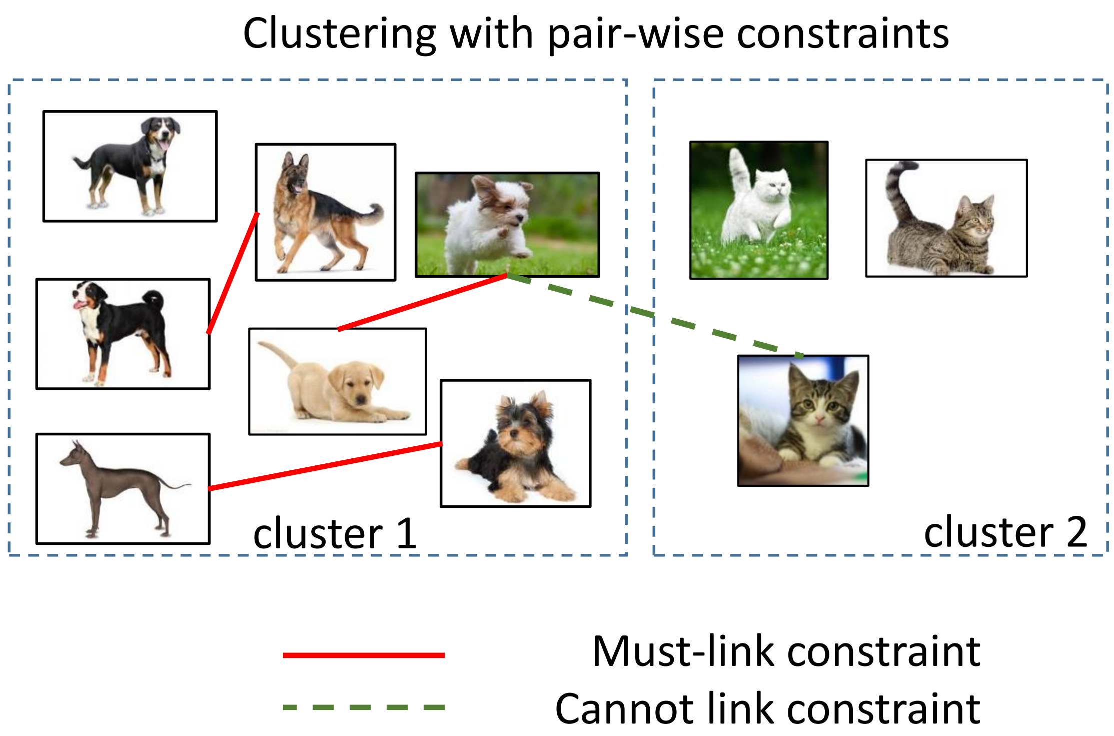
Instance-level constraints

Clustering with pair-wise constraints



Constrained Clustering

Instance-level constraints

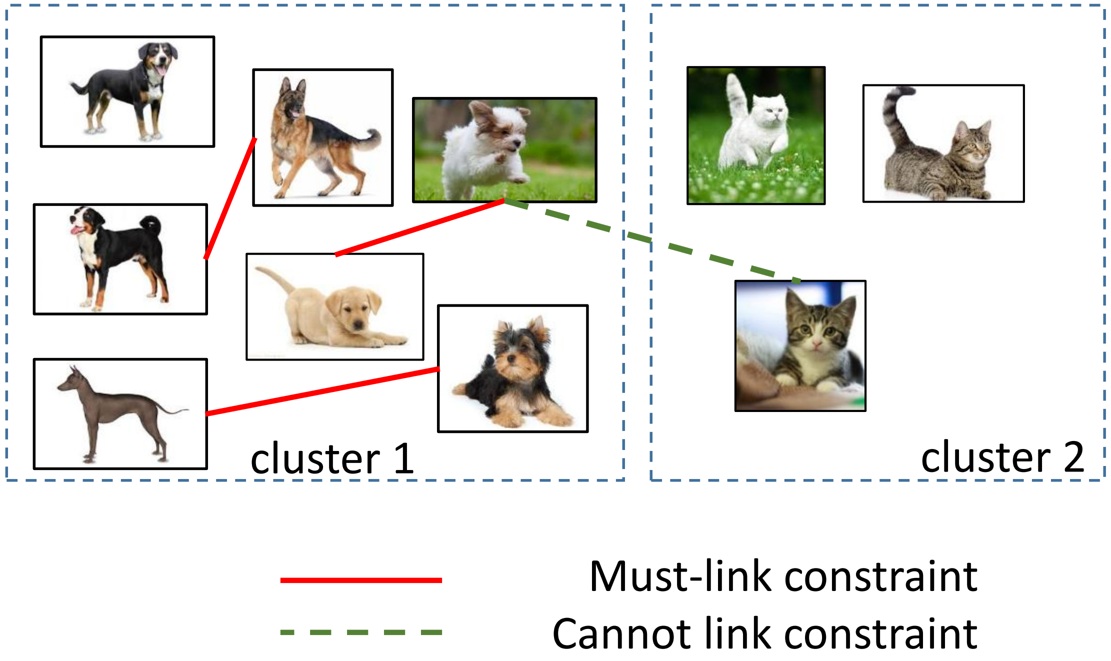


- *Well covered in literature* [Basu'SDM04, Bilenko'ICML04, Wagstaff'ICML01]

Constrained Clustering

Instance-level constraints

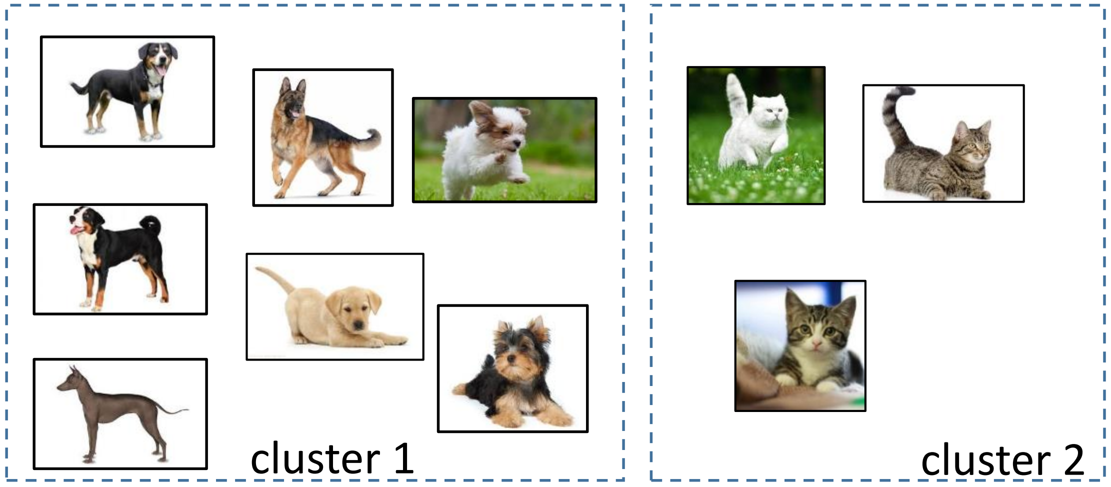
Clustering with pair-wise constraints



- *Well covered in literature* [Basu'SDM04, Bilenko'ICML04, Wagstaff'ICML01]

Group-level constraints

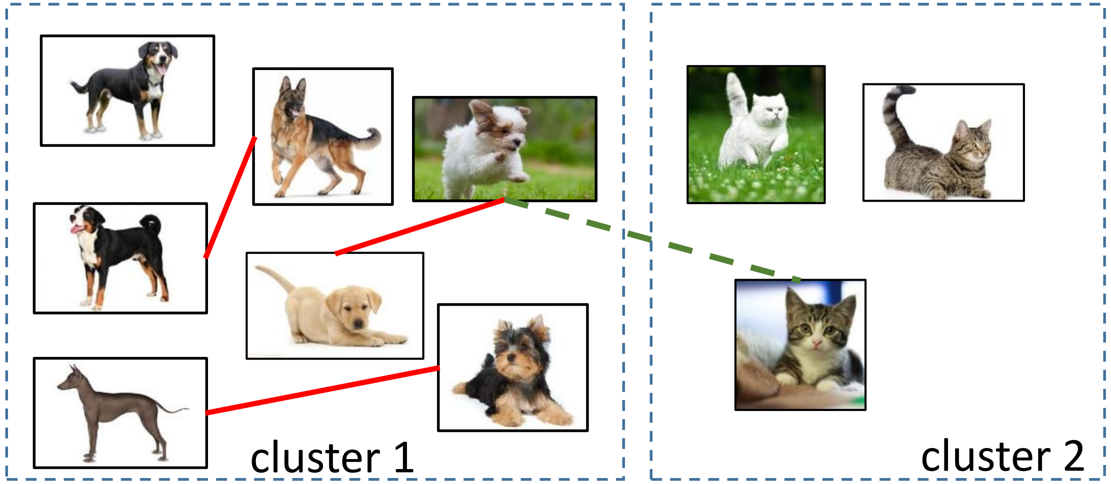
Clustering with cardinality constraints



Constrained Clustering

Instance-level constraints

Clustering with pair-wise constraints

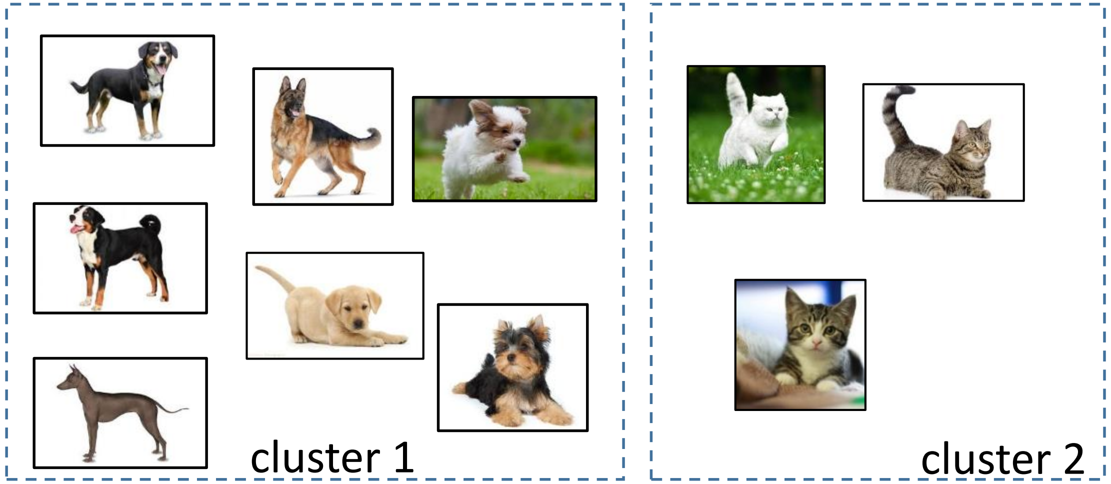


Must-link constraint
Cannot link constraint

- *Well covered in literature* [Basu'SDM04, Bilenko'ICML04, Wagstaff'ICML01]

Group-level constraints

Clustering with cardinality constraints

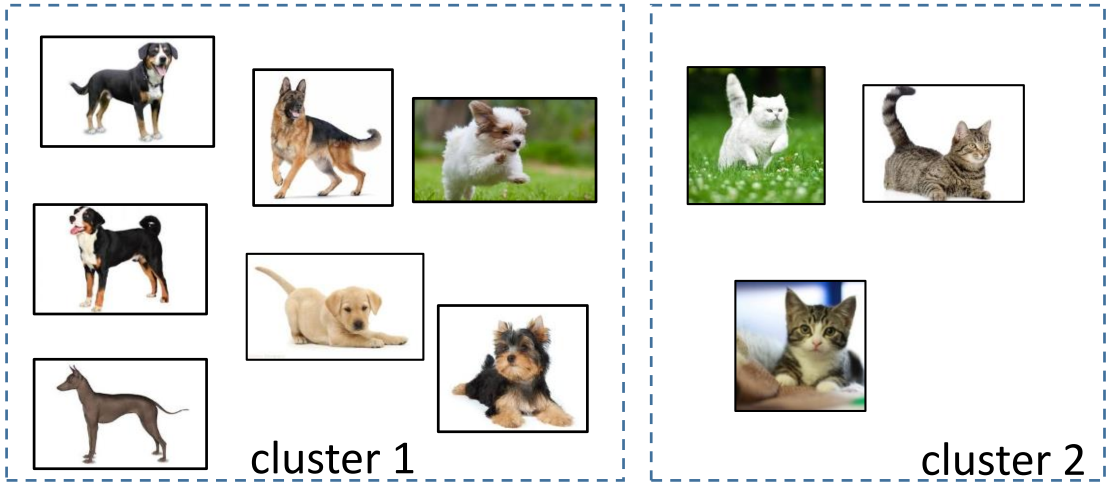


- *E.g., 7 images in cluster 1 and 3 images in cluster 2*

Constrained Clustering

Group-level constraints

Clustering with cardinality constraints



- *E.g., 7 images in cluster 1 and 3 images in cluster 2*
- *Limited coverage in literature*

<u>→ This work focuses on group-level constraints</u>

Applications

- Political election: [Quadrianto'JMLR09]

E.g., Clinton vs. Trump electoral map

State	**Date**	**Clinton**	**Trump**
Alaska	1/24	44	49
Arizona	4/26	42	35
California	5/2	56	34
Connecticut	4/12	48	40

Task: Cluster individuals by political affiliation

Applications

- Political election: [Quadrianto'JMLR09]

E.g., Clinton vs. Trump electoral map

State	Date	Clinton	Trump
Alaska	1/24	44	49
Arizona	4/26	42	35
California	5/2	56	34
Connecticut	4/12	48	40

Task: Cluster individual by political affiliation

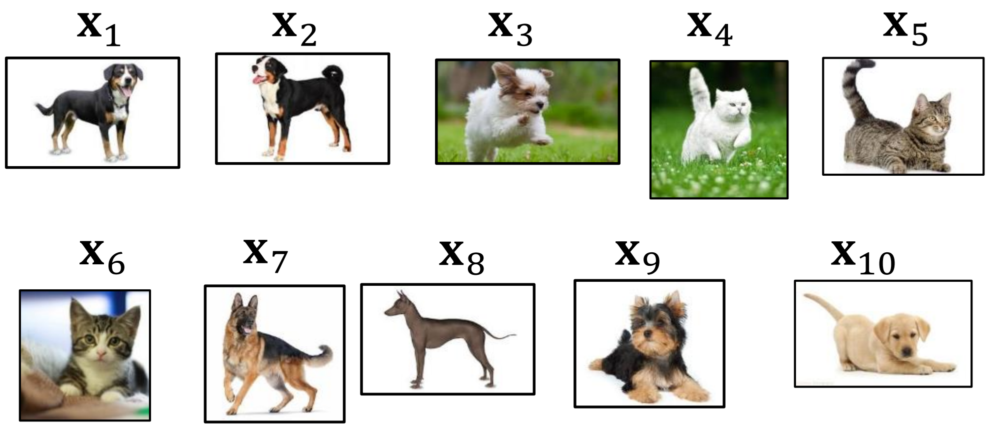
- Health-care data: [Yu'14]

E.g., Proportions of 2 types of diabetes

Task: Cluster type 1 versus type 2 diabetes (e.g., for drug recommendation)

Problem formulation

Observed data



$\mathbf{N} = [7, 3]$

- **Observed data:**
  - $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\mathrm{n}]$, where $\mathbf{x}_\mathrm{i} \in \mathbf{R}^\mathrm{d}$ denotes the i^{th} data point.
  - $\mathbf{N} = [\mathrm{N}_1, \mathrm{N}_2, \dots, \mathrm{N}_\mathrm{C}]$, where N_c indicates the number of samples in class c.

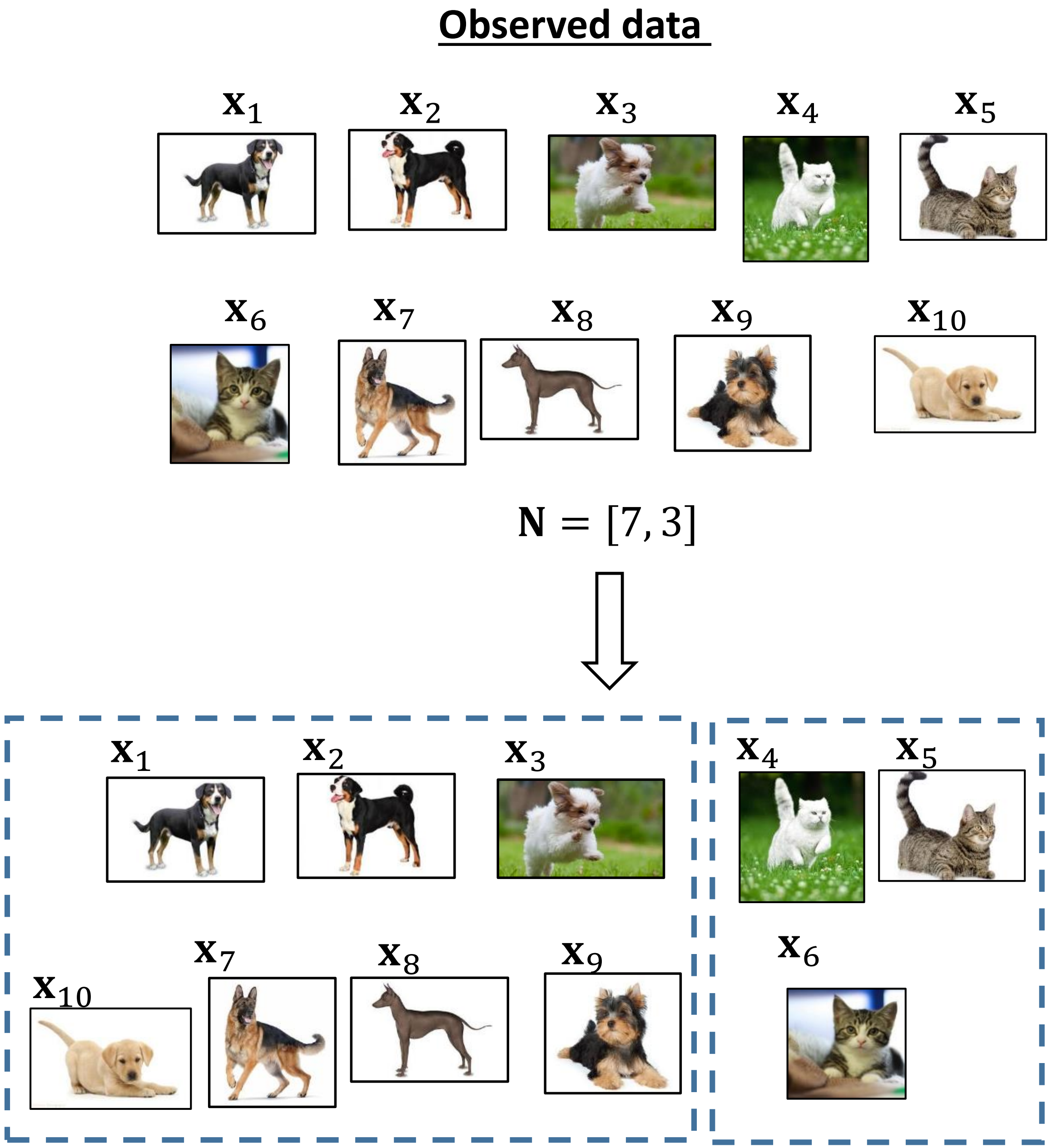
- **Hidden data:**
  - $\mathbf{Y} = [\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_\mathrm{n}]$ denotes the hidden label for each sample, $\mathrm{y}_\mathrm{i} \in \{1, 2, \dots, \mathrm{C}\}$.

Problem formulation

- **Observed data:**
  - $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$, where $\mathbf{x}_i \in \mathbf{R}^d$ denotes the i^{th} data point.
  - $\mathbf{N} = [N_1, N_2, \dots, N_C]$, where N_c indicates the number of samples in class c.

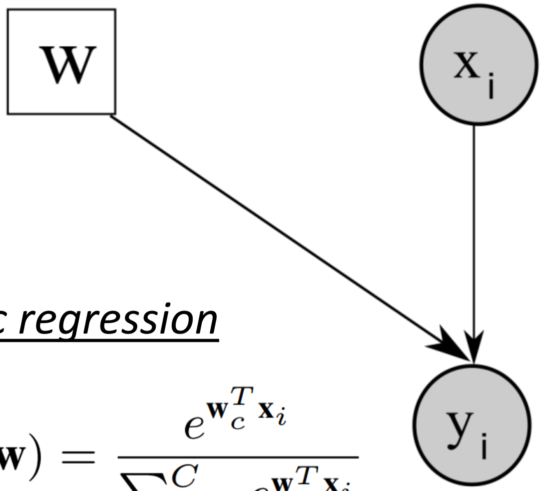
- **Hidden data:**
  - $\mathbf{Y} = [y_1, y_2, \dots, y_n]$ denotes the hidden label for each sample, $y_i \in \{1,2,\dots,C\}$.

- **Goal:**
  - Learn a mapping for each feature vector in $\mathbf{R}^d$ to a label in $\{1,2,\dots,C\}$.



Discriminative model with cardinality constraints

- Suppose label y_i is **known** for x_i, for all i

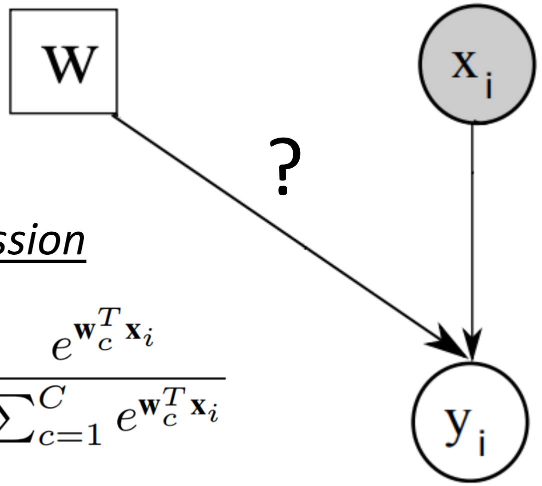


Logistic regression

$$p(y_i = c|\mathbf{x}_i, \mathbf{w}) = \frac{e^{\mathbf{w}_c^T \mathbf{x}_i}}{\sum_{c=1}^{C} e^{\mathbf{w}_c^T \mathbf{x}_i}}$$

Discriminative model with cardinality constraints (cont.)

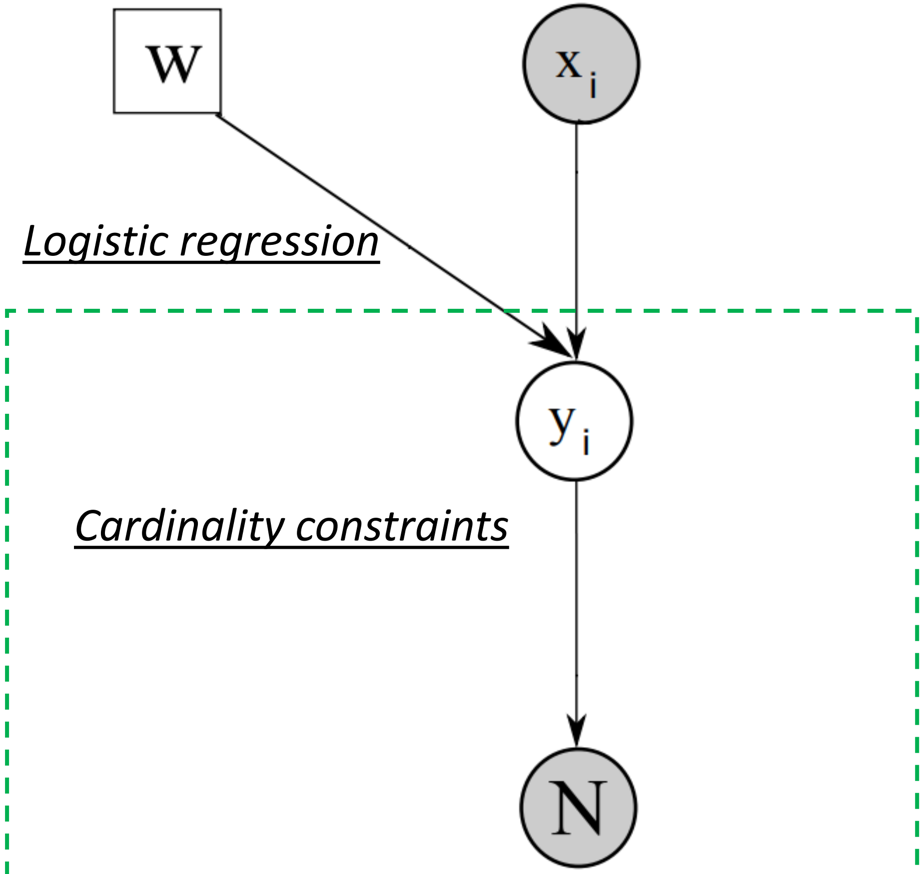
- However, y_i is **unknown**



Logistic regression

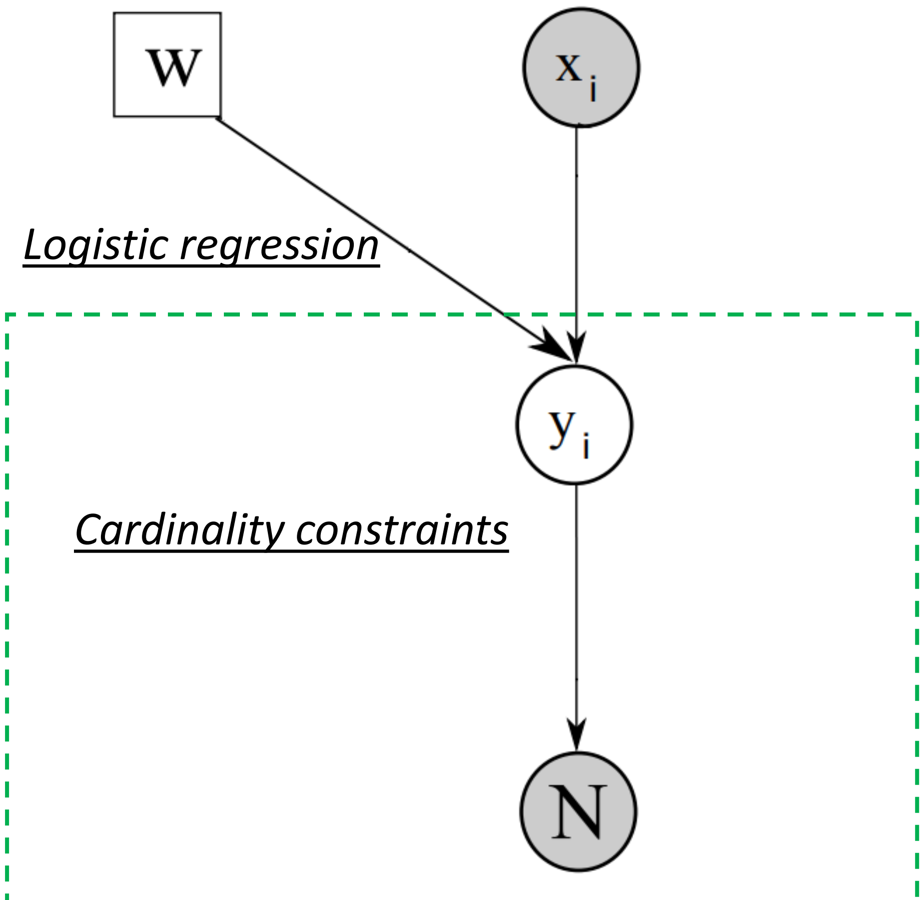
$$p(y_i = c|\mathbf{x}_i, \mathbf{w}) = \frac{e^{\mathbf{w}_c^T \mathbf{x}_i}}{\sum_{c=1}^{C} e^{\mathbf{w}_c^T \mathbf{x}_i}}$$

Discriminative model with cardinality constraints (cont.)



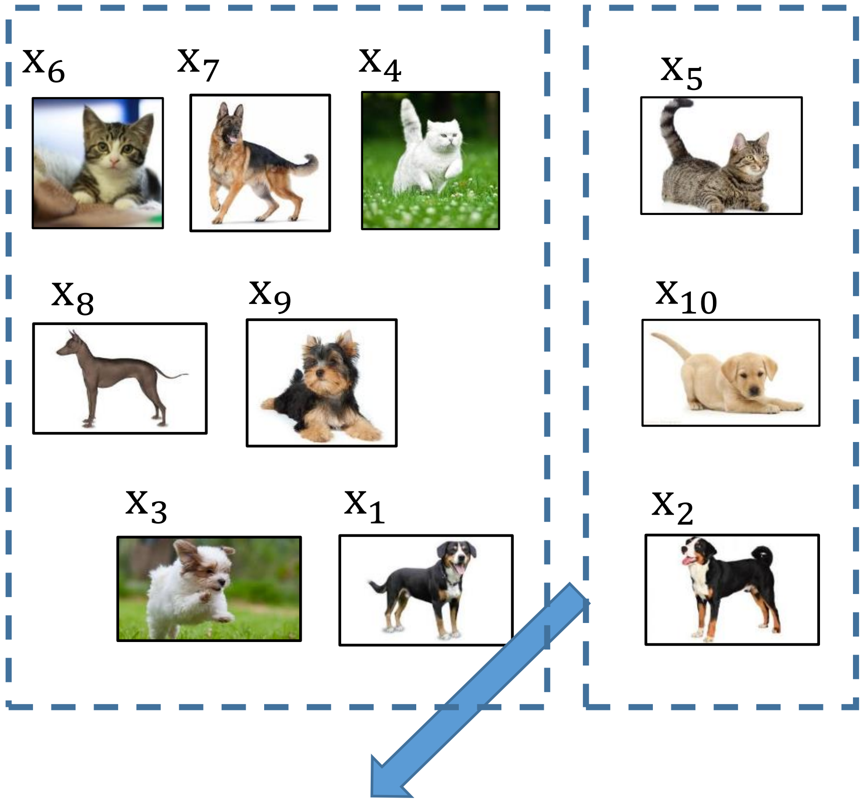
$$p(\mathbf{N}|\mathbf{y}) = \prod_{c=1}^{C} \mathbb{I}[N_c = \sum_{i=1}^{n} \mathbb{I}[y_i = c]\ ]$$

Discriminative model with cardinality constraints (cont.)



$$p(\mathbf{N}|\mathbf{y}) = \prod_{c=1}^{C} \mathbb{I}[N_c = \sum_{i=1}^{n} \mathbb{I}[y_i = c]\ ]$$

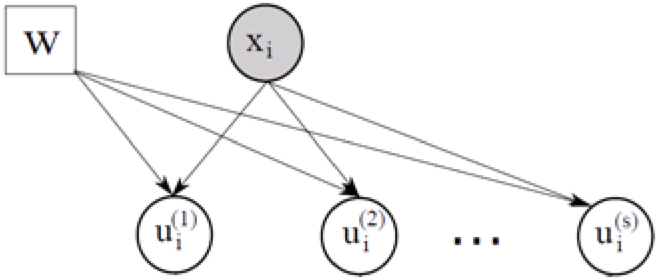
Challenge: Too many ways to partition given N (e.g., N = [7,3])



Crispness on the boundary may help

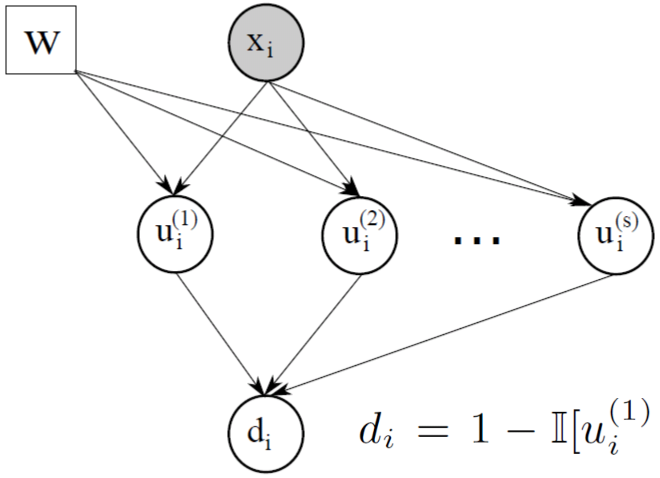
Model: Cluster crispness

- Generate s labels for each sample



Model: Cluster crispness

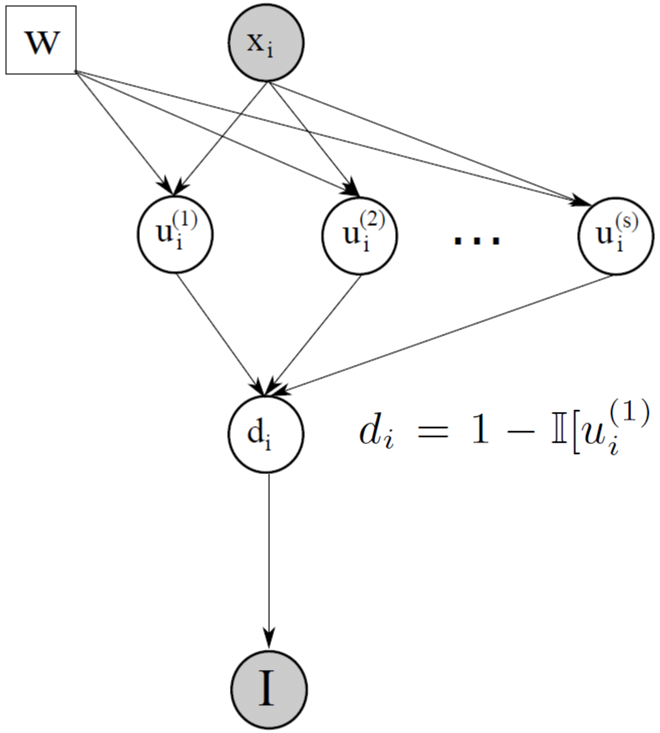
- Generate s labels for each sample
- Test if s labels disagree using d_i ($d_i \in \{0, 1\}$)
- Higher crispness, smaller no. of disagreements over the data



$$d_i = 1 - \mathbb{I}[u_i^{(1)} = u_i^{(2)} = \cdots = u_i^{(s)}]$$

Model: Cluster crispness

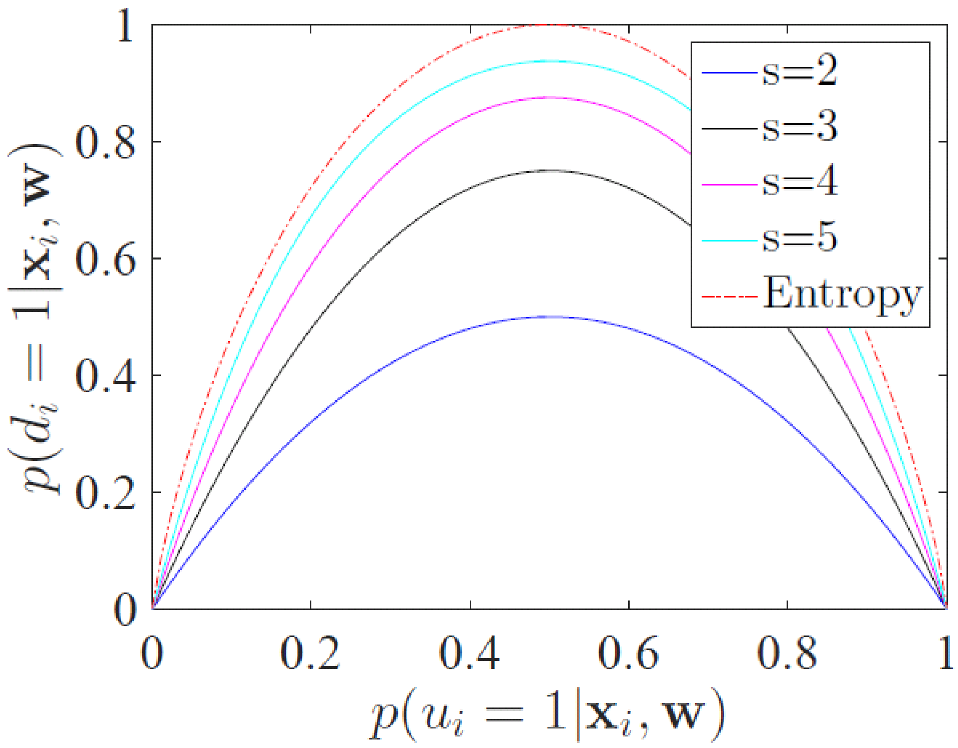
- Generate s labels for each sample
- Test if s labels disagree using d_i ($d_i \in \{0, 1\}$)
- Higher crispness, smaller no. of disagreements over the data
- m controls total crispness in all data points



$$d_i = 1 - \mathbb{I}[u_i^{(1)} = u_i^{(2)} = \cdots = u_i^{(s)}]$$

$$p(I = 1|\mathbf{d}) = \mathbb{I}[\sum_{i=1}^{n} d_i \leq m]$$ (m is a hyper-parameter)

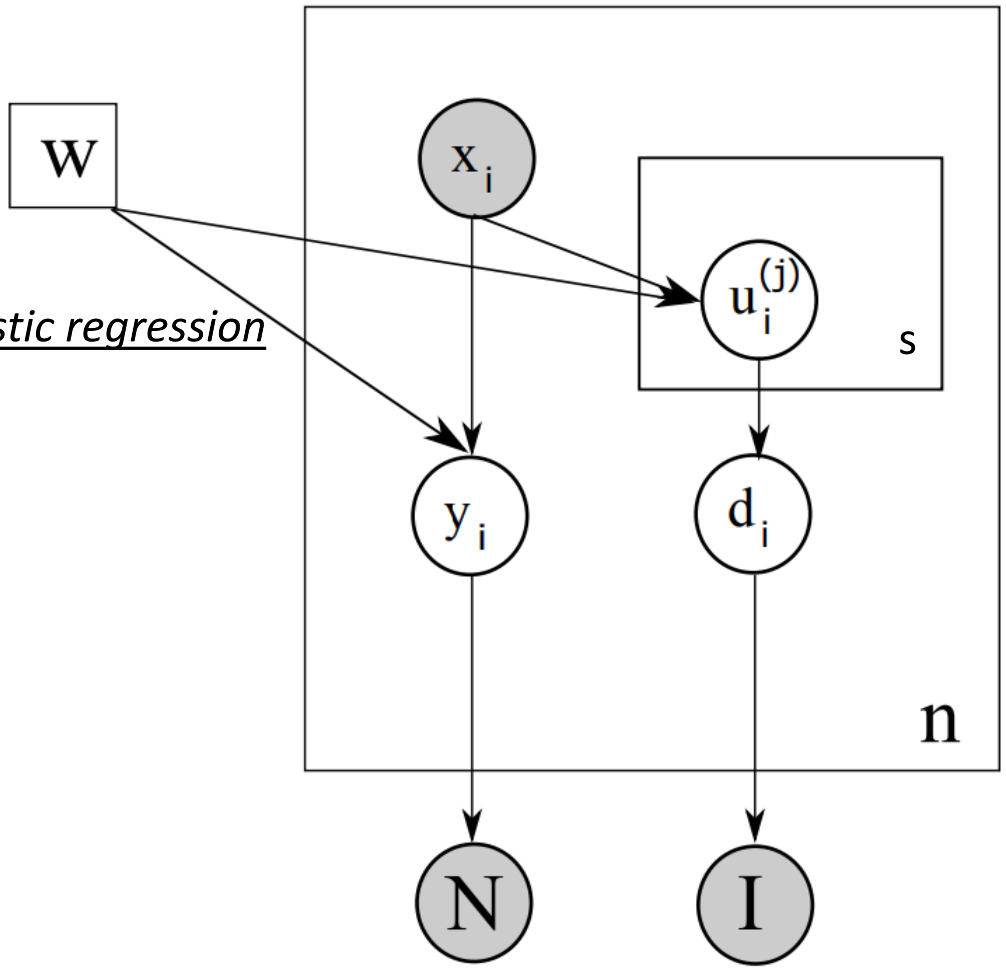
Cluster crispness vs. Entropy



Crispness vs. entropy

(two class)

Model



Logistic regression

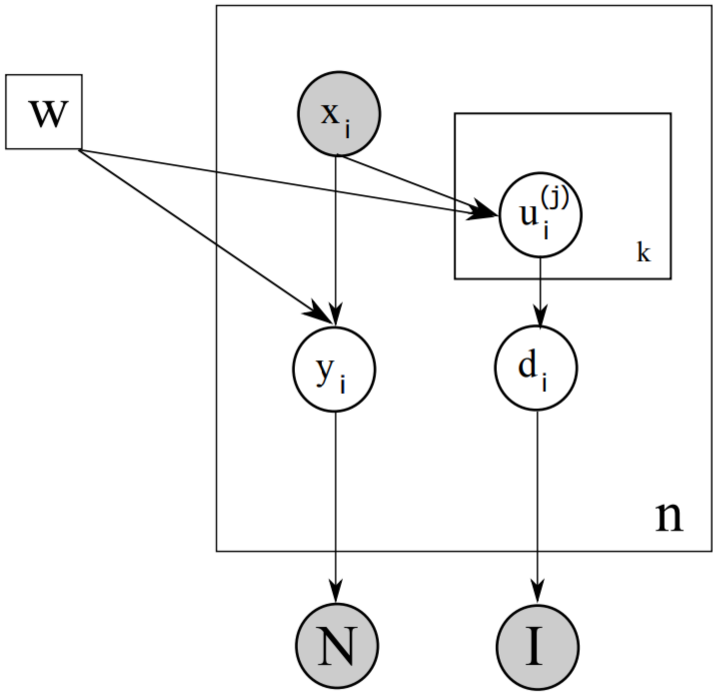
Hidden variables: y and u (marginalize d)

Cardinality constraints

Cluster crispness

Inference

Complete log-likelihood $\mathbf{L}_c(\mathbf{w}) = \log p(I, \mathbf{N}, \mathbf{y}, \mathbf{u} | \mathbf{X}, \mathbf{w})$



Auxiliary function $Q(\mathbf{w}, \mathbf{w}') = E_{\mathbf{y},\mathbf{u}|I,\mathbf{N},\mathbf{w}'}\left[\mathbf{L}_c(\mathbf{w})\right]$

$$= \zeta + \sum_{i=1}^{n}\Big[\big[\sum_{c=1}^{C} \boxed{p(y_i = c | \mathbf{N}, \mathbf{X}, \mathbf{w}')}\mathbf{w}_c^T \mathbf{x}_i - \log(\sum_{c=1}^{C} e^{\mathbf{w}_c^T \mathbf{x}_i})\big]$$

$$+ s \times \big[\sum_{c=1}^{C} \boxed{p(u_i = c | I, \mathbf{X}, \mathbf{w}')}\mathbf{w}_c^T \mathbf{x}_i - \log(\sum_{c=1}^{C} e^{\mathbf{w}_c^T \mathbf{x}_i})\big]\Big]$$

<u>E-step:</u> $p(y_i = c | \mathbf{N}, \mathbf{X}, \mathbf{w}') = \frac{p(y_i = c, \mathbf{N} | \mathbf{X}, \mathbf{w}')}{\sum_{l=1}^{C} p(y_i = l, \mathbf{N} | \mathbf{X}, \mathbf{w}')}$ where $\mathbf{w}' = \mathbf{w}^{(h)}$

Similarly for $P(u_i = c \mid I, X, w')$

<u>M-step:</u>

$$\mathbf{w}^{(h+1)} = \mathbf{w}^{(h)} + \eta \frac{\partial Q(\mathbf{w}, \mathbf{w}^{(h)})}{\partial \mathbf{w}}\Big|_{\mathbf{w} = \mathbf{w}^{(h)}}$$

Dynamic programming for E-step

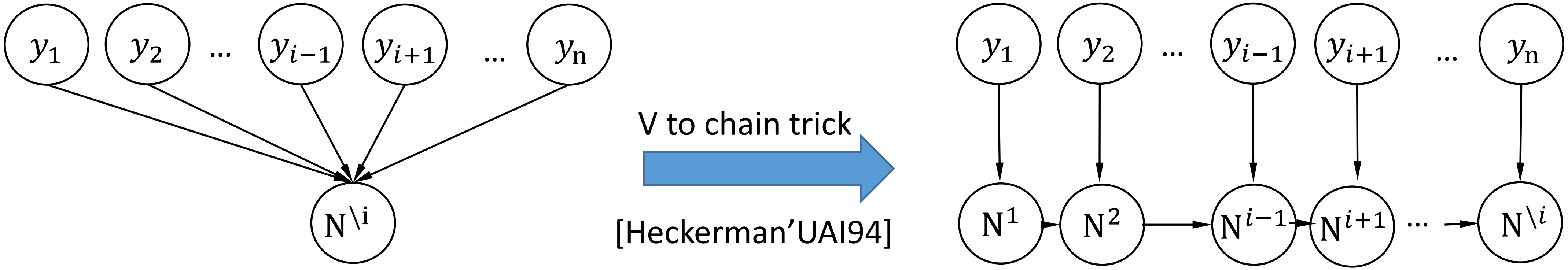
- $N_c^{\setminus i} = \sum_{j \neq i} I[y_j = c]$, $\quad p(y_i = c, \mathbf{N} = \mathbf{v} | \mathbf{X}, \mathbf{w}') = p(y_i = c | \mathbf{x}_i, \mathbf{w}') p(\mathbf{N}^{\setminus i} = \mathbf{v} - \mathbf{e}_c | \mathbf{X}, \mathbf{w}')$

- Compute $p(\mathrm{N}^{\setminus \mathrm{i}} | X, \mathrm{w}')$?

Dynamic programming for E-step

- $N_c^{\setminus i} = \sum_{j \neq i} I[y_j = c]$, $\quad p(y_i = c, \mathbf{N} = \mathbf{v} | \mathbf{X}, \mathbf{w}') = p(y_i = c | \mathbf{x}_i, \mathbf{w}') p(\mathbf{N}^{\setminus i} = \mathbf{v} - \mathbf{e}_c | \mathbf{X}, \mathbf{w}')$

- Compute $p(\mathrm{N}^{\setminus \mathrm{i}} | X, \mathrm{w}')$

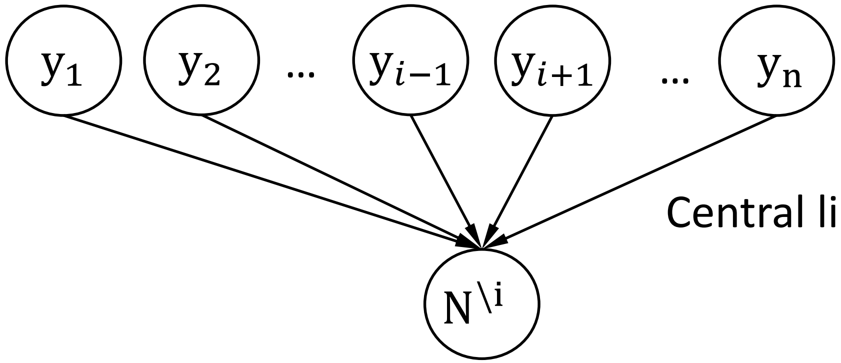


$O(C^n)$

$O(n^C)$!!!

- Infeasible for large C

Gaussian approximation for E-step



Central limit theorem $O(nC)$!!!

- $y_i \sim p(y_i = c|\mathbf{x}_i, w)$ and $y_1, y_2, \dots, y_n$ are independent given **X**
- $N_c^{\backslash i} = \sum_{j=1,\neq i}^{n} I[y_i = c], \forall c$
- $N^{\backslash i}$ follows central limit theorem when n is sufficiently large (true in real-world application)
- $N^{\backslash i}$ is multivariate normal with mean $\mu^{\backslash i} = \sum_{j=1,\neq i}^{n} \mu_i$ and variance $\Sigma^{\backslash i} = \sum_{j=1,\neq i}^{n} \Sigma_i$

Experiments on MNIST

- **Datasets:** MNIST with pairs of digits: uniform among two classes.
- **Baseline:** K-means, Maximum-margin clustering (MMC) [Xu'NIPS04], Regularized Information Maximization (RIM) [Krause'NIPS10] (*RIM uses cardinality constraints*).
- **Evaluation metric:** Normalized mutual information (NMI) [Jain'PRL10], averaged 10 times
- **Setting:**
  - MNIST is reasonably well separated, $m = 0, s = 2$
  - Consider both dynamic programming implementation and Gaussian approximation

Experiments on MNIST

- **Datasets:** MNIST with pairs of digits: uniform among two classes.
- **Baseline:** K-means, Maximum-margin clustering (MMC) [Xu'NIPS04], Regularized Information Maximization (RIM) [Krause'NIPS10] (*RIM uses cardinality constraints*).
- **Evaluation metric:** Normalized mutual information (NMI) [Jain'PRL10], averaged 10 times
- **Setting:**
  - MNIST is reasonably well separated, $m = 0, s = 2$
  - Consider both dynamic programming implementation and Gaussian approximation

Datasets	1vs.2	3vs.4	5vs.6	7vs.8	9vs.0
DCCC-D	0.70	**0.93**	**0.72**	**0.89**	**0.93**
DCCC-G	0.70	**0.93**	**0.72**	**0.89**	**0.93**

Experiments on MNIST

- **Datasets:** MNIST with pairs of digits: uniform among two classes.
- **Baseline:** K-means, Maximum-margin clustering (MMC) [Xu'NIPS04], Regularized Information Maximization (RIM) [Krause'NIPS10] (*RIM uses cardinality constraints*).
- **Evaluation metric:** Normalized mutual information (NMI) [Jain'PRL10], averaged 10 times
- **Setting:**
  - MNIST is reasonably well separated, $m = 0, s = 2$
  - Consider both dynamic programming implementation and Gaussian approximation

Datasets	1vs.2	3vs.4	5vs.6	7vs.8	9vs.0
DCCC-D	0.70	**0.93**	**0.72**	**0.89**	**0.93**
DCCC-G	0.70	**0.93**	**0.72**	**0.89**	**0.93**
RIM	**0.73**	0.89	0.69	0.88	**0.93**

Experiments on MNIST

- **Datasets:** MNIST with pairs of digits: uniform among two classes.
- **Baseline:** K-means, Maximum-margin clustering (MMC) [Xu'NIPS04], Regularized Information Maximization (RIM) [Krause'NIPS10] (*RIM uses cardinality constraints*).
- **Evaluation metric:** Normalized mutual information (NMI) [Jain'PRL10], averaged 10 times
- **Setting:**
  - MNIST is reasonably well separated, $m = 0, s = 2$
  - Consider both dynamic programming implementation and Gaussian approximation

Datasets	1vs.2	3vs.4	5vs.6	7vs.8	9vs.0
DCCC-D	0.70	**0.93**	**0.72**	**0.89**	**0.93**
DCCC-G	0.70	**0.93**	**0.72**	**0.89**	**0.93**
RIM	**0.73**	0.89	0.69	0.88	**0.93**
MMC	0.64	0.81	0.71	0.76	0.90
Kmeans	0.46	0.81	0.56	0.79	0.81

Experiments on real datasets

- **Datasets:**
  - **HJA bird-song dataset (13 classes):** each syllable is a sample
  - **MSCV2 (19 classes) + Voc12 are image annotation (20 classes) datasets:** each segment is a sample
- **Baseline:**
  - Consider Gaussian approximation $O(nC)$only due to the high complexity of dynamic programming $O(n^C)$
  - Skip MMC since MMC is not applicable for multi-class
- **Setting:**
  - $s \in \{2,3\}, m \in \{10,20,\dots,50\}$. Tuning based on likelihood on validation set wrt. N.

Datasets	HJA bird song	MSCV2	Voc12
DCCC-G	**0.40**	**0.31**	**0.12**
RIM	0.39	0.25	0.11
K-means	0.06	0.13	0.02

Conclusions

- We proposed a discriminative framework for clustering with cardinality constraints and high crispness.
- We proposed both exact and approximate inference.
- We verified the effectiveness of our method on synthetic and real world datasets.

?

References

- [Yu’14] “On learning from label proportions,” arXiv preprint arXiv:1402.5902, 2014.
- [Quadrianto’JMLR09] “Estimating labels from label proportions,” Journal of Machine Learning Research, vol. 10, no. Oct, pp. 2349–2374, 2009.
- [Musicant’ICDM07] “Supervised learning by training on aggregate outputs,” in International Conference on Data Mining, 2007, pp. 252–261.
- [Heckerman’UAI94] “A new look at causal independence,” in Proceedings of the Conference on Uncertainty in Artificial Intelligence, 1994, pp. 286–292.
- [Xu’NIPS04] “Maximum margin clustering,” in Advances in neural information processing systems, 2004, pp. 1537–1544.
- [Krause’NIPS10] “Discriminative clustering by regularized information maximization,” in Advances in neural information processing systems, 2010, pp. 775–783.
- [Jain’PRL10] “Data clustering: 50 years beyond k-means,” Pattern recognition letters, vol. 31, no. 8, pp. 651–666, 2010.
- [Basu’SDM04] “Active semi-supervision for pairwise constrained clustering,” in SIAM International Conference on Data Mining, 2004, pp. 333–344.
- [Bilenko’ICML04] “Integrating constraints and metric learning in semi-supervised clustering,” in International Conference on Machine Learning, 2004, pp. 11.
- [Wagstaff’ICML01] “Constrained k-means clustering with background knowledge,” in International Conference on Machine Learning, 2001, pp. 577--584 .